

**UCLA**

**UCLA Electronic Theses and Dissertations**

**Title**

Genetic Control of Expression and Splicing in the Developing Human Brain

**Permalink**

<https://escholarship.org/uc/item/0495m92k>

**Author**

Walker, Rebecca

**Publication Date**

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Genetic Control of Expression and Splicing in the Developing Human Brain

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

By

Rebecca Leah Walker

2019

© Copyright by  
Rebecca Leah Walker  
2019

## ABSTRACT OF THE DISSERTATION

Genetic Control of Expression and Splicing in the Developing Human Brain

by

Rebecca Leah Walker

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2019

Professor Daniel H. Geschwind, Chair

Neurodevelopmental and neuropsychiatric diseases, such as autism spectrum disorder (ASD) and schizophrenia (SCZ), are highly heritable, with hundreds of risk loci contributing to disease risk identified through large-scale genomic studies. The ability to interpret these susceptibility variants and their contributions to disease has been difficult due to the fact that many of these variants fall in non-coding regions of the genome, or in regions of high linkage disequilibrium. Given the non-coding nature of the majority of these variants, as well as their enrichment in known enhancers, many of these variants are predicted to regulate gene expression, which is known to be dependent on tissue, cell type and developmental stage.

Here, I have characterized functional genetic variation controlling transcriptional regulation in developing human brain to dissect common variation contributing to neurodevelopmental and early onset neuropsychiatric diseases, characterized by phenotypes originating in utero or early postnatal life. I have comprehensively profiled expression and splicing levels by RNA sequencing and high-density genotyping in 201 mid-gestational human

brains and have performed expression and splicing quantitative trait loci analysis, which is currently the largest eQTL study in the developing brain. I identified 7962 expression QTL (eQTL) and 4635 splice QTL (sQTL), including several thousand fetal-specific regulatory regions when compared to published QTL studies of the adult brain. I leveraged these eQTL and sQTL to identify splicing and transcriptional drivers affected by human genetic variation, by significant enrichment in experimentally determined transcription factor, DNA binding proteins, and RNA binding proteins binding sites. Further integration with experimental transcription factor knockdown data provide evidence that the regulatory regions identified through the eQTL and sQTL analysis are functional and validate that the changes seen in gene expression levels and/or splicing are likely due to the transcription factor's role in regulating that gene.

By integration with GWAS, I characterized the genes and isoforms contributing to specific neuropsychiatric disorders, including SCZ and ASD, as well as other cognitive or behavioral-related phenotypes. Specifically showing prenatal brain regulatory regions are significantly enriched for SCZ GWAS risk in a complimentary and additive manner to adult brain regulatory regions. I then perform gene co-expression network analysis and identify co-expressed modules of genes representing distinct biological processes in the developing brain. By integrating the QTL identified gene regulatory regions with co-expression modules and GWAS risk loci, I find SCZ and ASD impact distinct developmental gene co-expression modules. Yet, in both disorders, common and rare genetic variation converge. In ASD this convergence also implicates a specific cell type as well, superficial cortical neurons.

Additionally, integration of eQTL and sQTL with GWAS via transcriptome wide association identified dozens of novel candidate risk genes, highlighting shared and stage-specific mechanisms in SCZ. These analyses demonstrate the highly distinctive effects of transcriptional control, as well as divergent age-related contributions to disease. More broadly, these findings demonstrate the genetic mechanisms by which early developmental events have a striking and widespread influence on adult anatomical and behavioral phenotypes.

The dissertation of Rebecca Leah Walker is approved.

Eleazar Eskin

Roel Ophoff

Bogdan Pasaniuc

Daniel H. Geschwind, Committee Chair

University of California, Los Angeles

2019

## TABLE OF CONTENTS

|                                                                                             |           |
|---------------------------------------------------------------------------------------------|-----------|
| LIST OF TABLES AND FIGURES .....                                                            | viii      |
| ACKNOWLEDGEMENTS .....                                                                      | x         |
| VITA.....                                                                                   | xiii      |
| <b>CHAPTER 1: Introduction.....</b>                                                         | <b>1</b>  |
| 1.1: Genetic architecture of neuropsychiatric diseases .....                                | 2         |
| 1.2: Annotation of the non-coding genome .....                                              | 5         |
| 1.3: Expression and splicing quantitative trait loci analysis .....                         | 7         |
| 1.4: Transcriptomic profiling of the brain .....                                            | 9         |
| 1.5: Tissue and stage specificity for interpretation of neuropsychiatric disease loci ..... | 11        |
| 1.6: Integrative methods for GWAS and functional annotation.....                            | 13        |
| 1.7: Conclusions .....                                                                      | 15        |
| <b>CHAPTER 2: Robust identification of prenatal brain eQTL .....</b>                        | <b>19</b> |
| 2.1: Introduction.....                                                                      | 20        |
| 2.2: Methods.....                                                                           | 20        |
| 2.2.a: Data processing .....                                                                | 20        |
| 2.2.a.i Genotyping .....                                                                    | 22        |
| 2.2.a.ii: Gene expression quantification .....                                              | 23        |
| 2.2.a.iii: Sample Swap Identification .....                                                 | 24        |
| 2.2.b: Covariate Selection .....                                                            | 24        |
| 2.2.c: eQTL analysis .....                                                                  | 27        |
| 2.2.d: Correcting for ancestral differences .....                                           | 29        |
| 2.2.e: Enrichment in genome-wide annotations .....                                          | 31        |
| 2.2.f: Characterization of regulatory drivers .....                                         | 32        |

|                                                                          |           |
|--------------------------------------------------------------------------|-----------|
| 2.3: Results.....                                                        | 33        |
| <b>CHAPTER 3: Robust identification of prenatal brain sQTL .....</b>     | <b>39</b> |
| 3.1: Introduction.....                                                   | 40        |
| 3.2: Methods .....                                                       | 40        |
| 3.2.a: Splicing quantification .....                                     | 40        |
| 3.2.b: sQTL analysis .....                                               | 42        |
| 3.2.c: Enrichment in genome-wide annotations .....                       | 43        |
| 3.2.d: Characterization of splicing drivers .....                        | 43        |
| 3.2.e: Overlap with eQTLs .....                                          | 44        |
| 3.3: Results.....                                                        | 45        |
| <b>CHAPTER 4: Tissue specificity of QTLs .....</b>                       | <b>52</b> |
| 4.1: Introduction.....                                                   | 53        |
| 4.2: Methods .....                                                       | 53        |
| 4.2.a: Effect size correlations .....                                    | 53        |
| 4.2.b: Pi-hat analysis .....                                             | 54        |
| 4.2.c: eGene Overlap.....                                                | 56        |
| 4.2.d: Partitioned heritability by LD score regression .....             | 56        |
| 4.3: Results.....                                                        | 58        |
| <b>CHAPTER 5: eQTLs in the context of transcriptional networks .....</b> | <b>64</b> |
| 5.1: Introduction.....                                                   | 65        |
| 5.2: Methods .....                                                       | 65        |
| 5.2.a: WGCNA.....                                                        | 65        |
| 5.2.b: Module annotation .....                                           | 67        |

|                                                                |            |
|----------------------------------------------------------------|------------|
| 5.2.c: Module eQTL enrichment .....                            | 68         |
| 5.3: Results.....                                              | 70         |
| <b>CHAPTER 6: Transcriptome-wide association studies .....</b> | <b>75</b>  |
| 6.1: Introduction.....                                         | 76         |
| 6.2: Methods .....                                             | 76         |
| 6.2.a: Prenatal brain weights .....                            | 76         |
| 6.2.b: PsychEncode adult brain weights .....                   | 77         |
| 6.2.c: Published adult weights.....                            | 78         |
| 6.2.d: TWAS .....                                              | 78         |
| 6.2.e: High confidence adult brain implicated genes .....      | 79         |
| 6.2.f: TWAS overlap with GWAS loci .....                       | 79         |
| 6.2.g: Summary-based mendelian randomization .....             | 79         |
| 6.2.h: TWAS fine-mapping .....                                 | 79         |
| 6.3: Results.....                                              | 80         |
| 6.3.a: Schizophrenia .....                                     | 80         |
| 6.3.b: Intracranial volume .....                               | 89         |
| 6.3.c: Autism Spectrum Disorder .....                          | 93         |
| <b>CHAPTER 7: Conclusion and Future Directions.....</b>        | <b>95</b>  |
| <b>REFERENCES .....</b>                                        | <b>103</b> |

## LIST OF TABLES AND FIGURES

### Chapter 1

Figure 1.1: Study Design and Methodological Overview of QTL Analysis

### Chapter 2

Figure 2.S1: Overview of raw data processing and QC

Figure 2.S2: Sample demographics and dataset quality control

Figure 2.S3: Identification of optimal number of HCP selection and correlations of all covariates in final cis-eQTL analysis

Figure 2.S4: Ancestry correction

Table 2.S1: Expressed gene biotypes

Figure 2.1: Characterization of eQTLs

### Chapter 3

Figure 3.S1: Overview of raw data processing and QC

Figure 3.S2: Identification of optimal number of HCP selection and correlations of all covariates in final sQTL analysis

Figure 3.1: Characterization of sQTLs and Comparison to eQTLs

Figure 3.S3: Genes harboring both an eQTL and sQTL

Table 3.S1: Genes having an overlapping eQTL and sQTL

### Chapter 4

Figure 4.S1: Cross-tissue overlap of eQTLs.

Figure 4.1: Age Specificity of Brain eQTLs

Figure 4.2: eQTL and sQTL enriched for complex disease heritability

### Chapter 5

Figure 5.S1: WGCNA parameters and module preservation

Table 5.S1 : Rare variant disease associated gene count

Figure 5.S2: Permutations of module GWAS enrichment

Figure 5.1: Prenatal Brain Co-expression Networks

## Chapter 6

Figure 6.1: SCZ-PsychENCODE Adult Brain TWAS

Table 6.S1: Fetal brain-SCZ genes significant across TWAS and SMR

Figure 6.2: SCZ TWAS

Figure 6.3: ICV TWAS

Figure 6.4: ASD TWAS

## Chapter 7

There are no figures or tables in this section.

## ACKNOWLEDGEMENTS

I would like to thank my advisor, Dan Geschwind, for his excellent mentorship and guidance throughout my graduate career. This work would not have been possible without his support. He saw my potential as an entering student and provided me with the tools, resources and environment to learn, and fail, but ultimately succeed in becoming a true scientist. He set the bar high and pushed me past my perceived limits, yet knew when I needed words of encouragement the most. I am forever grateful for the invaluable opportunities he has provided me, as they have truly shaped the person I am today.

I would further like to express my gratitude to Gokul Ramaswami, Michael Gandal, and Christopher Hartl, who I also consider great mentors, co-authors, and friends of mine. They graciously put up with my endless questions, from sharing code to scientific discussions, editing manuscripts to (nonalcoholic) beer 5Ks. This work really wouldn't have been possible without their expertise and I cannot thank them enough for lending me their time, support, and encouragement along the way. I would like to thank Damon Polioudakis for making our office the best place to sit in the lab, indulging in many my work breaks and stupid jokes, lending a hand whenever Adobe Illustrator would betray me, and proofreading everything from emails to manuscripts and Will Pembroke, for the daily life chats and scientific discussions over the cheapest possible coffee option and invaluable feedback on presentations, figures, and manuscripts. I would also like to thank Elizabeth Ruzzo and Katie Kugler, lab mates turned close friends, who I look up to as incredibly strong, hard-working women in science. They have always provided great advice on navigating life both inside and outside academia, from dance floors to workout classes. I would like to thank the other current and former postdocs, students, and members of the Geschwind Lab, for creating an amazing environment to work in, sharing scientific insights and of course, socializing.

I would like to thank the co-authors of my paper, specifically Luis de la Torre-Ubieta and Jason Stein for conceptualizing much of this project. They collected and processed the samples and got me started when I first joined the lab. I am grateful they allowed me to carry this project through and I hope our *Cell* paper made them proud.

I would like to thank key people in the Bioinformatics Interdepartmental Program, including my two committee members Eleazar Eskin and Bogdan Pasaniuc. They both played crucial roles in my understanding of the fundamentals of bioinformatics, and have provided me with great feedback and input on my work along the way. Additionally, I'd like to acknowledge Kimberly Insigne, Rob Brown, Shan Sabri, Adriana Arneson, and Doug Arneson, who I am lucky to have gotten to know and learn from these past five years.

I would like to thank the Neurobehavioral Genetics community at UCLA, including my committee member Roel Ophoff, for being such an incredible group of students, postdocs, and faculty to share ideas and engage in meaningful scientific discussions with.

I would like to thank Rita Cantor and Paivi Pajukanta for allowing me the opportunity to teach a portion of their graduate course, Advance Human Genetics, for the past three years. They allowed me the freedom to create lectures around topics I am both passionate about and could provide valuable insights on. I had so much fun teaching and always looked forwards to the Friday's where I got to do so.

I would like to acknowledge the unconditional support of my incredible parents, who believed in me even when I doubted myself. I am thankful that have been my biggest cheerleaders and always knew that my childhood stubbornness would lead to determination and

perseverance, skills they have always encouraged and have been crucial for my success in graduate school.

I would like to thank my dear friends and family for their endless supported throughout graduate school- my brother Andrew, my cousin Jessica, my Aunt Sandra and Uncle Dan, my Aunt Carol and Uncle Jim, Ariela, Spencer, Danielle, Susanna, Monica, Kim, Eric. Thank you for being there for me, this would not be possible without you.

I would like to acknowledge that all chapters contain significant content from, “Genetic control of expression and splicing in developing human brain informs disease mechanisms”, a manuscript published October 17, 2019 in *Cell*, Volume 179, Issue 3, p 750-771. DOI: 10.1016/j.cell.2019.09.021. This work was co-authored by Gokul Ramaswami, Christopher Hartl, Nicholas Mancuso, Michael J. Gandal, Luis de la Torre-Ibieta, Bogdan Pasaniuc, Jason L. Stein, and Daniel H. Geschwind.

Chapter 6 includes work that I contributed to “Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder”, published on December 14, 2018 in *Science* Volume 362, Issue 6420. DOI: 10.1126/science.aat8127.

Tissue for this work was collected from the UCLA CFAR (5P30 AI028697). RNA-seq libraries were sequenced by the UCLA Neuroscience Genomics Core. This work was supported by the NIH (5R37 MH060233, 5R01 MH094714, 1R01 MH110927) to D.H.G, and the National Institute of Neurological Disorders And Stroke of the National Institutes of Health Training Grant (T32NS048004 R.L.W.)

## VITA

### Education

Doctoral Candidate in Bioinformatics, University of California, Los Angeles

September 2014 – December 2019

Bachelor of Science in Genetics, University of California, Davis

September 2008 - June 2012

### Publications

**Walker RL**, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L, Pasaniuc B, Stein JL, Geschwind DH. Genetic Control of Expression and Splicing in the Developing Human Brain Informs Disease Mechanisms. *Cell* 179.3 (2019): 750-771.

Gandal MJ, Zhang P, Chen C, **Walker RL**, Liu S, Won H, Hadjimichael E, van Bakel H, Haney J, Parhami S, Losada P, Xia Y, Dai R, Wang D, Yang Y, Xu M, Warrell J, Fitzgerald D, PsychENCODE Consortium, White K, Peters M, Gerstein M, Liu C, Iakoucheva L, Pinto D, Geschwind DH. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*. 2018 Dec 14;362(6420):eaat8127.

Cantor RM, Navarro L, Won H, **Walker RL**, Lowe JK, Geschwind DH. ASD restricted and repetitive behaviors associated at 17q21.33: genes prioritized by expression in fetal brains. *Molecular psychiatry*. 2018 Apr;23(4):993.

## **CHAPTER ONE**

### Introduction

## **1.1: Genetic architecture of neuropsychiatric diseases**

Neurodevelopmental and neuropsychiatric diseases are complex conditions with little understanding of their underlying molecular etiology (Geschwind and Flint, 2015). While family and twin based studies have long shown significant heritability for these diseases (Bolton et al., 1994; Kremen et al., 1994; Polderman et al., 2015; Ritvo et al., 1989), only recently have large-scale genomic studies been adequately powered to identify genetic variants associated with increased risk for neurodevelopmental and neuropsychiatric diseases (Risch and Merikangas, 1996), such as schizophrenia (SCZ), autism spectrum disorder (ASD), and attention-deficit hyperactivity disorder (ADHD) (Autism Spectrum Disorders Working Group of The Psychiatric Genomics, 2017; Gratten et al., 2014; Pardini et al., 2018). As the genetic architecture of these disorders continues to be unraveled, connecting genetic variation to mechanisms and functional neurobiology remains at the forefront of furthering our understanding of neuropsychiatric disorders.

The genetic contributions of these diseases can be broken down into two main classes: common and rare variation (Cargill et al., 1999; Chakravarti, 1999; Reich and Lander, 2001; Schork et al., 2009; Zwick et al., 2000). Additionally, variants can be classified by effect size, which is a function of what role the variant is playing in a biological process and is usually inversely related to their frequency (Goldstein, 2009; Manolio et al., 2009). The distinction is made between these types of variation not only because they implicate different kinds of functional interpretation but are detected through different methodology and study design (Gratten et al., 2014; Park et al., 2010).

Common variation, which signifies variants having a frequency of greater than 1% in the population, has become easy and cost-effective to assay genome-wide due to advances in single nucleotide polymorphism (SNP) microarray technology and underlying linkage patterns allowing

for efficient surveying of well-spaced SNPs throughout the entire genome (Barrett and Cardon, 2006; Pe'er et al., 2006). These advances have allowed tremendous progress to be made by identifying common variants that influence disease through genome-wide association studies (GWAS) — the approach of testing millions of single-nucleotide polymorphisms (SNPs) for statistical association with disease status. For neuropsychiatric diseases, the success of GWAS came to fruition when sample sizes became relatively large (Billingslea et al., 2014; Cross-Disorder Group of the Psychiatric Genomics, 2013; McCarthy et al., 2008; Psychiatric, 2011), currently reaching ~100,000 individuals for SCZ (Pardinas et al., 2018; Ripke et al., 2014) and identifying 145 loci associated with SCZ risk. Yet for autism spectrum disorder, one of the most phenotypically and etiologically heterogeneous disorders (Geschwind, 2011), there have been very few loci identified through this methodology, most likely due to smaller sample sizes (~46,000) compared to SCZ studies (Autism Spectrum Disorders Working Group of The Psychiatric Genomics, 2017; Grove et al., 2019). However, common variation is a large contributor of genetic risk for both SCZ and ASD (Gaugler et al., 2014; Klei et al., 2012) and with larger samples sizes and the use of endophenotypes, GWAS will begin to reveal more significant loci (Robinson et al., 2016).

Rare variation, on the other hand, represents variants that occur at frequencies less than 1% in the population (and often <0.1%) and are either new (*de novo*) or have arisen in only the past few generations. Therefore, rare variants have not been subject to selection in most cases and have an increased chance of being under negative selection due to deleteriousness or reduced fitness. Given this framework, rare variation tends to have larger effect sizes with respect to disease risk than common variation (Gorlov et al., 2008; Pritchard, 2001). Whole exome and whole genome sequencing in families with an affected child have enabled the detection of rare *de novo* and inherited variants including single nucleotide variants (SNVs), small chromosomal arrangements, indels, inversions, and copy number variants (CNVs) associated with disease risk.

Specifically, sequencing efforts have led to the identification of rare mutations in protein-coding domains and therefore individual genes, and likely causal mutations, contributing to risk (Iossifov et al., 2014).

For ASD, there has been a large focus on family based studies to identify rare *de novo* and inherited protein-disrupting SNVs and CNVs that increase risk for ASD (De Rubeis et al., 2014; Iossifov et al., 2014; Iossifov et al., 2012; Levy et al., 2011; Marshall et al., 2008; Neale et al., 2012; O'Roak et al., 2014; O'Roak et al., 2012; Ruzzo et al., 2019; Sanders et al., 2011; Sanders et al., 2012; Sebat et al., 2007). These studies have been fairly successful in gene discovery, with currently 137 genes identified harboring protein-disrupting SNVs associated with ASD (Ruzzo et al., 2019; Satterstrom et al., 2019), though the variants in these known genes only account for 3-5% of ASD liability (Constantino et al., 2010; Gaugler et al., 2014). Additionally, many studies show a high rate of *de novo* CNVs associated with ASD (Levy et al., 2011; Marshall et al., 2008; Pinto et al., 2010; Sanders et al., 2011; Sanders et al., 2012), with several regions strongly implicated: 16p11.2, 7q11.23, 15q11-13, 22q11.2, and 1q21.1 (Glessner et al., 2009; Jacquemont et al., 2006; Levy et al., 2011; Marshall et al., 2008; Sanders et al., 2011; Sebat et al., 2007; Weiss et al., 2008). However, these CNV regions tend to be large, containing multiple genes, and therefore identifying the causal gene or genes remains a challenge.

For SCZ, studies show a significant role for rare, large CNVs contributing to risk (Buizer-Voskamp et al., 2011; Sebat et al., 2009; Walsh et al., 2008). But, in contrast to ASD, very few genes identified carrying protein-truncating SNVs have been associated with the disorder so far (Geschwind and Flint, 2015; Girard et al., 2011; Xu et al., 2012). A large deletion at 22q11.21 has long been associated with SCZ risk (Karayiorgou et al., 1995), with more recent genome-wide studies implicating variants at 15q13.3, 1q21.1, 16p11.2, 3q29, 2p16.3, and 7q36.3 (Fromer et al., 2014; Ingason et al., 2011; Kirov et al., 2009; Levinson et al., 2011; McCarthy et

al., 2009; Moreno-De-Luca et al., 2010; Mulle et al., 2010; Purcell et al., 2014; Vacic et al., 2011). While CNVs contribute significantly to disease risk, they have high phenotypic heterogeneity, with several being associated with other neuropsychiatric disorders such as ASD, Bipolar disorder (BP), depressive disorder (MDD) and intellectual disability (ID), revealing the complexity of interpretation for these types of variants (Cooper et al., 2011; Grozeva et al., 2010; Malhotra et al., 2011; Mefford et al., 2008; Moreno-De-Luca et al., 2010).

These studies establish genetics as the largest contributor to disease liability for ASD and SCZ, among other neuropsychiatric disorders, with risk variants spanning the allelic spectrum. They also indicate that common variation contributes the most with estimates around 50% for ASD and 35% of disease liability for SCZ (de la Torre-Ubieta et al., 2016; Gandal et al., 2016; Gaugler et al., 2014; Geschwind and Flint, 2015; Ripke et al., 2014; Sullivan et al., 2012). While the success of GWAS was made possible by harnessing the linkage disequilibrium structure of the genome for efficient assaying of SNPs, it is the largest hinderance in the ability to biologically interpret significantly associated loci and pinpoint causal mutations (Spain and Barrett, 2015), along with the fact that 80% of GWAS associations for common disease lie outside coding regions (Farh et al., 2015). This has motivated the “post-GWAS era” to focus on methods to fine map casual mutations as well as identify their functional impact (Cannon and Mohlke, 2018; Gallagher and Chen-Plotkin, 2018).

## **1.2: Annotation of the non-coding genome**

The assembly of the human genome revealed the genetic landscape and distribution of genes, transposable elements, and CpG islands, revealing an estimated 30,000-40,000 protein coding genes contributing to only 2% of the genome sequence (Lander et al., 2001; Venter et al., 2001). While the remaining 98% has been referred to as “junk DNA” (Ohno, 1972), there has been countless lines of evidence implicating its many roles in driving evolution (Biemont and

Vieira, 2006), to gene regulation either by producing non-coding transcripts (Ghildiyal and Zamore, 2009) or encoding regulatory DNA elements such as enhancers and repressors (The GTEx Consortium, 2015; The GTEx Consortium et al., 2017). Additionally, many cross-species studies find many regions of evolutionary conservation in non-coding elements further implicating an important functional role (Bejerano et al., 2004; Pennacchio et al., 2006; Rands et al., 2014; Siepel et al., 2005).

In response to the assembly of human genome, the ENCyclopedia Of DNA Elements (ENCODE) project was established to annotate all functional elements in the human genome (Encode Project Consortium, 2004, 2011; Encode Project Consortium et al., 2007), which used a combination of both high-throughput experimental ( Hi-C, ATAC-seq, ChIP-seq, bisulfate sequencing, and RNA-seq) and computational approaches to identify genes, promoters, enhancers, transcription factor binding sites, methylation sites, DNase I hypersensitivity sites, and chromatin modifications, among others (Encode Project Consortium, 2004). While the genome is consistent throughout most cells in the body, it is the dynamic epigenetic landscape that controls gene expression and cell type specific biological function (Jones, 2012; Rivera and Ren, 2013; Smith and Meissner, 2013; Zhou et al., 2011). There have been multiple large-scale projects, including Roadmap Epigenomics, FANTOM and GTEx (Ernst et al., 2011; Roadmap Epigenomics Consortium et al., 2015; The FANTOM Consortium and the RIKEN PMI and CLST et al., 2014; The GTEx Consortium, 2015) with the goal of annotating regulatory regions in a tissue and cell-type specific manner. While these efforts provide a catalog of functional genomic elements, many of the experiments are performed only in a few samples and do not capture the variability among individuals in the human population and many cell types have yet to be profiled. Therefore, while these annotation resources may provide putative function to genetic variants, there is still a very incomplete picture of regulatory function of the genome (Lappalainen, 2015).

Given the non-coding nature of the significant GWAS loci associated with neurodevelopmental and neuropsychiatric diseases, interpreting the role of these variants has relied on these efforts to functionally annotate the genome (Gandal et al., 2016; Nica and Dermitzakis, 2013; Schaid et al., 2018). Further, it has been shown that GWAS identified variants enrich in known functional regions of disease relevant tissue types and developmental stage as inferred through chromatin accessibility (Cockerill, 2011; de la Torre-Ubieta et al., 2018), evolutionary conserved regions (Siepel et al., 2005), and signature histone modifications (Schaub et al., 2012; Visel et al., 2009), indicating many of these variants likely function through transcriptional regulation (Li et al., 2016; Maurano et al., 2012; Nicolae et al., 2010; Ward and Kellis, 2012). Furthermore, disease associated variants that alter gene expression, referred to as expression quantitative trait loci (eQTL), contribute disproportionately more to disease heritability compared to other variant types (Gamazon et al., 2019; Hormozdiari et al., 2018; Torres et al., 2014)

### **1.3: Expression and splicing quantitative trait loci analysis**

Expression quantitative trait loci (eQTL) analysis is a common approach for characterizing the effects of genetic variation on levels of gene expression while also defining gene regulatory regions (Battle et al., 2014; Emilsson et al., 2008; Grundberg et al., 2012; Lappalainen et al., 2013; Montgomery et al., 2010; Pickrell et al., 2010a; Rockman and Kruglyak, 2006). The first studies of this kind were carried out in yeast, showing connections could be identified linking genetic loci to the genes they modulate both directly and indirectly (Brem et al., 2002), which provided a framework for assessing these relationships in human populations (Bestman et al., 2015; Cheung et al., 2005; Stranger et al., 2007). Given transcriptome-wide gene expression measures from microarray or RNA-sequencing and genotyping data, eQTL analysis identifies common genetic loci associated with changes in gene expression in a tissue-specific manner, as

eQTL relationships are highly dependent on cell type and developmental stage (Dimas et al., 2009; Flutre et al., 2013; Gerrits et al., 2009; Nica et al., 2011; The GTEx Consortium, 2015). They have also unveiled properties of cis regulation, from consistency across different human populations (Stranger et al., 2012), to providing a molecular mechanism of local human adaptation driven by gene expression changes (Fraser, 2013), as well as their proximity to genes, often located in promoters and enhancers (Gaffney et al., 2012).

eQTL analysis has been carried out in numerous large-scale human studies including the GTEx Consortium (The GTEx Consortium, 2015; The GTEx Consortium et al., 2017), Geuvadis Consortium (Lappalainen et al., 2013), CommonMind (Fromer et al., 2016), PsychENCODE (Wang et al., 2018c), assaying expression across most tissue types including the brain. As sample sizes of these studies have increased (200-5,000 individuals per tissue type), with the largest seen in more easily accessible tissues such as blood and adipose (Grundberg et al., 2012; Zhang et al., 2015; Zhernakova et al., 2017), so have the number of genes with detected eQTLs. For neuronal tissues, the latest v8 GTEx release includes 13 different brain regions from 100-250 adult donors (Aguet et al., 2019). The largest brain dataset currently is represented in PsychENCODE, which aggregated 1,866 post-mortem prefrontal cortex samples across multiple studies including CommonMind to identify 33,000 genes with a significant eQTL, nearly approaching all expressed genes in the adult brain (Wang et al., 2018c).

QTL analysis has also been applied to many other quantitative molecular traits with a defined genomic locus, such as splicing (sQTL) (Gutierrez-Arcelus et al., 2015; Lalonde et al., 2011; Pickrell et al., 2010a; Pickrell et al., 2010b), chromatin state (dsQTL when assaying DNase I sensitivity) (Degner et al., 2012), methylation (mQTL) (Banovich et al., 2014; Bell et al., 2011; Gutierrez-Arcelus et al., 2013; Heyn et al., 2013), and protein levels (pQTL) (Battle et al., 2015). As these other cellular properties have also become more easily quantified due to advances and

scalability in molecular assays, these various QTL studies have also become more mainstream and have revealed significant genetic control throughout these many layers of regulation (Gaffney, 2013). Together, they provide insight into how biochemical changes in epigenetic state can lead to downstream changes in gene expression and protein abundance, and the combinatorial nature of the numerous regulatory stages for normal cellular function (Gutierrez-Arcelus et al., 2013; Pai et al., 2012; Pai et al., 2015).

#### **1.4: Transcriptomic profiling of the brain**

Taking a step back from the genetic control of gene expression, expression profiling alone can provide many insights into biology. As an orthogonal approach to studying neurodevelopmental and neuropsychiatric diseases through a genetic lens, transcriptomic studies, which assay levels of gene and isoform expression from high-throughput RNA sequencing or microarray, have identified many genes and pathways altered in disease (Gandal et al., 2018a). Genes and pathways altered in disease were initially identified through differential gene expression (DGE), testing one gene at a time and resulting in a gene list based off of significance differences in mean expression between groups. The field has since moved toward network-based methods that organize transcriptional data into co-expression modules based on gene connectivity, highlighting relationships between genes and providing context for analyzing genes in a larger biological system (Geschwind and Konopka, 2009).

Many transcriptome-wide studies have profiled gene expression in the brain throughout the human developmental trajectory, brain regions, cell types, and disease status, and used DGE and co-expression network based approaches to identify genes whose expression profiles differ between state or condition (Hawrylycz et al., 2012; Kang et al., 2011; Miller et al., 2014; Oldham et al., 2008; Parikshak et al., 2013; Purcell et al., 2001; Sunkin et al., 2013). Transcriptional studies of post-mortem brains of individuals with ASD and SCZ consistently find dysregulated

pathways associated with disease (Altar et al., 2005; Garbett et al., 2008; Ginsberg et al., 2012; Hakak et al., 2001; Hashimoto et al., 2008; Parikshak et al., 2016; Purcell et al., 2001). Studies using weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) show expression abnormalities in genes responsible for cortical patterning, neuronal signaling, immune and inflammatory response, transcriptional regulation and synaptic development in ASD (Gupta et al., 2014; Parikshak et al., 2016; Voineagu et al., 2011), as well as developmental dysregulation in modules associated with neurite outgrowth, neuronal differentiation, and cellular signaling for SCZ (Chen et al., 2013; Torkamani et al., 2010).

In addition to identifying genes whose expression differs between cases and controls, networks provide biological context for interpretation of disease associated genes identified through genomic studies. Networks of the human brain transcriptome are highly reproducible and layout gene expression trajectories from stage and region-specific biological processes spanning early brain development to adulthood (Hawrylycz et al., 2015; Kang et al., 2011; Oldham et al., 2008). These networks have been extremely valuable for interpretation of ASD associated genes. Studies mapping ASD candidate-genes identified through large family-based studies and GWAS signal onto network modules find enrichment in neuronal cell-type specific modules (Ben-David and Shifman, 2012). In addition, ASD risk genes harboring rare *de novo* mutations converge onto modules which relate to processes specific to early and mid-fetal brain development, providing strong evidence of neurodevelopmental origin of the disorder (Parikshak et al., 2013; Willsey et al., 2013). Additionally, both co-expression and protein interaction networks have aided in the interpretation of many SCZ risk genes, largely from rare *de novo* gene-disrupting CNVs, and also provide evidence that disease risk similarly converges on genes expressed in the prefrontal cortex during fetal development (Gilman et al., 2012; Gulsuner et al., 2013; Xu et al., 2012).

Pre-mRNA splicing can also be assessed using RNA-sequencing based transcriptomic profiling at the level of isoforms and splice-junctions (Bray et al., 2016; Li and Dewey, 2011; Li et al., 2018; Shen et al., 2014; Trapnell et al., 2009). Splicing is widely used in the brain, where specialized isoforms are known to affect key events in neuronal development such as cell-fate determination, axon guidance, and synaptogenesis, with evidence of isoform-switches between developing and mature neurons (Darnell, 2013; Li et al., 2007; Raj and Blencowe, 2015; Zheng and Black, 2013). Furthermore, many splicing defects have been implicated in neurodevelopmental and neurodegenerative diseases, including ASD and SCZ (Da Cruz and Cleveland, 2011; Fromer et al., 2016; Irimia et al., 2014; Jaffe et al., 2018; Ling et al., 2013; Parikshak et al., 2016). Further, networks built on isoform level quantifications show larger genetic enrichment and disease specificity compared to gene level quantification, refining the disease signal of ASD and SCZ and implicating splicing regulation to be of great importance to further our understanding of disease mechanisms (Gandal et al., 2018b).

Transcriptomics studies have aided the understanding of neurologic disease mechanisms by contextualizing gene and isoform expression changes and mapping of previously implicated variants and genes into biological context, revealing convergent molecular pathology (Parikshak et al., 2015). Furthermore, disease-associated modules in both ASD and SCZ studies provide strong evidence of transcriptional and splicing abnormalities that are of developmental origin, suggesting spatiotemporal specificity of relevant disease biology.

### **1.5: Tissue and stage specificity for interpretation of neuropsychiatric disease loci**

Integrating GWAS hits with gene expression data through eQTL analysis has allowed for identification of new functional loci without having previous knowledge of regulatory regions (Nica and Dermitzakis, 2013). This is most effective when the eQTL analysis is performed in a tissue relevant to the associated phenotype (Hernandez et al., 2012; Nica and Dermitzakis, 2008).

Transcriptome-wide studies profiling gene expression in the brain throughout human developmental trajectory show prominent patterns of expression that differ spatiotemporally (Colantuoni et al., 2011; Kang et al., 2011; Sunkin et al., 2013). Previous studies (De Rubeis et al., 2014; Gilman et al., 2012; Gulsuner et al., 2013; McCarthy et al., 2014; Parikshak et al., 2013; Weinberger, 1987; Willsey et al., 2013) have suggested that genetic disruption of these patterns, particularly during prenatal cortical development, leads to developmental and neuropsychiatric disease. These findings highlight the need to map regulatory variation within this critical time point.

There have been a few studies that assay the epigenetic landscape of the fetal brain with the goal of contextualizing neuropsychiatric disease loci. Long-range 3D chromosomal interaction maps during human corticogenesis have provided annotations of putative distal regulatory regions and were shown to link SCZ GWAS loci to candidate genes, many of which were not previously implicated (Won et al., 2016). Chromatin accessibility, by high throughput ATAC-sequencing, conducted in the germinal zones and cortical plate of developing brain showed significant enrichment of genetic loci associated with SCZ, attention deficit hyperactivity disorder (ADHD), depressive symptoms, neuroticism, and intracranial volume (ICV) in open regions of the germinal zone (de la Torre-Ubieta et al., 2018). Additionally, in their efforts to identify global and tissue specific regulatory elements, the Roadmap Epigenetics Consortium (Roadmap Epigenomics Consortium et al., 2015) assayed both fetal and adult brain samples for multiple histone modifications, DNA accessibility, and DNA methylation. These annotations have been incorporated into ChromHmm, a machine-learning model which learns chromatin-states based off aggregation of multiple genomic signatures, and provide genome-wide annotations of known classes of functional genomic elements (Ernst and Kellis, 2012, 2017), which have been used widely for interpretation of non-coding regions (Andersson et al., 2014; Hibar et al., 2015; Won et al., 2016).

However, these previous epigenetic analyses of fetal brain have only included a few samples each. Developmental tissues are very difficult to collect, and therefore existing fetal brain eQTL analyses are relatively small (Jaffe et al., 2018; O'Brien et al., 2018) and none include analysis of splicing. All sQTL analyses to date have been performed in adult brain (Fromer et al., 2016; Ramasamy et al., 2014; Takata et al., 2017; The GTEx Consortium, 2015). However, little is known about how human allelic variation affects putative regulatory interactions during brain development, which is crucial for human higher cognition and brain evolution (Geschwind and Rakic, 2013; Nord et al., 2015; Ward and Kellis, 2012). Therefore, to dissect the functional genetic variation of neurodevelopmental and early onset neuropsychiatric diseases characterized by phenotypes originating in utero or early postnatal life (Parikshak et al., 2015), studying eQTLs and sQTLs in fetal brain tissue is crucial.

### **1.6: Integrative methods for GWAS and functional annotations**

Given published GWAS summary statistics and the appropriate functional annotations from disease relevant tissue and cell types, there have been numerous statistical methodologies established to leverage both types of data in order to gain insight into the genetics of complex diseases. These methodologies include GWAS fine-mapping, heritability enrichments of functional annotations and gene identification through colocalization and gene-based association tests (Pasaniuc and Price, 2017), which have shown increases in power when functional data have been combined with previous approaches (Finucane et al., 2015; Kichaev et al., 2014).

Given the highly polygenic nature of neuropsychiatric diseases and the continued identification of variants with small effect, efforts have been made to estimate the heritability explained by these disease-associated SNPs. A common method to estimate heritability, LD Score Regression (Bulik-Sullivan et al., 2015), leverages linkage-disequilibrium patterns to

correct for genomic inflation of GWAS association statistics in order to distinguish polygenic signals from confounding from population stratification and cryptic relatedness. Furthermore, given the availability of functional annotations, models of heritability have been extended to evaluate enrichment of polygenic signal of disease heritability within a specific functional category. Stratified LD Score Regression (Finucane et al., 2015), among other methods to estimate polygenic enrichment of genomic annotations (Pickrell, 2014), have since shown significant enrichment in cell type specific functional annotations for disease heritability.

Methods for fine mapping of GWAS identified loci have also shown benefits of incorporating functional annotation in their models (Chen et al., 2016b; Chung et al., 2014; Kichaev and Pasaniuc, 2015; Kichaev et al., 2014; Li and Kellis, 2016). Given that marks of active gene regulation have been shown to be enriched for disease associated variants (Maurano et al., 2012; Trynka et al., 2013), fine-mapping methods have been developed to integrate functional annotations as prior information. By leveraging these annotations, likely causal variants can be prioritized based on predicted function as well as strength of association between the genetic variant and the phenotype. These methods have increased accuracy and resolution of resulting credible sets of SNPs, decreasing the burden of any further experimental validation of true casual variants (Kichaev and Pasaniuc, 2015).

Lastly, given GWAS identified loci are enriched for variants affecting gene expression (Nicolae et al., 2010), there have been many methods developed to link non-coding disease variants to genes by leveraging QTL data. The first class of methods, such as eCAVIAR (Hormozdiari et al., 2016) and coloc (Giambartolomei et al., 2014), look for colocalization of association signals, estimating shared casual variants between the molecular QTL and GWAS loci. More recently, studies have leveraged predicted expression, made possible by large eQTL studies, to perform association testing between disease status of GWAS and imputed cis-

regulated gene expression through methods such as TWAS (Gusev et al., 2016), PrediXcan (Gamazon et al., 2015), and summary-data-based Mendelian Randomization (SMR) (Zhu et al., 2016). By limiting the search space from millions of SNPs to thousands of genes, TWAS methods leverage power to detect disease associations along with the ability to aggregate signal to identify genes in regions that did not pass genome-wide significance in the original GWAS (Pavlidis et al., 2016).

While these gene identification methods, which identify disease risk converging at the gene-level, have made substantial advances to the field, it is common for large discrepancies to exist among results of different methodologies with the same input GWAS and eQTL datasets (Liu et al., 2018), and even more so when eQTL datasets are used from different tissues (Gusev et al., 2018). This ambiguity of computationally implicated genes emphasizes the need for experimental validation of candidate genes and variants (Gallagher and Chen-Plotkin, 2018). However, TWAS implicated genes from disease relevant tissues and fine-mapped credible SNPs would be a great place to start.

## **1.7: Conclusions**

In this chapter I covered the current state of neuropsychiatric genetics and transcriptomics, focusing on ASD and SCZ. I describe the efforts the field has made to functionally annotate the non-coding genome, which has shown significant tissues and cell-type specificity and incomplete especially in brain tissues. I introduced methodology used to combine known disease risk with functional data to further dissect the genetic architecture of these disease and implicate risk genes, and set up the motivation for conducting expression and splicing QTL studies specifically during fetal brain development in order to interpret disease loci.

In Chapters 2 and 3, I performed a well powered expression QTL and splice QTL analyses in human cortex at mid-gestation to understand how functional genetic variation impacts disorders characterized by phenotypes that likely originate in utero or early postnatal life, such as neurodevelopmental and neuropsychiatric diseases (Hannon et al., 2016; Jaffe et al., 2016; Parikshak et al., 2015; Weinberger, 1987). I focused on mid-gestation, an epoch that captures crucial stages of neural progenitor proliferation and neurogenesis (Geschwind and Rakic, 2013; Johnson et al., 2009; Silbereis et al., 2016). By combining these data with transcription and splicing factor binding, I identify candidate transcriptional and splicing regulatory networks whose activity is predicated to be impacted by human common genetic variation.

In Chapter 4, I contrast the genetic control of fetal and adult expression and splicing, which as expected, do show substantial differences. I also profiled the relationship between expression time points and disease risk. Similarly, I find that both eQTL and sQTL contribute substantially to disease risk, but show mostly non-overlapping patterns.

In Chapter 5, I performed weighted co-expression network analysis and identify modules relating to specific fetal cell types and developmental biological processes. I use eQTL mapped regulatory regions to show common GWAS risk variants for ASD and SCZ coalesce at the module level and compare module enrichment of genes implicated by rare variant studies, which also converge.

In Chapter 6, I further integrate eQTL and sQTL with GWAS via transcriptome wide association (Gusev et al., 2016) and identified new putative mechanisms undetected in adult brain data sets. I describe the creation of both adult brain expression weights from PsychENCODE and fetal brain expression weights, as well as compared and contrast implicated genes from both time points when performing TWAS for SCZ, ICDV, and ASD. These data provide insights into genes

and isoforms relevant to developmental risk factors, and establishing which aspects of disease risk affect early corticogenesis, as compared to later developmental processes (**Figure 1.1**).

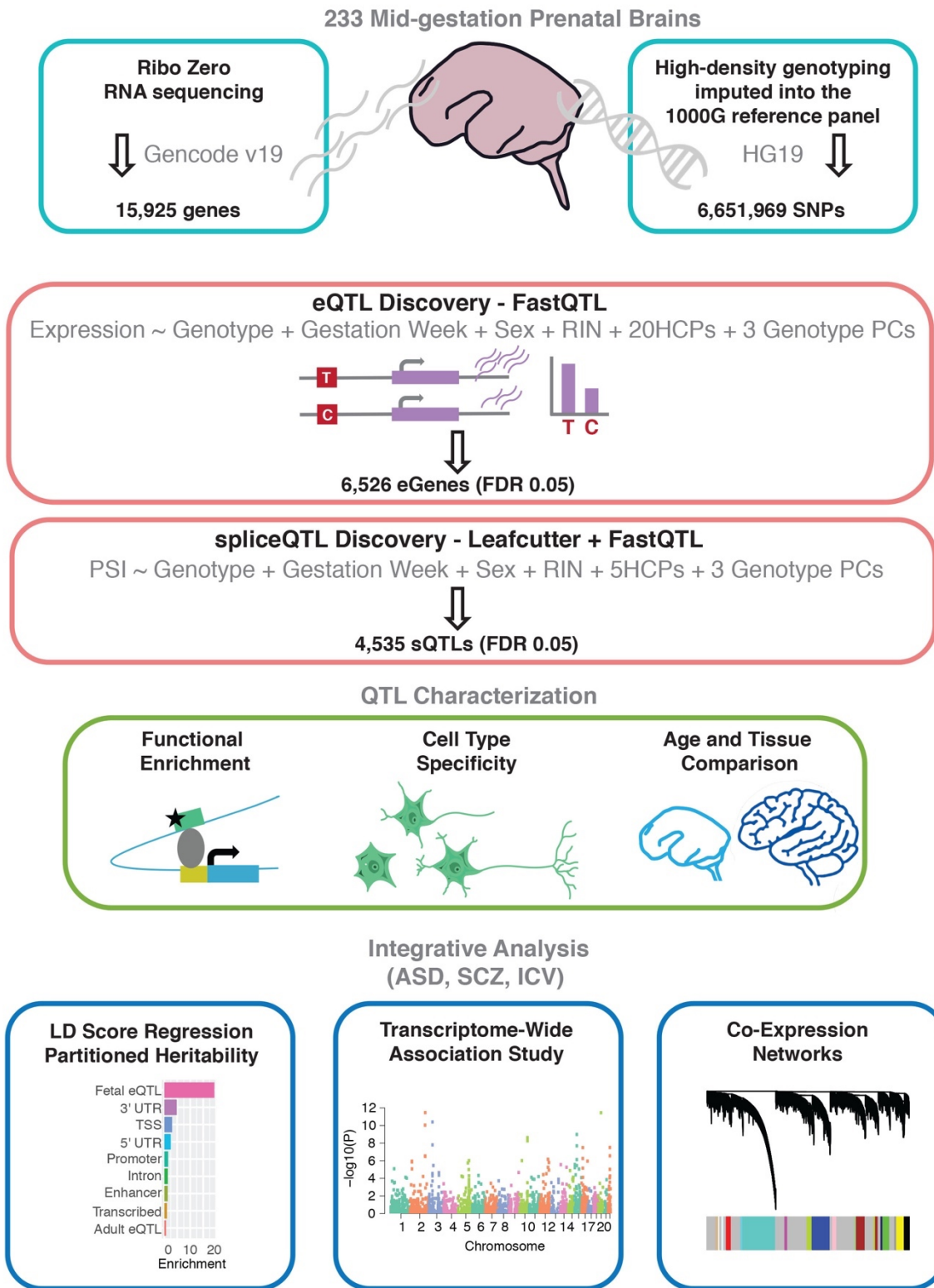


**Figure 1.1: Study Design and Methodological Overview of QTL Analysis.**

RNA sequencing data from 233 fetal cortical samples was exhaustively QC'ed leaving data from 201 samples that was integrated with genome-wide imputation of common genetic variation

based on high-density genotyping for expression and splicing QTL discovery. eQTL and sQTL were independently calculated using standard methods of covariate correction (Mostafavi et al., 2013). QTLs were characterized based on functional enrichment, cell type specificity, and compared to mature brain and non-brain tissue QTLs. I integrated these fetal brain QTLs with GWAS of neuropsychiatric disorders, as well as other human brain phenotypes, such as intracranial volume and educational attainment, performing LDSR, TWAS, and gene co-expression network analysis to identify developmental disease risk and important biological processes under genetic control.

## **CHAPTER TWO**

Robust identification of prenatal brain eQTL

## **2.1: Introduction**

Tissue specific regulatory regions harbor substantial genetic risk for disease. Since brain development is a critical epoch for neuropsychiatric disease susceptibility, and all large brain eQTLs have been performed in adult post-mortem samples, there is a critical need to assess eQTLs in developing brain. Here, I have performed the largest eQTL analysis in the developing human cortex to date, identify genetic loci that regulate gene expression during this critical time point. I describe thorough quality control of RNA-seq and genotype data, as well as robust identification of eQTLs using multiple methodology to correct for population stratification, as these sample come from diverse ancestries. I then characterized genomic properties of eQTLs as well as enrichment in orthogonal data, such as fetal brain Hi-C defined chromosome looping interactions, fetal brain ATAC-seq defined open regions, and fetal brain ChromHMM chromatin state predictions. Lastly, to explore potential mechanisms by which eQTLs may be influencing gene expression, I examined whether eQTLs are enriched in transcription factor (TF) and DNA binding protein (DBP) binding sites, and use experimental data to implicate CHD8's role in regulating *SRR* expression.

## **2.2: Methods**

### *2.2.a: Data processing*

Fetal tissue was obtained from the UCLA Gene and Cell Therapy core according to IRB guidelines from 233 donors (post-conception weeks: 14-21) following voluntary termination of pregnancy. This study was performed under the auspices of the UCLA Office of Human Research Protection, which determined that it was exempt because samples are anonymous pathological specimens. Full informed consent was obtained from all of the parent donors.

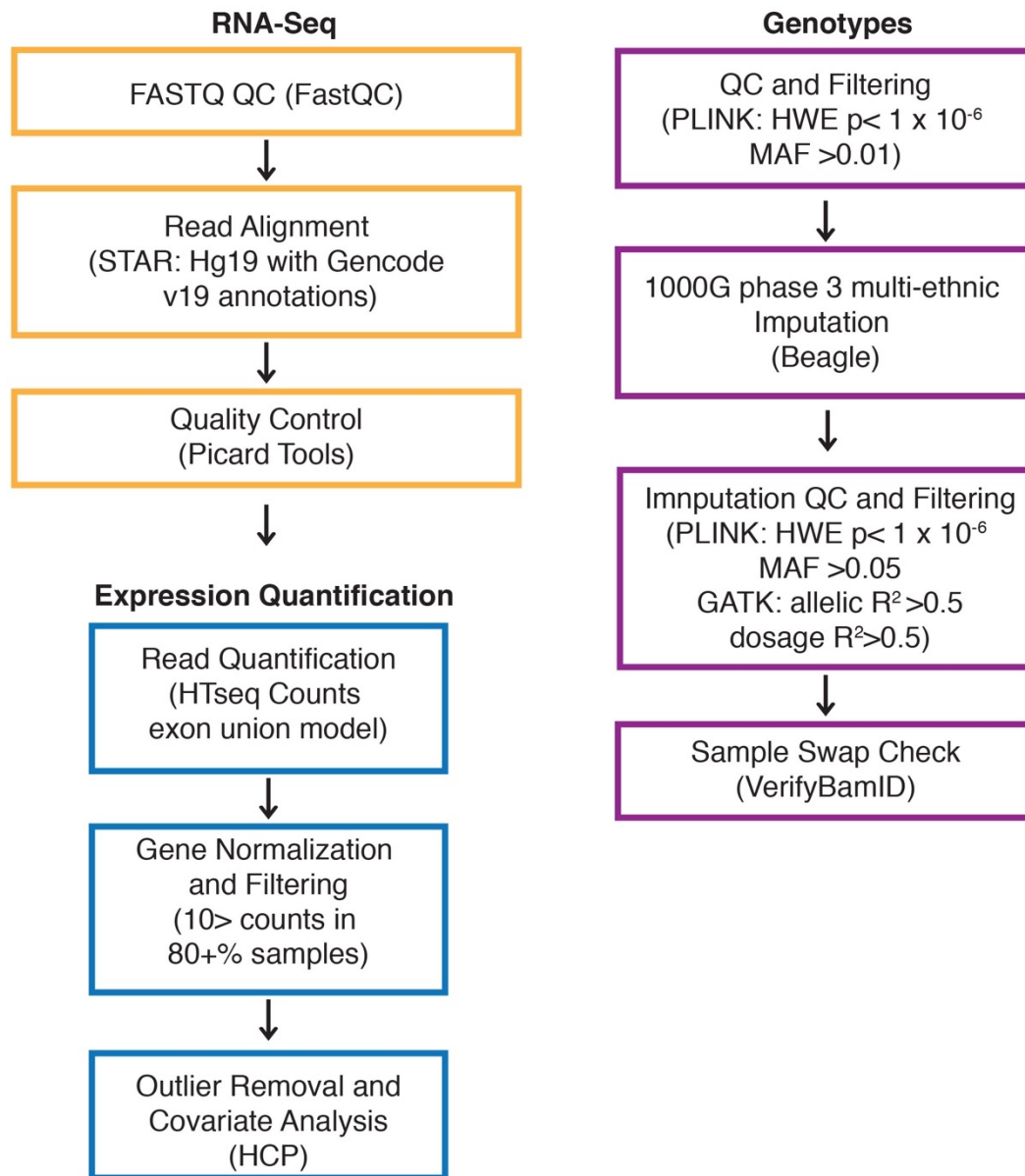


**Figure 2.S1: Overview of raw data processing and QC**

Overview of methods and QC pipeline for processing RNA sequencing FASTQ files to gene quantifications and genotype imputation, QC, and filtering. Output data from this pipeline was used as inputs for further eQTL analysis.

Total RNA and genomic DNA from human fetal brain tissue from PCW 14-21 that visually appeared to be cortical was extracted using miRNeasy-mini (Qiagen) and DNeasy Blood and Tissue Kit (DNA) or were extracted using trizol with glycogen followed by column purification. Library preparation via Illumina Stranded TruSeq kit with Ribozero Gold ribosomal RNA depletion library prep was followed by sequencing on 233 brains and genotype array data on 212 brains was generated at the UCLA Neurogenomics Core. Pseudo-randomization to decrease correlation between sequencing lane and biological variables like sex and gestation week were performed. RNA samples were pooled, randomized, and run on 4 lanes. Ribozero, ribosome depleted, 50 bp paired-end RNA sequencing was performed with mean sequencing depth of 60 million reads on an Illumina HiSeq2500.

### *2.2.a.i Genotyping*

Genotyping was performed at the UCLA Neurogenomics Core (UNGC) on either Illumina HumanOmni2.5 or HumanOmni2.5Exome platform in 8 batches. SNP genotypes were exported into PLINK format. Batches were merged and markers that did not overlap genotyping platforms were removed. SNP marker names were converted from Illumina KGP IDs to rsIDs using the conversion file provided by Illumina. Quality control was performed in PLINK v1.9 (Chang et al., 2015). SNPs were filtered based on Hardy-Weinberg equilibrium ( --hwe 1e6), minor allele frequency ( --maf 0.01), individual missing genotype rate ( --mind 0.10), variant missing genotype rate ( --geno 0.05) resulting in 1,799,583 variants (**Figure 2.S1**).

PLINK genotype files were converted to vcf files using Plinkseq v0.10 (<https://www.atgu.mgh.harvard.edu/plinkseq>). Genotypes were imputed into the 1000 Genomes Project phase 3 multi-ethnic reference panel (The 1000 Genomes Project Consortium et al., 2015) by chromosome using Beagle v4.1 (Browning and Browning, 2016) and subsequently merged. Multiallelic sites were removed using GATK v3.5 (Van der Auwera et al., 2013).

Imputed genotypes were filtered for Hardy-Weinberg equilibrium  $p$ -value  $< 1 \times 10^{-6}$  and minor allele frequency (MAF) 5%. Imputation quality was assessed filtering variants where allelic  $R$ -squared  $> 0.5$  and dosage  $R$ -squared  $> 0.5$  by GATK, resulting in ~6.6 million autosomal SNPs. I restricted to only autosomal due to sex chromosome dosage, as commonly done (The GTEx Consortium, 2015).

### *2.2.a.ii: Gene expression quantification*

All raw RNAseq fastq files, 4 per sample run on different lanes, were run through FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). FastQC output was visually inspected and sequencing lanes where the “per tile sequence quality” was red were removed, there was no sample with more than one sequencing lane removed. Fastq files were aligned to the GRCh37.p13 (hg19; Homo\_sapiens.GRCh37.75.dna.primary\_assembly.fa: [ftp://ftp.ensembl.org/pub/release-75/fasta/homo\\_sapiens/dna/](ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/)) reference genome using STAR v2.4 (Dobin et al., 2013). SAM files were sorted, indexed, converted to BAM files and merged across lanes from the same sample using Samtools v1.2 (Li et al., 2009). Gene quantifications were calculated using HTSeq-counts v0.6.0 (Anders et al., 2015) using an exon union model on the basis of Gencode v19 comprehensive gene annotations (Harrow et al., 2012). Quality control metrics were calculated using PicardTools v1.139 (<http://broadinstitute.github.io/picard>) and Samtools. A sex incompatibility check was also performed using XIST expression and Y chromosome non-pseudoautosomal expression which are known to show different patterns of expression in males and females. A scatter plot of XIST expression versus the sum of expression of genes in the non-pseudoautosomal region of the Y chromosome showed no gender mismatches. (**Figure 2.S1**)

Gene counts were compiled from HTSeq Count (Anders et al., 2015) quantifications and imported into R version 3.2.1 for downstream analyses. Gene counts were put through quality control, removing genes that were not expressed in 80% of samples with 10 counts or more.

Expression was then corrected for GC content, gene length, and quantile normalized to a standard normal distribution, as commonly done in QTL studies (Battle et al., 2015). Sample outliers were removed based on standardized sample network connectivity Z scores  $< 2$  (Zhang and Horvath, 2005) (**Figure 2.S2 F**). ComBat batch correction was performed (Johnson et al., 2007). After quality control and normalization, there remained 201 samples with 15,930 genes expressed (on the basis of Gencode v19 annotations) at sufficient levels.

### *2.2.a.iii: Sample Swap Identification*

QC'd genotypes and sample BAM files were used to identify any sample identity swaps between the RNA and DNA experiments using VerifyBamID v1.1.2 (Jun et al., 2012). I identified 4 samples where [CHIPMIX]  $\sim 1$  AND [FREEMIX]  $\sim 0$ , indicative of nonmatching RNA and DNA, which were removed in the VCF file.

### *2.2.b: Covariate Selection*

To evaluate and remove global effects of gene expression, I used PLINK v1.9 (Chang et al., 2015) to run multidimensional scaling on the QC'd imputed genotypes and to verify ancestral backgrounds of the samples. I aggregated the final 201 samples with HapMap3 (International HapMap, 2003) of 1397 samples across 11 populations (87 ASW, 165 CEU, 137 CHB, 109 CHD, 101 GIH, 113 JPT, 110 LWK, 86 MXL, 184 MKK, TSI 102, YRI 203). A plot of the first two MDS components of the merged data shows the genetic ancestry of our samples among a diverse reference population (**Figure 2.S2 H**). For eQTL analysis, the top 3 MDS components calculated only in the fetal brain samples were used as covariates.

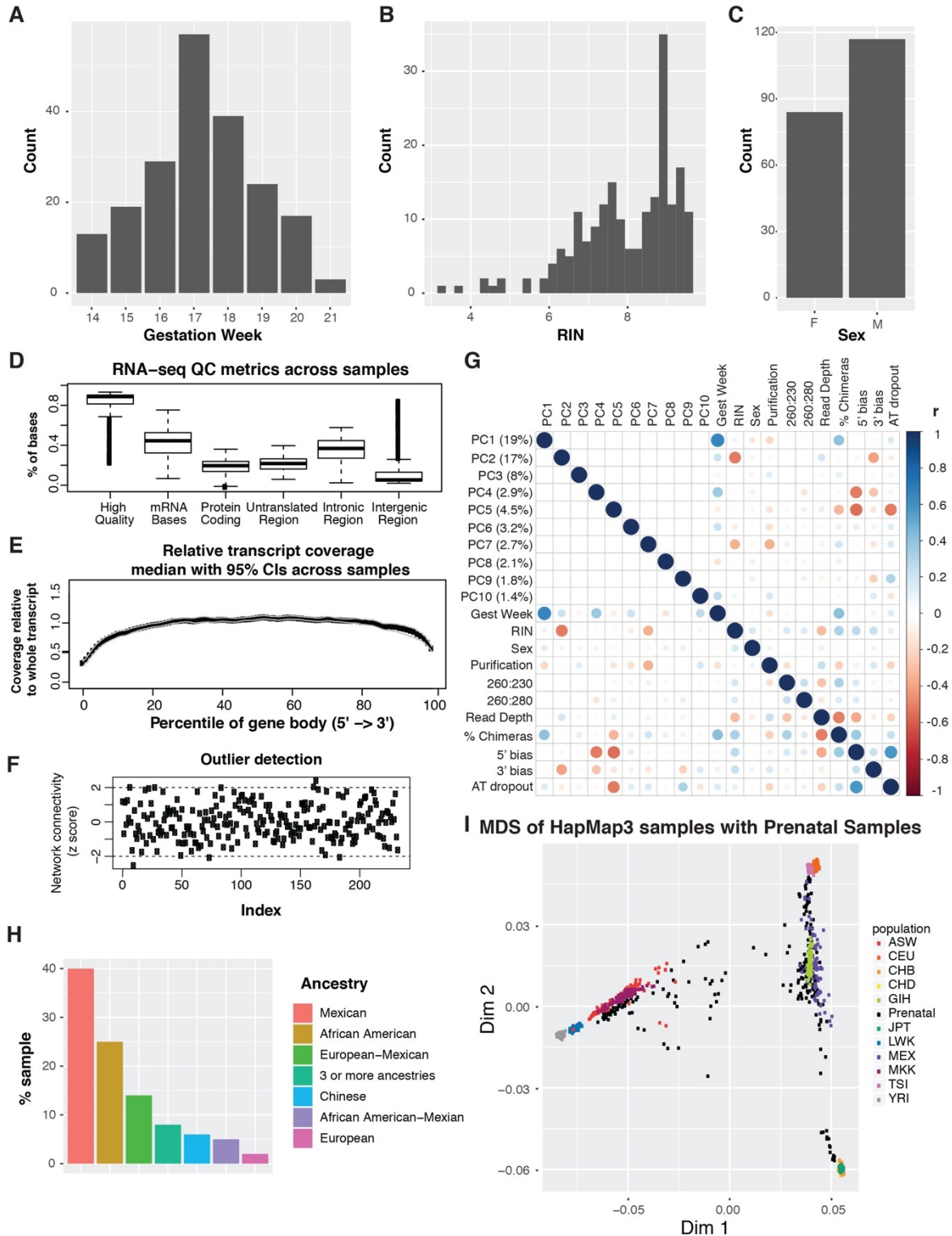


**Figure 2.S2: Sample demographics and dataset quality control**

**(A)** Distribution of age (gestation week) of our samples, showing our samples span mid-gestation. **(B)** Distribution of RNA integrity number (RIN) among our samples. **(C)** Distribution of sex among our samples. **(D)** RNA sequencing metrics from Picard tools showing the majority of

our reads are of high quality and correspond to mRNA. **(E)** Relative transcript coverage across the body of a gene, showing good coverage along the entire gene body, representative of using a ribo-zero library preparation. **(F)** Sample outlier detection determined from WGCNA's network connectivity Z-score. Samples with a Z-score of greater than 2 or less than -2 are removed. **(G)** Correlation of the top 10 expression principle components (PCs) with gestation week, RIN, sex, Picard tool metrics: read depth, % chimeras, 5' bias, 3' bias, AT dropout, and technical measurements: purification method and purity assessments 260:230 and 260:280. Gestation Week and RIN show high correlations with the top 2 PCs. **(H)** Distribution of ancestry among our samples. **(I)** MDS plot of genotypes from our fetal brain samples merged with the HapMap3 samples from 11 populations, shows the diversity across the fetal samples.

I also assessed the correlation of known measured biological covariates, measured technical covariates, as well as RNA quality control metrics from Picard tools (gestation week, RIN, sex, purification method, 260:230 ratio, 260:280 ratio, read depth, percent chimeras, 5' bias, 3' bias, AT dropout) with the top 10 principle components of the expression data and find the top principal component corresponds to the age of the sample (gestation week) and the second component corresponds to the RNA integrity number (RIN) (**Figure 2.S2 G**).

To measure hidden batch effects and confounders, hidden covariate analysis was performed using Hidden Covariates with a Prior (HCP) (Mostafavi et al., 2013). Hidden factors were calculated given the known measured factors. HCP was run separately for varying number of inferred hidden components: 5, 10, 15, 20, 25, 30. I included 20 HCPs in our eQTL model (which I found to maximized eGene discover), along with gestation week, RIN, and sex (**Figure 2.S3 A**). I correlated the 20 HCPs along with gestation week, RIN, sex, and top 3 genotype PCs (all covariates used in the final model) with the measured factors and Picard metrics, as well as the top 20 PCs of expression to gain insight to meaning of the HCPs. I see each inferred hidden component's relationship to the known variables is complex and distributed across variables.

### 2.2.c: eQTL analysis

I performed cis-eQTL mapping using FastQTL (Ongen et al., 2016), a defined cis window of 1 megabase up- and down-stream of the transcription start site for 15,930 expressed genes, and correction for the following covariates: gestation week, RIN, sex, 20 HCPs, 3 genotype PCs. FastQTL (Ongen et al., 2016) was run in the permutation pass mode (1000 permutations) to identify the best nominal associated SNP per phenotype and with a beta approximation to model the permutation outcome (**Figure 2.S3 C**) and correct for all SNPs in LD with the top SNP per phenotype. Beta approximated permutation p-values were then multiple test corrected using the q-value Storey and Tibshirani FDR correction (Storey and Tibshirani, 2003). I define eQTL containing genes (eGenes) by having an FDR q-value  $\leq 0.05$ . Secondary, independent eQTLs were identified by rerunning permutation tests in FastQTL for every eGene conditioning on the primary eSNP.

To assess inflation in the Q-Q plot, I randomly chose 10 genes to run eQTL analysis in trans, testing all SNPs genome-wide with association with gene expression using MatrixEQTL (Shabalin, 2012). I corrected for the same covariates as in the cis-eQTL analysis. The Q-Q plot of the trans-eQTLs shows no inflation, an indication that our eQTL results are not confounded by population stratification. (**Figure 2.S3 D**)

To validate a few eQTLs, I selected eGenes with the highest effect sizes in order to be able to detect a difference by qPCR and stratify by genotype, in which I find concordant significant eQTL signal. qPCR was performed on 5 samples of genotype 0 and 5 samples of genotype either 1 or 2 for each gene. (**Figure 2.S3 E,F,G**)

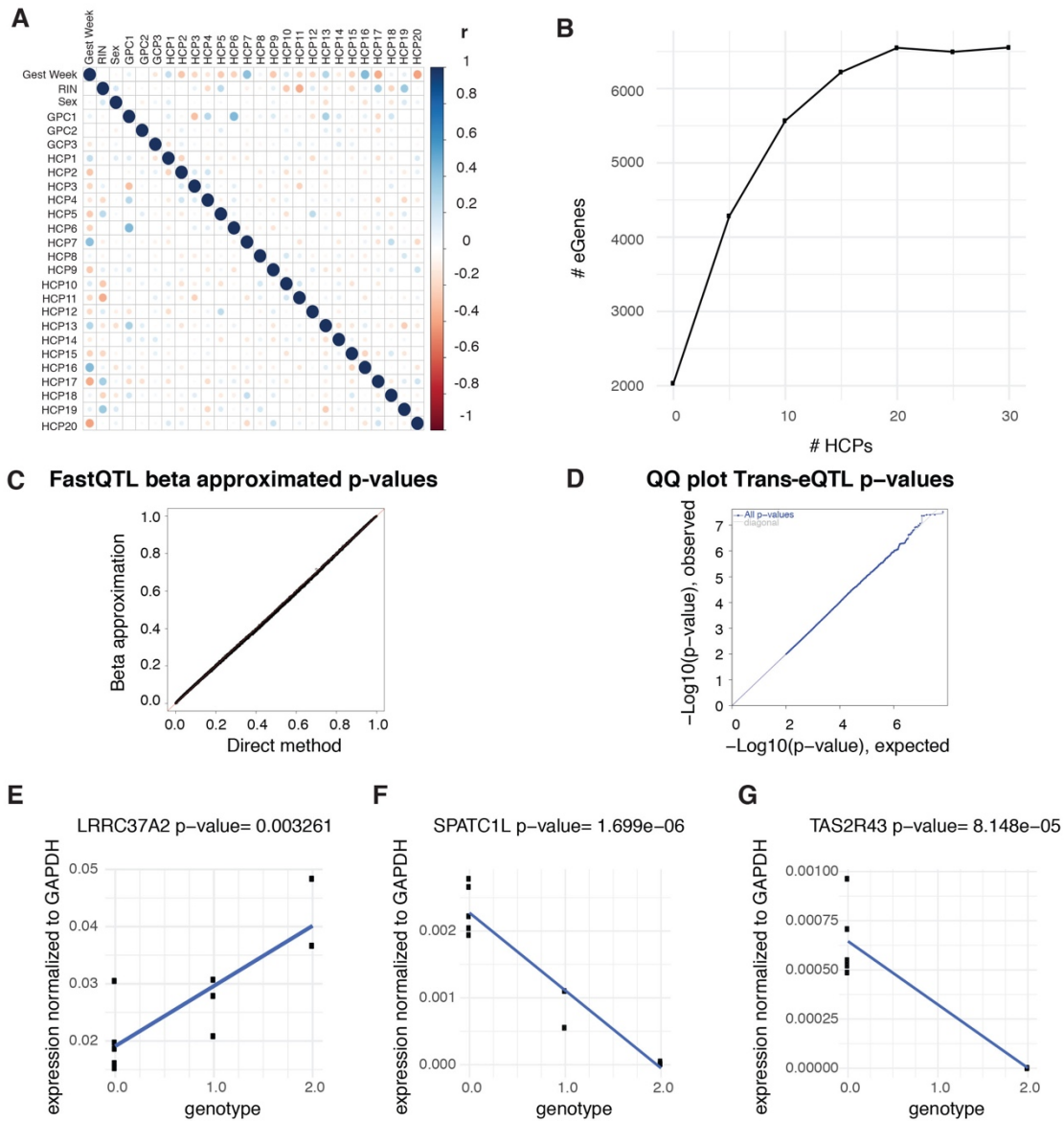


**Figure 2.S3: Identification of optimal number of HCP selection and correlations of all covariates in final cis-eQTL analysis**

**(A)** Pearson correlation ( $r$ ) between all covariates used in the final eQTL analysis. **(B)** The number of hidden covariates (HCP) was chosen to maximize cis-eGene discovery. Cis-eQTL analysis was performed correcting for gestation week, RIN, sex, the top 3 genotype PCs and then increments of 5 HCPs using FastQTLs permutation scheme. The number of HCPs included in the eQTL analysis is shown on the X axis with the number of significant eGenes is shown on the Y axis. The number of eGenes was determined by FDR q-value  $\leq 0.05$ . Based on these results, 20 HCPs were selected for the final eQTL analysis. **(C)** FastQTL implements a beta approximation for permutation p-values. Empirical p-values are on the X axis with beta approximated p-values are on the Y axis. This shows the beta approximated p-values are well calibrated given the empirical p-values. **(D)** QQ plot for trans-eQTL analysis that was run for 10

randomly chosen genes, expected p-values on the X axis and observed p-values on the Y axis, indicating there is no systematic inflation. **(E,F,G)** qPCR validation of three eQTLs of high effect size for genes *LRRC37A*, *SPATC1L*, and *TAS2P43*. For each eQTL, the genotype of samples is shown on the X axis, where there are 5 samples with the genotype of 0 and 5 samples of genotype 1 or 2 for each gene. Expression normalized to *GAPDH* is shown on the Y axis.

#### *2.2.d: Correcting for ancestral differences*

To further check that our eQTLs are not due to the population differences of our samples, I ran cis-eQTL analysis with EMMAX (Kang et al., 2008), which accounts for population structure using a genetic relationship matrix. I used the emmax-kin function (-v -h -s -d 10) to create the IBS kinship matrix. EMMAX was run for each gene with a cis-window of +/- 1MB around the TSS, correcting for the same covariates in the FastQTL analysis. Nominal EMMAX p-values were corrected for multiple testing using the q-value Storey and Tibshirani FDR correction (Storey and Tibshirani, 2003).

To assess overlap between FastQTL and EMMAX, FastQTL was also run in the nominal pass mode to obtain nominal p-values for all cis-SNPs tested per gene. FastQTL nominal p-values were also corrected for multiple testing using the q-value Storey and Tibshirani FDR correction. I discover 920,356 nominal eQTLs at a 5% FDR threshold. To compare nominal results, I defined eGenes as a gene containing a significant SNP association at  $FDR \leq 0.05$ . I found 93.8% of eGenes from the FastQTL analysis was an eGene in the EMMAX analysis. Additionally, I compared all SNP-gene pairs tested and found 92.8% of significant SNP-gene associations from nominal FastQTL to be significant in the EMMAX analysis (**Figure 2.S4 D,E**).

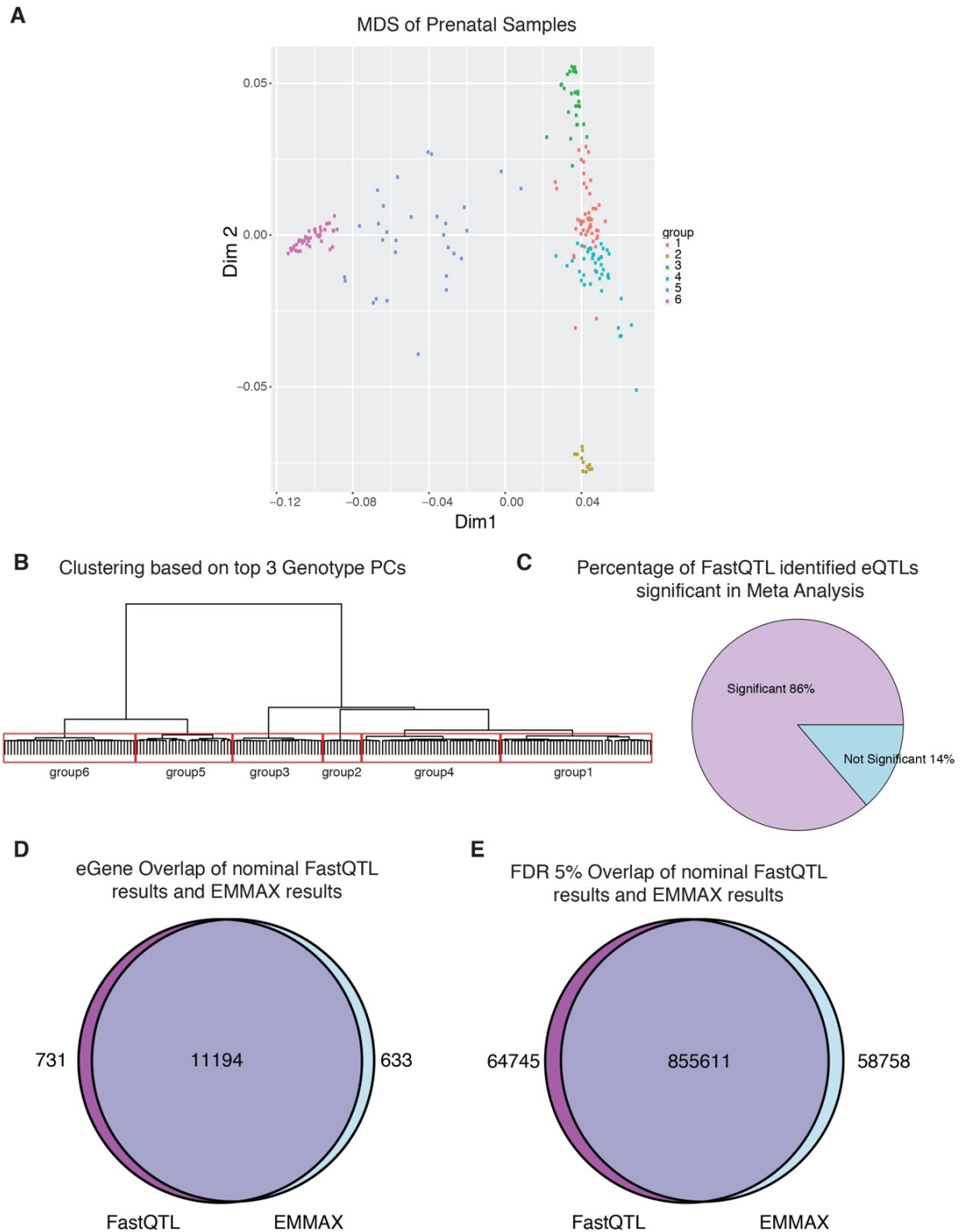


**Figure 2.S4: Ancestry correction**

(A) MDS plot of genotypes from just prenatal brain samples colored by group assignments from (B). (B) Sample clustering based on hierarchical clustering of the top 3 genotype PCs groups

prenatal brain samples into 6 subpopulation groups. **(C)** Meta-Analysis of eQTLs (top SNP per gene) performed across the 6 subpopulation groups finds 86% of eQTLs significant. **(D)** 94% of eGenes in FastQTLs nominal pass overlap significant eGenes from EMMAX. **(E)** 93% of significant (FDR  $\leq 0.05$ ) SNP-gene pairs from the nominal FastQTL pass overlap with significant SNP-gene pairs from the EMMAX eQTL analysis.

I divided our sample into six groups based on hierarchical clustering of the top 3 PCs of the genotype data (**Figure 2.S4 A,B**). The size of each group ranged from 12 samples to 47 samples and each group corresponded to distinct ancestries based on the MDS plots of samples merged with HapMap3. I performed association testing between the top SNP per gene, identified by FastQTL permutation pass eQTL analysis, within group using the `lm()` function in R, correcting for gestation week, RIN, age, and 20HCPs. A fixed effect meta-analysis was then run between groups using METAL v3.25.2011 (Willer et al., 2010), which implements a Cochran's Q test for heterogeneity. I find significant heterogeneity at 10% of our eQTLs and find 87% of our eQTLs are significant in the meta-analysis at an FDR < 0.05% (q-value) strongly suggesting that our results are not due to population stratification (**Figure 2.S4 C**).

#### *2.2.e: Enrichment in genome-wide annotations*

Fetal brain ATAC-seq peaks were obtained from (de la Torre-Ubieta et al., 2018). I annotated eQTLs as being supported by ATAC if the LD block ( $r^2 > 0.8$  PLINK) around its eSNP overlapped an open chromatin region. To test for significance, I created a null set of eQTLs (q-value > 0.2), annotated overlap with ATAC peaks, and then ran a Fisher's exact test.

Fetal brain CP and GZ Hi-C topological association domain bed files were obtained from (Won et al., 2016). eSNPs located within 10kb of the eGene TSS were removed, as Hi-C cannot detect any chromosomal interaction less than 10kb apart. I defined any remaining eQTL as

overlapping Hi-C if the LD block ( $r^2 > 0.8$  PLINK) around its eSNP fell in one 10kb TAD bins and the corresponding eGene  $\pm 2$ kb overlapped with the other 10kb TAD bin in either CP or GZ. To test for significance, I created a null set of eQTLs (q-value  $> 0.2$ ), annotated overlap with Hi-C, and then ran a Fisher's exact test.

I performed functional enrichment of eQTLs using GREGOR (Genomic Regulatory Elements and Gwas Overlap algoRithm) (Schmidt et al., 2015) to evaluate enrichment of variants in genome-wide annotations using the 25 state ChromHMM model BED files from the Roadmap Epigenetics Project (Ernst and Kellis, 2015; Roadmap Epigenomics Consortium et al., 2015), generated from a set of 5 core chromatin marks assayed in fetal brain. GREGOR evaluates the enrichment of QTL variants in these genomic annotations by estimating the significance of observed overlap of the eSNP relative to the expected overlap using a set of matched control variants. GREGOR creates a list of possible causal SNPs by extending the list of eSNPs (index SNPs) to all SNPs in high linkage disequilibrium ( $r^2 > 0.7$ ). A set of matched control SNPs (SNPs are selected based on matching the index SNP for number of variants in LD, minor allele frequency, and distance to nearest gene/intron) is then created, and enrichments are calculated based on the observed and expected overlap within each annotation.

### *2.2.f: Characterization of regulatory drivers*

I downloaded consensus transcription factor and DNA-binding protein binding site BED files (Arbiza et al., 2013), which called consensus binding sites from multiple cell types from Encode CHIP-seq data which was used to computationally annotate all possible genome-wide sites for 78 binding proteins. I filtered to 62 binding proteins that showed cortical brain expression in BrainSpan (BrainSpan, 2013). I used GREGOR to evaluate the enrichment of eSNPs in the transcription factor binding sites.

I obtained CHD8 CHIP-seq peaks performed in human mid-fetal brain (Cotney et al., 2015) and expression counts for CHD8 knockdown in neural progenitors (Sugathan et al., 2014). CHIP-seq peaks were overlapped with fetal eQTLs using the GenomicRanges package in R. eGenes were identified in the knockdown and control expression data and plotted using ggplot.

### 2.3: Results

To identify genetic variants regulating gene expression in the developing brain, I performed high-throughput RNA sequencing and high-density genotyping at 2.5 million sites in a set of 233 fetal brains. After stringent quality control and normalization of gene expression quantifications and genotype imputation into the 1000 Genomes Project phase 3 multi-ethnic reference panel (Methods, **Figure 2.S1**; (The 1000 Genomes Project Consortium et al., 2015), I obtained a starting dataset of 15,925 expressed genes (gene biotypes include 12,943 protein coding, 767 long noncoding RNAs, among others, see **Table 2.S1** for complete listing) and 6.6 million autosomal single nucleotide polymorphisms (SNPs) from 201 individuals. PCA-based (principle component analysis) analysis of ancestry (Methods) indicate that the donors in our study come from admixed ancestries of 40% Mexican 25% African American, 14% European-Mexican, 8% admixed of 3 or more ancestries, 6% Chinese, 5% African American- Mexican, 2% European descent (**Figure 2.S2**). The resulting dataset is the first population-level fetal brain expression and splicing dataset.

I identified cis-eQTLs by testing all SNPs within a 1MB window from the transcription start site (TSS) of each gene using a permutation procedure implemented in FastQTL (Ongen et al., 2016), while adjusting for known (RIN, sex, age, and genotype PCs) and inferred covariates (Methods, **Figure 2.S3**), which have been shown to greatly increase sensitivity for cis-eQTL

detection (Kang et al., 2008; Leek and Storey, 2007; Mostafavi et al., 2013). I identified 6,546 genes with a cis-eQTL at a 5% false discovery rate (FDR), hereafter referred to as eGenes, 82% of which corresponded to protein coding genes (**Table 2.S1**). To compare with studies that did not perform a permutation procedure, we tested all SNP-gene pairs and discover 893,813 significant eQTLs at a 5% FDR corresponding to 11,625 eGenes. To identify additional, independent cis-eQTLs for each eGene, I conducted the analysis while conditioning on the primary eSNP, identifying an additional 1,416 secondary eQTLs for a total number of 7,962 eQTL identified, which represents nearly 45% of expressed genes (Methods). To ensure that the eQTLs identified are not driven by ancestry differences within our dataset, I ran eQTL analysis using two alternative methods: EMMAX, which controls for population stratification using a genetic relationship matrix (Kang et al., 2008) and METAL, which conducts a meta-analysis of the top SNP per gene across subpopulations of the samples (Willer et al., 2010). Both methods demonstrated high reproducibility of eQTLs identified by FastQTL (94% of eGenes detected by FastQTL were also detected by EMMAX and 84% of eQTLs were still significant when a subpopulation meta-analysis was performed despite substantially reduced power in the subpopulation analyses; see **Figure 2.S4**), which indicates that ancestry differences are not driving the eQTL signal (**Figure 2.S4**; Methods).

| Gene Biotype           | Frequency |
|------------------------|-----------|
| 3' overlapping ncRNA   | 5         |
| Antisense              | 722       |
| IG V pseudogene        | 1         |
| lincRNA                | 767       |
| miRNA                  | 48        |
| miscRNA                | 60        |
| Polymorphic pseudogene | 8         |
| Processed transcript   | 148       |
| Protein coding         | 12,942    |
| Pseudogene             | 802       |
| Sense intronic         | 246       |
| Sense overlapping      | 75        |
| snoRNA                 | 71        |
| snRNA                  | 30        |

**Table 2.S1: Expressed gene biotypes**

Frequency of gene biotypes of fetal brain expressed genes

Positional enrichment analyses of the most significant SNP (eSNP) at each eQTL (primary and secondary) shows that over 20% of significant eQTLs cluster within 10kb of the TSS of its target gene (**Figure 2.1 B and 2.1 C**), concordant with previous studies showing that promoter variants have a large influence on cognate gene expression (Kim et al., 2014; Stranger et al., 2007; Strunz et al., 2018). The overall distribution of eQTL signal is consistent with previous work (The GTEx Consortium, 2015; Veyrieras et al., 2008), with 70% of significant eSNPs located within 100kb of the TSS, as well as a slight upstream bias (~56%) of significant eQTLs (**Figure 2.1 B**).

I reasoned that eQTL should be highly overlapping with putative regulatory regions defined by other methods and sought to test this as an external validation (**Figure 2.1 A**). Indeed, I find eQTLs are significantly enriched within regions of open chromatin identified by ATAC-seq from developing brain ( $OR=4.42$ ,  $p\text{-value} < 2.2 \times 10^{-16}$ ) (**Figure 2.1 D**; (de la Torre-Ubieta et al., 2018). Distal eQTLs ( $> 10Kb$  from TSS) are also enriched within 3D chromatin conformation contacts detected at the same stage of brain development ( $OR=2.96$ ,  $p\text{-value} < 2.2 \times 10^{-16}$ ) (**Figure 2.1 E**; (Won et al., 2016) strengthening confidence that eSNPs are in accessible regions where transcription factors preferentially bind and distal eSNPs regulate their associated target eGenes (Methods).

To further characterize the identified eQTLs, for each eGene, I annotated the most significant variant with chromatin state predictions from fetal brain tissue from the Roadmap Epigenetics Consortium (Roadmap Epigenomics Consortium et al., 2015), using GREGOR (Schmidt et al., 2015) Methods) to test for enrichment of eQTLs among the 25 chromatin states. As expected, I observed that eSNPs were most significantly enriched in transcription start sites, promoters, and transcribed regulatory promoter or enhancers (**Figure 2.1 F**). Enrichment in these regulatory regions provides further evidence, in addition to the ATAC and Hi-C, that discovered

eQTLs fall in defined regions of the genome functionally relevant to the regulation of gene expression, as would be predicted (Ward and Kellis, 2012).

Since changes in constrained genes expressed in the developing brain are likely to have a higher impact on fecundity than those that are not highly constrained (Samocha et al., 2014), we reasoned that genes perturbed by regulatory variation would also be more tolerant to protein-disrupting variation (Lek et al., 2016). Indeed, comparing genes harboring an eQTL to genes that do not have a significant eQTL, we find that eGenes are more tolerant to loss of function mutations and less constrained (Wilcoxon rank sum test p-value  $< 2.2 \times 10^{-16}$ ; **Figure 2.1 G**). Additionally, fetal eGenes are significantly more intolerant to loss of function mutations than adult eGenes identified in adult cortex in GTEx (Wilcoxon rank sum test p= 0.000194)(The GTEx Consortium et al., 2017).

To understand the mechanisms by which eQTLs may be influencing gene expression, I examined whether genomic regions tagged by these eSNPs are enriched in experimentally-determined transcription factor (TF) and DNA binding protein (DBP) binding sites, based on chromatin immunoprecipitation data from the ENCODE project and Cotney *et al.* (Methods; (Arbiza et al., 2013; Cotney et al., 2015). I found genome-wide significant binding site enrichments within eQTL containing regions for 39 TFs and DBPs (**Figure 2.1 H**), many with prominent known roles in brain development and patterning, including *ELK4*, *NRF1*, *SMARCC1*, *SMARCC2*, and *CHD8* (Bestman et al., 2015; Durak et al., 2016; Eising et al., 2019; Lai et al., 2001; Ojeda et al., 1999; Preciados et al., 2016). Enrichment of eQTLs in these binding sites provides further support that these non-coding regions possess functional qualities compared to the genome background, primarily, but not exclusively, consistent with regulatory enhancer properties (Encode Project Consortium, 2012).

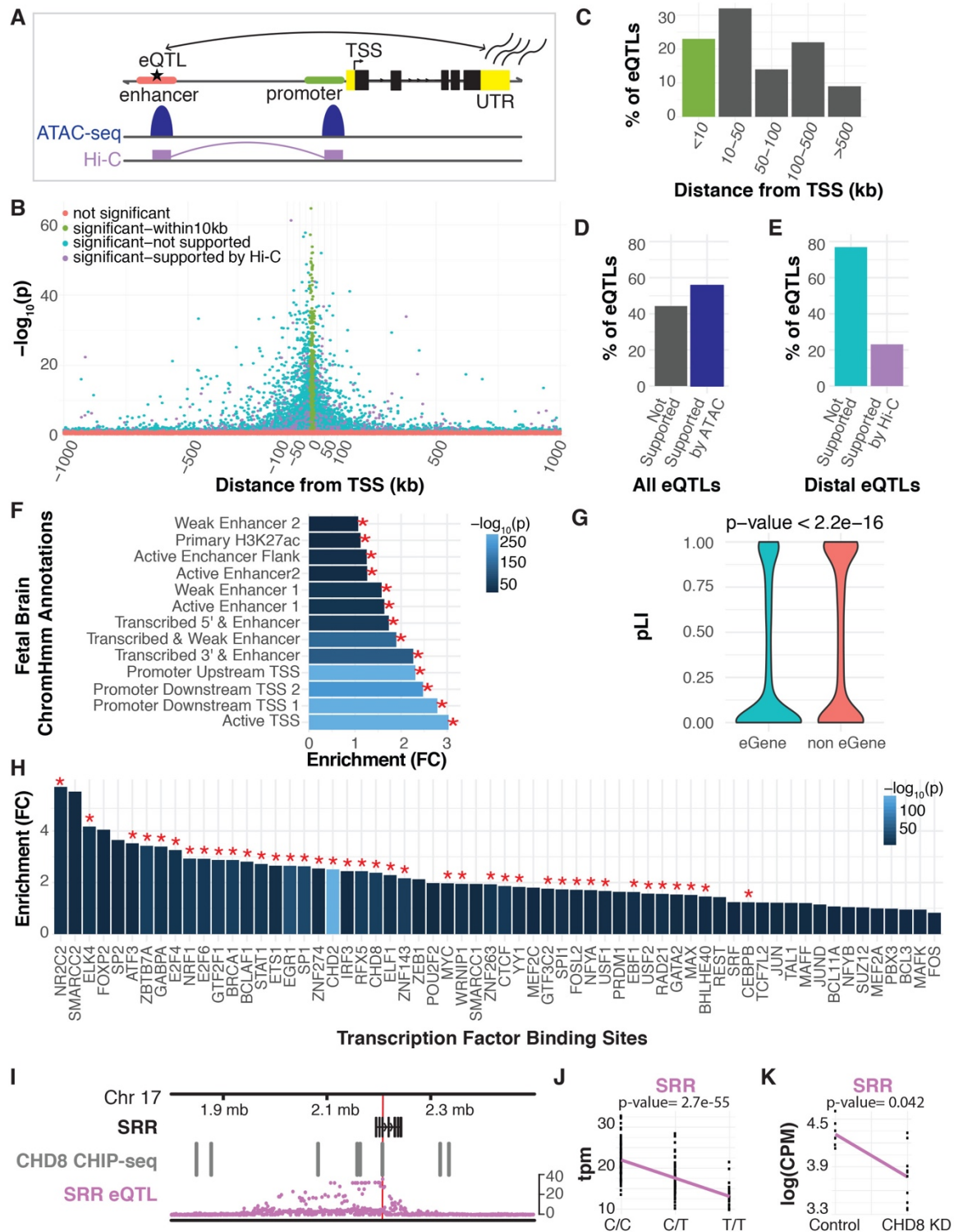


**Figure 2.1: Characterization of eQTLs**

(A) A schematic showing that chromatin accessibility (ATAC-seq) and chromatin interactions (Hi-C) are overlapping and support eQTLs in linking genes to distal regulatory elements. (B) Position of eSNPs in relation to the eGene TSS. Significant eQTLs are annotated based on support from prenatal brain Hi-C and distance from the TSS. (C) Distribution of all primary and secondary eQTLs. eQTLs that lie within 10kb of the TSS are colored green, as depicted in (B).

**(D)** Fraction of eQTLs with prenatal brain ATAC-seq support. **(E)** Fraction of distal eQTLs (>10kb from the TSS) with prenatal brain Hi-C support. eQTLs that show Hi-C support are colored purple, while those not supported by Hi-C are colored blue, as depicted in (B). **(F)** Fold enrichment of eSNPs by their distribution within prenatal brain chromatin states, which are listed on the Y axis. **(G)** eGenes are more tolerant to loss of function, as compared with those genes expressed in prenatal brain without significant eQTL. (Wilcoxon Rank Sum  $P < 2.2 \times 10^{-16}$ ). **(H)** Fold enrichments of eSNPs in TF binding sites (Arbiza et al., 2013). A red asterisk indicates significance at 5% FDR. **(I)** Whole-gene view of *SRR* highlighting the association  $-\log^{10}(P)$  for all cis-SNPs with expression of *SRR*, the eSNP is denoted by the red line. Additionally, the eSNP falls within a *CHD8* CHIP-seq binding site. **(J)** The eQTL for *SRR*, showing the distribution of expression values of *SRR* per genotype. **(K)** The expression of *SRR* in control samples versus *CHD8* knockdown.

As a specific example, I next chose *CHD8* ( $FC = 2.4$ ,  $p\text{-value} = 7.8 \times 10^{-23}$ ), which was of interest because of its strong association with ASD by harboring *de novo* loss of function mutations (Bernier et al., 2014; Neale et al., 2012; O’Roak et al., 2014; Sanders et al., 2012). We obtained CHiP-seq binding from human mid-fetal brain and differential expression after knockdown of *CHD8* in neural progenitors (Cotney et al., 2015; Sugathan et al., 2014). Only a small fraction of eGenes exhibit significant differential expression when *CHD8* expression is reduced in neural progenitors, consistent with published studies (Cusanovich et al., 2014). However, I was able to identify an eSNP falling directly within a *CHD8* CHIP-seq peak for the gene serine-racemase (*SRR*) (**Figure 2.1 I,J**), which has been shown to play a role in glutamatergic neurotransmission and modulates neuropsychiatric phenotypes (Basu et al., 2009). Indeed, *CHD8* knockdown in neural progenitors led to a significant change in *SRR* expression (**Figure 2.1 K**) (Sugathan et al., 2014). Thus, experimental data from fetal brain and neural progenitors *in vitro* supports the mechanistic role of *CHD8* in regulating this gene.

## **CHAPTER THREE**

Robust identification of prenatal brain sQTL

### 3.1: Introduction

Splice QTLs (sQTL) are known to contribute, comparably, if not more so than eQTL to complex disease risk and further the understanding of the link between genetic variation and gene regulation (Li et al., 2016). Since there has been no systematic analysis of splicing regulation in human fetal brain development, such analysis is critical to further the understanding of the link between genetic variation and disease during development. There are multiple approaches for quantifying splicing and therefore sQTLs. These include junction-level quantification methods that calculate the ratio of reads spanning splice junction or percent spliced in index (Li et al., 2018; Shen et al., 2014), transcript level quantification methods that use expectation maximization algorithms to probabilistically assign reads to a given transcript (Bray et al., 2016; Trapnell et al., 2009), to using isoform ratios to perform QTL analysis (Monlong et al., 2014), though all are limited by properties of short read RNA sequencing technology. However, since isoform annotations in the brain are known to be incomplete, I have used splice junction quantifications for sQTL analysis.

### 3.2: Methods

#### 3.2.a: *Splicing quantification*

I used Leafcutter (Li et al., 2018) to leverage information from reads that span introns to quantify clusters of variably spliced introns. From the already aligned FASTQ files by STAR, output bam files were converted into junction files. Intron clustering was performed using default settings of 50 reads per cluster and a maximum intron length of 500kb. The Leafcutter `prepare_genotype_table` script was then used to calculate intron excision ratios and to filter out introns used in less than 40% of individuals with almost no variation. Intron excision ratios were then standardized and quantile normalized (**Figure 3.S1**). Leafcutter's `leafviz` annotation code were used to annotate detected introns as annotated or cryptic.

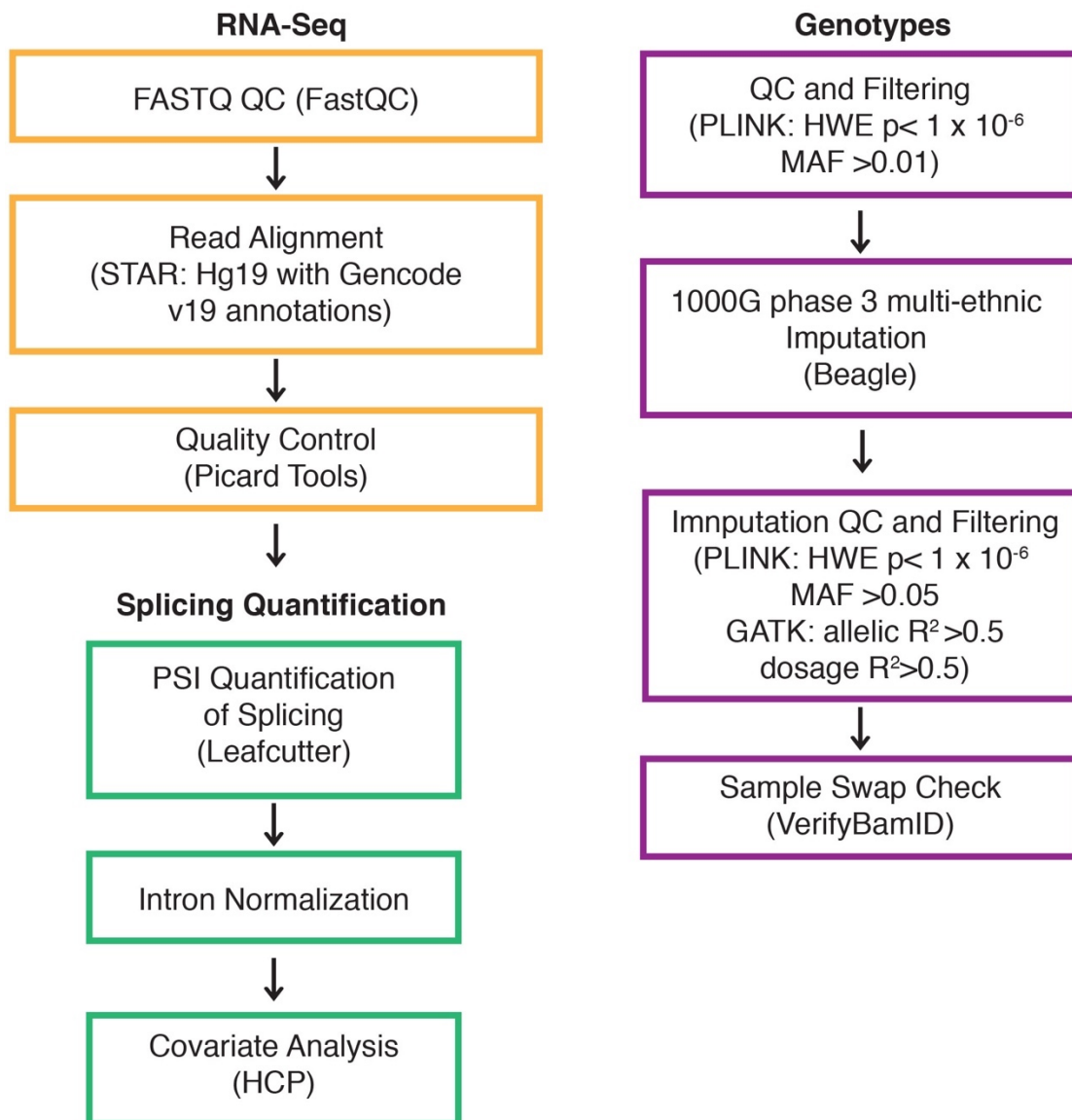


**Figure 3.S1: Overview of raw data processing and QC**

Overview of methods and QC pipeline for processing RNA sequencing FASTQ files to intron quantifications and genotype imputation, QC, and filtering. Output data from this pipeline was used as inputs for further sQTL analysis.

### 3.2.b: sQTL analysis

Standardized and normalized intron excision ratios calculated by leafcutter was used as the phenotype for sQTL mapping. FastQTL (Ongen et al., 2016) was used to test for association between SNPs within a cis-region of  $\pm 100\text{kb}$  of the intron cluster and intron ratios within cluster. Hidden covariate analysis was performed using Hidden Covariates with a Prior (HCP) (Mostafavi et al., 2013) on intron excision ratios given the same known covariates used for eQTL HCP calculations. I included 5 HCPs in our spliceQTL model (which we found to maximized intron QTL discovery; **Figure 3.S2**) and also included gestation week, RIN, and sex as additional covariates. FastQTL was run in the permutation pass mode (1000 permutations). Beta approximated permutation p-values were then multiple test corrected using the q-value Storey and Tibshirani FDR correction. I define sQTL as an intron having an FDR q-value  $\leq 0.05$ , and an sGene as a gene containing a significant sQTL at any intron.

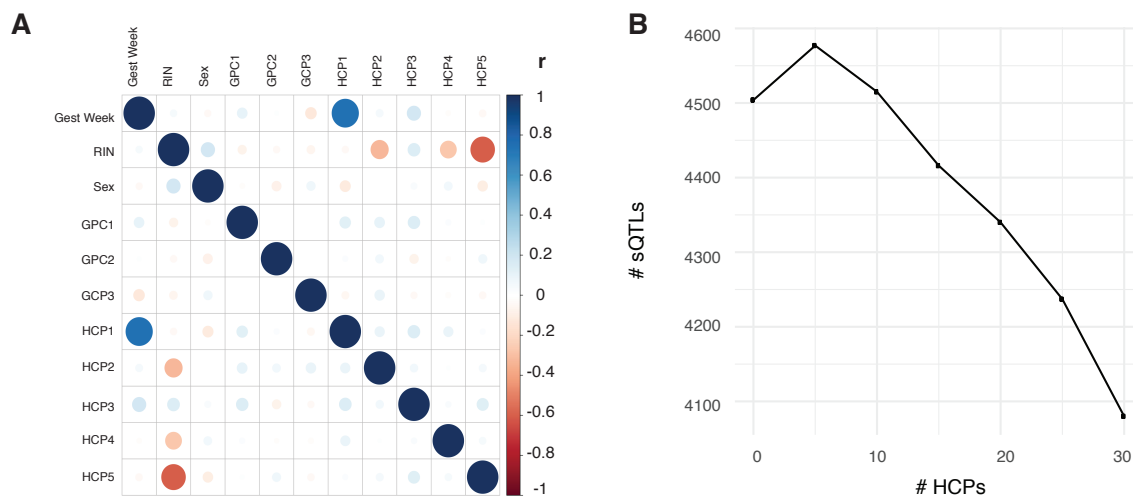


**Figure 3.S2: Identification of optimal number of HCP selection and correlations of all covariates in final sQTL analysis**

**(A)** Pearson correlation ( $r$ ) between all covariates used in the final sQTL analysis. **(B)** The number of hidden covariates (HCP) was chosen to maximize sQTL discovery. sQTL analysis was performed correcting for gestation week, RIN, sex, the top 3 genotype PCs and then increments of 5 HCPs using FastQTLs permutation scheme. The number of HCPs included in the sQTL analysis is shown on the X axis with the number of significant sGenes is shown on the Y axis. The number of sGenes was determined by FDR q-value  $\leq 0.05$ . Based on these results, 5 HCPs were selected for the final sQTL analysis.

### *3.2.c: Enrichment in genome-wide annotations*

I performed functional enrichment of sQTLs using GREGOR (Genomic Regulatory Elements and Gwas Overlap algoRithm) (Schmidt et al., 2015) to evaluate enrichment of variants in genome wide annotations. I downloaded the 25 state ChromHMM model BED files from the Roadmap Epigenetics Project (Ernst and Kellis, 2015; Roadmap Epigenomics Consortium et al., 2015), generated from a set of 5 core chromatin marks assayed in fetal brain. GREGOR evaluates the enrichment of QTL variants in these genomic annotations by estimating the significance of observed overlap of the sSNP relative to the expected overlap using a set of matched control variants. GREGOR creates a list of possible causal SNPs by extending the list of sSNPs (index SNPs) to all SNPs in high linkage disequilibrium ( $r^2 > 0.7$ ). A set of matched control SNPs (SNPs are selected based on matching the index SNP for number of variants in LD, minor allele frequency, and distance to nearest gene/intron) is then created, and enrichments are calculated based on the observed and expected overlap within each annotation.

### *3.2.d: Characterization of splicing drivers*

I used human RNA binding protein (RBP) binding site BED files from CLIPdb (Yang et al., 2015) database of publicly available cross-linking immunoprecipitation (CLIP)-seq datasets from 51 RBPs to evaluate the enrichment of sSNPs in the RNA binding protein binding sites using GREGOR.

I obtained SRRM4 CLIP-seq peaks (Yang et al., 2015) and PSI values from SRRM4 overexpression in human 293T cells from averaged from 3 overexpression samples and 3 control samples (Parikshak et al., 2016; Raj et al., 2014). CLIP-seq peaks were overlapped with fetal sQTLs using the GenomicRanges package in R. The leafcutter identified intron of the sGene was identified in the overexpression and control PSI data and plotted using ggplot.

### 3.2.e: Overlap with eQTLs

I used Storey's  $\pi_1$  statistic described in Nica et al (Nica et al., 2011), to assess the proportion of true associations among sQTLs that were also detected by the eQTL analysis and eQTLs that were also detected by the sQTL analysis. The overlap was assessed by taking all significant SNP-gene associations from the eQTLs and estimating the proportion of true associations ( $\pi_1$ ) on the distribution of corresponding p-values of the overlapping SNP-gene pairs in the sQTL data set and vice versa. This is done by first estimating  $\pi_0$ , the proportion of true null associations based on their distribution. Then  $\pi_1 = 1 - \pi_0$  estimates the lower bound of true positive associations.

Ensembl's Variant Effect Predictor (VEP) version 90 (McLaren et al., 2016) was used to annotate the effects of variants of significant QTLs on genes, transcripts, protein sequence, and regulatory regions. VEP annotations are based off of a wide range of reference data including Ensembl database version 92, GRCH37.p13 genome assembly, Gencode 19 gene annotations, RefSeq 2015-01, PolyPhen 2.2.2, SOFT 5.2.2, dbSNP 150, COSMIC 81, ClinVar 2017-06, and gnomAD r2.0.

Cell type enriched genes were obtained from (Polioudakis et al., 2019) single-cell RNA-seq dataset of GW17-18 human fetal cortex. Briefly, Drop-seq was run on single cells isolated from human fetal neocortex according to the online Drop-seq protocol v.3.1 (<http://mccarrolllab.com/download/905/>) and the methods published in Macosko et al. (Macosko et al., 2015). The raw Drop-seq data was processed using the Drop-seq tools v1.12 pipeline from the McCarroll Laboratory (<http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf>). Normalization was performed using Seurat v2.0.1 (Butler et al., 2018). Raw counts were read depth normalized by dividing by the total number of UMIs per cell, then multiplying by 10,000, adding a value of 1, and log transforming ( $\ln(\text{transcripts} + 1)$ ).

per-10,000 + 1)) using the Seurat function 'CreateSeuratObject'. To identify cell type enriched genes, differential expression analysis was performed for each cluster individually versus all other cells in the dataset for genes detected in at least 10% of cells in the cluster. Differential expression analysis was performed using a linear model implemented in R as follows:  $\text{lm}(\text{expression} \sim \text{number\_of\_UMI} + \text{donor} + \text{lab\_batch})$ . P-values were then Benjamini-Hochberg corrected (Benjamini and Hochberg, 1995). Genes were considered enriched if they were detected in at least 10% of cells in the cluster,  $>0.2 \log_2$  fold enriched, and Benjamini-Hochberg corrected p-value  $< 0.05$ . Cell type enriched genes were annotated based on the gene harboring an eQTL in either the fetal brain dataset or GTEx adult cortex dataset (The GTEx Consortium et al., 2017).

### 3.3: Results

I quantified intron clusters using Leafcutter (Li et al., 2018), identifying 92,449 intronic excision clusters which mapped to 10,926 genes. A major strength of Leafcutter is that it is annotation free, allowing for alternative exon discovery, which is especially important in the context of brain isoforms, for which annotations are known to be incomplete. Of the 92,449 detected introns, 25,024 are new, unannotated, introns based on Gencode hg19 gene annotations. Per sample intron abundances (PSI, percent spliced in) were then used as a quantitative molecular trait for local spliceQTL (sQTL) discovery in FastQTL (Ongen et al., 2016); Methods, **Figure 3.S1**), identifying a total of 4,635 significant sQTLs (5% FDR) in 2,132 genes (sGenes; **Figure 3.1 A**). Of the 4,637 significant sQTLs, 3,295 were annotated introns (71%) and 1,255 (27%) new cryptic introns (cryptic 5', cryptic\_3', cryptic unanchored, novel annotated pair; Methods).

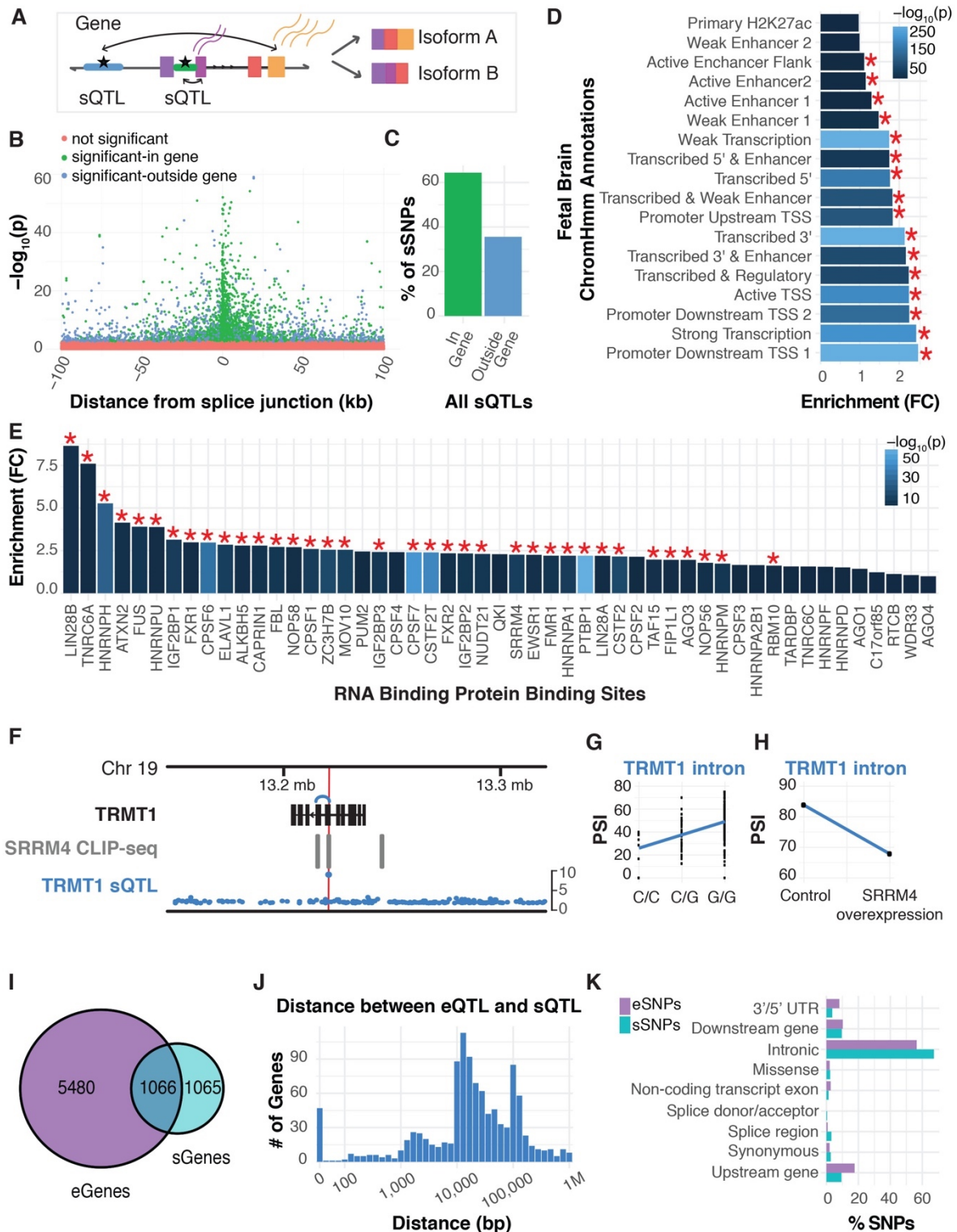


**Figure 3.1: Characterization of sQTLs and Comparison to eQTLs.**

**(A)** A schematic of sQTL detection and intron excision when genes are spliced into multiple isoforms. **(B)** Position of sQTLs in relation to the splice junction. Significant sQTLs are annotated based on whether the sSNP lies within (green) or outside (blue) of the corresponding

sGene. **(C)** Fraction of sQTLs where the sSNP lies within vs outside its sGene. **(D)** Fold enrichment of sSNPs by their distribution within prenatal brain chromatin states, which are depicted on the Y axis. **(E)** Fold enrichments of sSNPs in experimentally discovered RBP binding sites (Yang et al., 2015). A red asterisk indicates significance at 5% FDR. **(F)** Whole-gene view of *TRMT1* highlighting the association  $-\log_{10}(P)$  for all cis-SNPs with PSI values for a *TRMT1* intron, the sSNP is denoted by the red line. Additionally, the sSNP falls within a *SRRM4* CLIP-seq binding site. **(G)** The sQTL for the *TRMT1* intron, showing the distribution of PSI values of *TRMT1* per genotype. **(H)** The average PSI values of the *TRMT1* intron in 3 control samples versus 3 *SRRM4* overexpression samples. **(I)** Venn diagram showing the overlap of genes containing an eQTL vs sQTL. **(J)** Distribution of the distance in base pairs between eSNP and sSNP for genes harboring both. **(K)** Relative proportions of VEP predicted SNP effects between eSNPs and sSNPs.

Positional enrichment of the most significant SNP per sQTL (sSNP) shows clustering around the splice junction, with 42% of sSNPs within 10 kb of the splice junction (**Figure 3.1 B**), demonstrating that variants proximal to splicing junctions have a large effect. In contrast to eQTL, the majority of sSNPs (64%) lie within the gene body (**Figure 3.1 C**), consistent with expectations based on data from other tissues (Li et al., 2016). Splice QTLs were most strongly enriched in promoters and transcribed regions of the Roadmap Epigenetics Consortium chromatin states from fetal brain tissue (**Figure 3.1 D**), consistent with regions functionally relevant to gene splicing mechanisms (Lappalainen et al., 2013; Takata et al., 2017).

To validate that the identified sQTLs are tagging splicing regulatory regions, I evaluated sSNP enrichment in RNA binding protein (RBP) binding sites using the CLIPdb database of publicly available cross-linking immunoprecipitation (CLIP)-seq datasets representing 51 RBPs (Yang et al., 2015). We identified experimentally determined binding sites for 36 of the RBPs to be significantly enriched genome-wide among sSNPs (**Figure 3.1 E**; Methods), the majority of which do not have well characterized roles in CNS function. Among the identified RBPs with known roles in neurodevelopment are *HNRNPH*, *ATXN2*, and *SRRM4*. *HNRNPH* is involved in many aspects of neurodevelopment from alternative splicing of *TRF2*, which is implicated in neuronal differentiation (Grammatikakis et al., 2016), to oligodendrocyte differentiation (Wang et

al., 2007). *ATXN2* plays roles in proliferation and cell growth and known mutations in *ATXN2* cause neuropsychiatric phenotypes, from spinocerebellar ataxia and ALS to schizophrenia (Almaguer-Mederos et al., 2018; Kim et al., 2018; Zhang et al., 2014a). *SRRM4* was recently shown to regulate the splicing of the majority of microexons, many of which exhibit a high degree of evolutionary conservation and are spliced in a neuron-specific manner (Irimia et al., 2014). Microexons regulated by *SRRM4* are dysregulated in post mortem brains in patients with ASD (Irimia et al., 2014) and loss of function in mice has been linked to numerous neurodevelopmental deficiencies (Quesnel-Vallieres et al., 2015).

I obtained differential splicing changes from *SRRM4* overexpression (Parikshak et al., 2016; Raj et al., 2014), and as an example, focused on a single gene, *TRMT1*, which has been implicated in intellectual disability and whose sSNP falls within a *SRRM4* CLIP-seq peak (Yang et al., 2015)(**Figure 3.1 F**). The strong sQTL signal in *TRMT1* within this CLIP-seq peak (**Figure 3.1 G**) suggested that *SRRM4* would regulate this intron. Indeed, *SRRM4* overexpression leads to significant differential splicing of the same intron for *TRMT1* in human cells in vitro (Parikshak et al., 2016; Raj et al., 2014)(**Figure 3.1 H**). Thus, these splicing factors exhibiting genome-wide significant enrichment with fetal sSNPS provide functional links between human allelic variation at these loci and the factors that modulate alternative splicing. The subsequent changes in expressed protein sequence are likely to have considerable consequences for downstream neuronal functioning and provide high priority candidates for subsequent mechanistic investigation.

To assess sharing of eQTL and sQTLs, I next analyzed the overlap of genes harboring a significant eQTL compared to those with sQTLs. I found that about half of all genes with a sQTL were also an eGene (1,066 genes out of the 6,546 eGenes; **Figure 3.1 I**). To determine the extent in which eQTL and sQTL might have overlapping effects, we performed a threshold free

comparison of significant sQTLs from Leafcutter with eQTL summary statistics to estimate the proportion of non-null associations among the eQTLs (Storey and Tibshirani, 2003); Methods). I find that 67% of sQTLs also affect total levels of gene expression; however, only 22% of eQTLs affect splice-junction usage, suggesting mostly independent regulation of expression and splicing events, parallel with observations in other tissues (Li et al., 2016). This independence of expression and splicing regulation is also observed when comparing eSNPs and sSNPs for the same gene. Many of the tagged regulatory regions for expression and splicing of the same gene are distinct, as evidenced by the large distances between eSNP and sSNPs influencing the same gene (**Fig 3.1 J**), with 30% of the genes exhibiting a  $D' < 0.7$  between the eSNP and sSNP (**Figure 3.S3**). At the same time, we note that there are a minority of SNPs that affect both total expression and splicing: only 47 genes have directly overlapping sQTL and eQTL (**Figure 3.1 J**; **Table 3.S1**). For validation, we also used Ensembl's Variant Effect Predictor (VEP) (McLaren et al., 2016); Methods) which annotates SNP function, to compare the distributions of our splicing and expression related variants, finding very similar results. For example, we observe more sSNPs identified as an intronic variant than eSNPs, and more eSNPs being identified as upstream gene variants (**Figure 3.1 K**).

Linkage between eSNPs and sSNPs

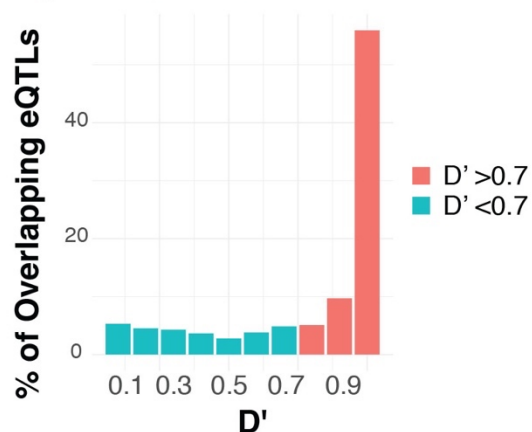


**Figure 3.S3: Genes harboring both an eQTL and sQTL**

Linkage disequilibrium ( $D'$ ) between the eSNP and sSNP of genes with both an eQTL and sQTL.

**Table 3.S1: Genes having an overlapping eQTL and sQTL**

| Gene      | ENSID           | SNP              | sQTL pval  | eQTL pval  |
|-----------|-----------------|------------------|------------|------------|
| TTC27     | ENSG00000018699 | 2_32903548_G_A   | 0.00237433 | 8.69E-20   |
| FECH      | ENSG00000066926 | 18_55238820_A_G  | 0.00014177 | 2.02E-19   |
| TTC38     | ENSG00000075234 | 22_46691400_T_C  | 0.00905827 | 4.74E-13   |
| CDC7      | ENSG00000097046 | 1_91965850_C_T   | 0.00082423 | 1.35E-31   |
| FAM21A    | ENSG00000099290 | 10_51808725_C_A  | 1.23E-07   | 1.04E-25   |
| KLHDC4    | ENSG00000104731 | 16_87764267_A_T  | 4.45E-06   | 0.00131365 |
| DFNA5     | ENSG00000105928 | 7_24746022_A_G   | 6.75E-08   | 0.01045235 |
| NPHP3     | ENSG00000113971 | 3_132436967_A_G  | 0.04207235 | 7.98E-13   |
| DNAJC27   | ENSG00000115137 | 2_25218054_T_G   | 0.01168616 | 0.00044877 |
| MTERFD2   | ENSG00000122085 | 2_242033696_T_C  | 5.74E-15   | 5.12E-29   |
| FAM64A    | ENSG00000129195 | 17_6353679_A_C   | 3.61E-06   | 0.00025836 |
| TEP1      | ENSG00000129566 | 14_20881527_T_C  | 7.36E-06   | 2.78E-09   |
| FIGNL1    | ENSG00000132436 | 7_50515627_G_A   | 9.23E-24   | 4.00E-42   |
| NEK3      | ENSG00000136098 | 13_52721067_C_T  | 2.65E-22   | 4.85E-11   |
| ZDHHC4    | ENSG00000136247 | 7_6617247_C_G    | 0.00030411 | 6.64E-14   |
| NARS2     | ENSG00000137513 | 11_78285875_A_T  | 1.14E-05   | 3.16E-19   |
| POC1B     | ENSG00000139323 | 12_89873975_A_T  | 0.00086057 | 2.87E-18   |
| SEMA6C    | ENSG00000143434 | 1_151126966_G_A  | 0.01473195 | 1.18E-08   |
| IMMP1L    | ENSG00000148950 | 11_31515477_C_G  | 5.60E-09   | 0.00105085 |
| TMCO3     | ENSG00000150403 | 13_114192125_A_G | 2.23E-12   | 0.00089985 |
| MPPE1     | ENSG00000154889 | 18_11905953_C_T  | 3.65E-08   | 7.28E-16   |
| TMED4     | ENSG00000158604 | 7_44588042_A_C   | 8.23E-07   | 0.00126401 |
| CAPN2     | ENSG00000162909 | 1_223949349_T_C  | 3.49E-17   | 9.07E-12   |
| SUCLG1    | ENSG00000163541 | 2_84618582_G_T   | 0.009083   | 0.03477109 |
| NSA2      | ENSG00000164346 | 5_74002514_G_C   | 2.92E-05   | 8.49E-33   |
| DIS3L     | ENSG00000166938 | 15_66585907_C_A  | 0.0035181  | 5.84E-21   |
| RPSA      | ENSG00000168028 | 3_39448804_G_A   | 1.92E-13   | 9.07E-49   |
| NBPF9     | ENSG00000168614 | 1_144894697_C_T  | 0.00190516 | 1.28E-13   |
| RNF181    | ENSG00000168894 | 2_85824073_T_C   | 1.79E-09   | 0.00108711 |
| ZFAND4    | ENSG00000172671 | 10_46167847_C_T  | 2.95E-10   | 7.75E-09   |
| SNX32     | ENSG00000172803 | 11_65601560_G_C  | 0.00247304 | 1.48E-11   |
| LRRC63    | ENSG00000173988 | 13_46802177_A_G  | 0.00124643 | 0.00017842 |
| EIF3F     | ENSG00000175390 | 11_8014479_C_T   | 4.58E-12   | 2.13E-11   |
| ZNF135    | ENSG00000176293 | 19_58572979_G_A  | 0.00489929 | 1.02E-13   |
| LINC00471 | ENSG00000181798 | 2_232378792_C_T  | 0.03320858 | 5.09E-18   |
| SKA2      | ENSG00000182628 | 17_57208752_T_C  | 0.00281451 | 5.74E-08   |
| TTC32     | ENSG00000183891 | 2_20155545_A_T   | 0.01026047 | 1.98E-11   |

|           |                 |                  |            |            |
|-----------|-----------------|------------------|------------|------------|
| GSTT1     | ENSG00000184674 | 22_24334948_A_C  | 0.04514398 | 2.51E-17   |
| EP400NL   | ENSG00000185684 | 12_132588733_G_A | 0.0262901  | 6.37E-14   |
| POLR1D    | ENSG00000186184 | 13_28196033_G_C  | 0.00011736 | 8.07E-15   |
| C2orf80   | ENSG00000188674 | 2_209036712_T_C  | 6.15E-05   | 0.00030176 |
| KIAA0895L | ENSG00000196123 | 16_67219107_G_C  | 2.56E-10   | 7.04E-05   |
| ECI2      | ENSG00000198721 | 6_4119516_A_G    | 5.22E-05   | 0.0035567  |
| ZNF468    | ENSG00000204604 | 19_53357172_C_T  | 4.76E-05   | 0.00806517 |
| DNLZ      | ENSG00000213221 | 9_139260551_C_T  | 0.02687819 | 1.23E-09   |
| C5orf54   | ENSG00000221886 | 5_159824464_A_G  | 3.32E-11   | 0.00019314 |
| SNHG8     | ENSG00000269893 | 4_119200274_C_T  | 3.13E-31   | 6.50E-17   |

## **CHAPTER FOUR**

### Tissue specificity of QTLs

## 4.1: Introduction

To begin to unravel the question of age and tissue specificity of gene regulation in the brain, I compared fetal derived eQTLs and sQTLs to previously published adult QTL data sets, including GTEx, CommonMind and PsychENCODE, as well as non-brain eQTL data sets from GTEx. I have compared these data by eGene and sGene overlap as well as effect size and p-value based methods to evaluate overlap of SNP-gene pairs between datasets, and further describe properties of fetal specific eQTLs versus eQTLs shared across tissue types. I then evaluated the relevance of fetal QTLs compared to adult QTLs, in terms of contribution to disease heritability through LD score partitioned heritability in eQTL and sQTL defined regulatory regions. Lastly, I compared eQTL to sQTL contribution to disease heritability for multiple brain disorders and traits such as ASD, SCZ, ADHD, head circumference, intracranial volume, educational attainment, major depressive disorder, and Alzheimer's disease.

## 4.2: Methods

### 4.2.a: *Effect size correlations*

I downloaded GTEx v7 eQTL summary statistics for all 48 tissue types (The GTEx Consortium et al., 2017). To compare effect sizes consistently between studies, we calculated effect size by running a linear model with scaled log tpm expression values for significant fetal eQTLs and calculated the same way for corresponding SNP-gene pairs in the GTEx data to obtain a beta value from non-standard normalized expression. Significant fetal eQTLs were identified as fetal specific if the corresponding SNP-gene pair was not found in any GTEx tissue or as shared if it was found in at least one GTEx tissue.

Effect size correlations between all GTEx tissues and PsychENCODE (The GTEx Consortium et al., 2017; Wang et al., 2018b) were calculated by first obtaining all  $FDR \leq 0.05$  (q-value) nominal fetal eQTLs. Nominal eQTL analysis was run in FastQTL, using the same

input for the permutation pass, to obtain all SNP-gene pairs tested. Spearman's  $\rho$  correlations were calculated per tissue on the absolute value of the slope from FastQTL output of all FDR $\leq$ 0.05 fetal eQTLs and corresponding absolute value of slope from SNP-gene pairs in GTEx nominal associations. The absolute value of the slope was used for all correlations to control for strand-flips. For each tissue compared in this analysis, I indicate how many of the fetal eGenes were found in the tissue of comparison by the sample size of that tissue (**Figure 4.S1 C**). A scatterplot of the fetal brain versus PsychEncode adult brain effect sizes shows corresponding eQTL regression beta values from each dataset (**Figure 4.S1 D**).

#### *4.2.b: Pi-hat analysis*

Additionally, I used Storey's Qvalue software (Storey and Tibshirani, 2003) to assess overlap between fetal brain eQTLs and the eQTLs from the GTEx v7 tissues (The GTEx Consortium et al., 2017). The proportion of true associations ( $\pi_1$ ) was estimated by looking up significant fetal brain eQTLs in each of the GTEx tissues, creating a distribution of corresponding p-values of the overlapping SNP-gene pairs used to calculate  $\pi_0$ , the proportion of true null associations based on their distribution. Then  $\pi_1=1-\pi_0$  estimates the lower bound of true positive associations. I performed the reciprocal overlap by looking up GTEx significant eQTLs per tissue in the fetal brain dataset (**Figure 4.S1 A,B**)

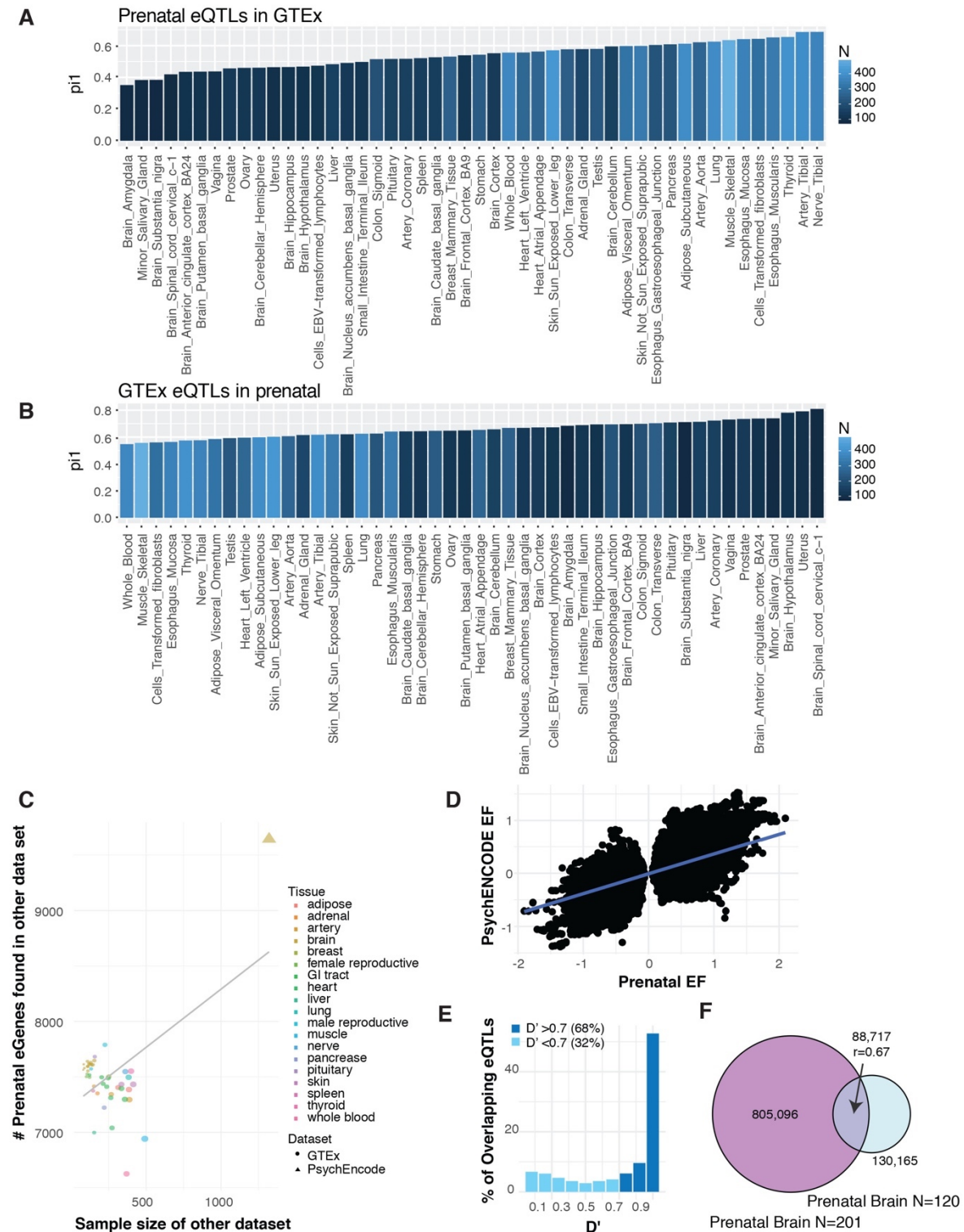


**Figure 4.S1: Cross-tissue overlap of eQTLs.**

**(A)** Pi1 values (proportion of true positive p-values) for significant prenatal brain eQTLs in GTEx tissues, colored by sample size (N) of GTEx tissue dataset. **(B)** Pi1 values (proportion of true positive p-values) for significant GTEx eQTLs per tissue in the prenatal brain eQTLs, colored by

sample size (N) of GTEx tissue dataset. **(C)** Scatterplot showing the number of prenatal eGenes that overlap cross-tissue datasets versus the sample size of the other dataset. Sample size is shown on the X axis, eGenes overlap is shown on the Y axis, each point is colored by the tissue type, with the shape of the point indicating which dataset it came from. **(D)** Scatterplot of effect sizes (regression beta) of overlapping eQTLs (FDR 5%) from prenatal brain vs PsychEncode. **(E)** Linkage disequilibrium ( $D'$ ) between the eSNPs of genes with overlapping eQTLs between prenatal brain and GTEx cortex. **(F)** Overlap of FDR-significant nominal eQTLs of prenatal brain eQTL datasets of varying sample sizes (N=201 vs N=120). The effect size correlation is shown for overlapping eQTLs.

#### 4.2.c: eGene Overlap

Overlap between fetal brain eGenes and GTEx Cortex eGenes was performed by intersecting all significant eGenes detected by permutation test followed by FDR correction. Overlap between fetal brain eGenes and PsychENCODE eGenes was performed by intersection all significant eGenes detected by all nominal eQTLs with an  $FDR \leq 0.05$ . Overlap between fetal brain sGenes and adult sGenes was performed by intersecting the genes of significant sQTLs by permutation test followed by FDR correction.

Overlap between our fetal brain eQTL dataset and a fetal brain eQTL dataset consisting of 120 samples (O'Brien et al., 2018), were performed by intersecting all nominal eQTLs with an  $FDR \leq 0.05$ . Spearman's  $\rho$  correlations were calculated for eQTL effect sizes (slope) for overlapping FDR-significant eQTLs (**Figure 4.S1 F**).

#### 4.2.d: Partitioned heritability by LD score regression

Partitioned heritability was measured using LD Score Regression v1.0.0 (Finucane et al., 2015) to identify enrichment of GWAS summary statistics among functional genomic annotations by accounting for LD, specifically eQTL and sQTL regulatory regions. The full baseline model of 53 functional categories was downloaded from Finucane et al. (<https://github.com/bulik/ldsc/wiki/Partitioned-Heritability>). Fetal and Adult eQTL annotations from both GTEx Cortex and PsychENCODE, and fetal sQTL and adult sQTL annotations

(Takata et al., 2017), were created by taking a 500bp window (+/-250) around each eSNP or sSNP. This resulted in 6,163 fetal eQTL annotations, 5,690 adult eQTL GTEx annotations, 32,944 adult PsychENCODE eQTL annotations, 4,635 fetal sQTL annotations, and 8,966 adult CMC sQTL annotations. An annotation file was then created by marking all HapMap3 (International HapMap, 2003) SNPs that fell within the QTL annotations. LD scores were calculated for the QTL annotation SNPs using an LD window of 1cM using LD reference panel 1000 Genomes European Phase 3 (The 1000 Genomes Project Consortium et al., 2015). This LD reference panel was chosen due to the CLOZUK+PGC SCZ GWAS (Pardinas et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics, 2014) being comprised of mainly European ancestry samples (Schizophrenia Working Group of the Psychiatric Genomics, 2014). Baseline LD-scores and QTL LD-scores were simultaneously included in computation of partitioned heritability. Enrichment for each annotation was calculated by the proportion of heritability explained by each annotation divided by the proportion on SNPs in the genome falling in that annotation category. Enrichment p-values were then Bonferroni corrected. All fetal eQTLs and sQTLs were tested separately for enrichment in genetic variants associated with schizophrenia (Pardinas et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics, 2014), major depressive disorder (CONVERGE Consortium, 2015), intracranial volume (Adams et al., 2016), inflammatory bowel disease (Jostins et al., 2012), head circumference (Taal et al., 2012), epilepsy (International League Against Epilepsy Consortium on Complex Epilepsies, 2014), educational attainment (Rietveld et al., 2013), cognitive performance (Benyamin et al., 2014), Autism (Grove et al., 2019), Alzheimer's disease (Lambert et al., 2013), and attention deficit hyperactivity disorder (Demontis et al., 2019).

### 4.3: Results

To examine eQTL sharing among tissue types, I examined the correlation of effect sizes of fetal brain eQTLs (Nica et al., 2011) to those from 48 different tissue types from the Genotype-Tissue Expression Consortium (GTEx) (Methods; (The GTEx Consortium et al., 2017). Fetal eQTLs found in one or more GTEx tissue types showed significantly lower effect sizes ( $p\text{-value} < 2.2 \times 10^{-16}$ ) than those that were fetal specific, consistent with previous studies that show eQTLs that are active globally across tissue types are more constrained, whereas tissue specific eQTLs have a greater magnitude of effect sizes (Mohammadi et al., 2017; The GTEx Consortium et al., 2017); **Figure 4.1 A**). In contrast to their regulatory regions, fetal brain-specific eGenes themselves show lower tolerance to loss of function mutations as measured by pLI (Lek et al., 2016), compared with those eGenes shared between one or more GTEx tissue types (**Figure 4.1 B**). This highlights a model whereby tissue specific regulatory control can provide a subtler means for evolution to vary protein expression levels, versus the often larger and disruptive effects of protein coding variation. To assess fetal eQTL sharing between CNS and non-CNS tissues, we correlated the effect sizes of all significant fetal eQTLs to the same eQTL SNP-gene pair across all GTEx tissues and PsychENCODE prefrontal cortex (Methods). We observed the strongest Spearman's  $\rho$  correlations of fetal brain eQTLs with adult brain tissue and proliferative epithelial containing tissues (**Figure 4.1 C**).

Another central question is related to developmental stage: how do fetal brain eQTLs compare with adult brain eQTLs? To address this, I first compared genes that harbor an eQTL in our dataset to that of GTEx adult cortex eQTLs (N=136 individuals, eGenes=6,146), a dataset of similar sample size, experimental design, and number of significant eQTLs (The GTEx Consortium et al., 2017). We found 2,532 eGenes that overlapped between fetal and adult, accounting for slightly more than one third of both datasets (**Figure 4.1 D**). Of the eGenes that overlapped, I examined how many eSNPs were tagging the same region by calculating the LD

between the top primary SNPs in the fetal and adult data in the 1000 Genomes Project phase 3 multi-ethnic reference panel (Methods; (The 1000 Genomes Project Consortium et al., 2015). We found that 68% of overlapping eQTLs tag the same regulatory region, even if the top SNP differed (**Figure 4.S1 E**). Additionally, we compared our 11,625 eGenes (5% FDR all SNP-gene pairs) to eGenes from the PsychENCODE adult brain eQTLs (N=1,866, eGenes= 32,944), the largest adult brain dataset to date -- with a sample size nearly 10 times larger than this cohort (Wang et al., 2018a). I found 10,233 eGenes to overlap between fetal and adult. However, I still find more than 1,000 eGenes specific to the fetal dataset (**Figure 4.1 E**). I note that this is likely an underestimate of the true number of fetal specific eQTL, given the fetal brain dataset's ten-fold smaller sample size.

I also compared genes that harbor a sQTL in our dataset to sQTLs identified in an adult brain prefrontal cortex dataset consisting of 206 control individuals from the CommonMind Consortium (CMC) (Takata et al., 2017). Though the methods used to detect splicing differ between studies (annotation-free intron excision events in the current study vs annotation based alternative splicing events in the CMC; Methods), I reasoned that comparison at the gene level would give insight to the potential stage specificity of genes that undergo splicing. We found that although the data sets were the same size, only 24% of the sGenes identified in fetal brain were sGenes in the adult brain, which accounts for 39% of the adult brain sGenes (**Figure 4.1 F**), suggesting that many splicing events may be stage specific. This emphasizes the need for more direct and comparable splicing comparisons across different stages of brain development.

To further examine the differences in eGenes at the different developmental time points, eGenes were annotated based on fetal cell type markers identified through differential expression of single cell sequencing identified cell clusters (Methods; (Polioudakis et al., 2019). I found that many more fetal specific eGenes are fetal cell type markers, compared with GTEx

eGenes that are adult specific (**Figure 4.1 G; Methods**), which is consistent with expectations given the cellular composition of each epoch. One interesting exception are markers for microglia, which is a neural-immune cell present in fetal and adult brain, but that showed enrichment for fetal eGenes.

Previous work based on chromatin accessibility during fetal development had suggested that liability for SCZ was significantly enriched in fetal brain, especially in neural progenitor containing zones (de la Torre-Ubieta et al., 2018). Given the many differences between fetal and adult brain eQTLs at both the level of SNPs and genes, we reasoned that partitioning disease risk imparted by common genetic variation could inform the question as to the timing of genetic contributions to disease risk. To test this, I created eQTL and sQTL annotation regions for fetal cortex and adult cortex by considering a 500bp window (+-250) around each eSNP or sSNP, resulting in a comparable number of annotations for each epoch (6,163 fetal eQTL annotations, 5,690 adult eQTL GTEx annotations, 32,944 adult PsychENCODE eQTL annotations, 4,635 fetal sQTL annotations, 8,966 adult CMC sQTL annotations, Methods). We found that SNP-based heritability for SCZ shows fetal brain eQTLs, followed by sQTLs to be the highest enriched among all significant functional categories, with 0.2% of eSNPs explaining an estimated 3.7% of SNP heritability for SCZ ( $p\text{-value} = 9.1 \times 10^{-4}$  for enrichment) and 0.1% of sSNPs explaining 2.1% of SNP heritability for SCZ ( $p\text{-value} = 6.6 \times 10^{-3}$  for enrichment) respectively. In contrast, adult brain eQTLs from both GTEx and PsychENCODE do not reach significance (**Figure 4.1 H**), suggesting that fetal brain enriched regulatory regions harbor a greater proportion of SCZ risk variant than those enriched in adult. However, when the fetal and adult annotations are combined, the proportion of SNP heritability explained for SCZ increases in an additive manner; the combined set of fetal eQTLs and PsychENCODE eQTLs explain 6.6% SNP heritability for SCZ. This further supports the other overlap analyses indicating that fetal and adult regulatory regions are distinct and therefore complementary.

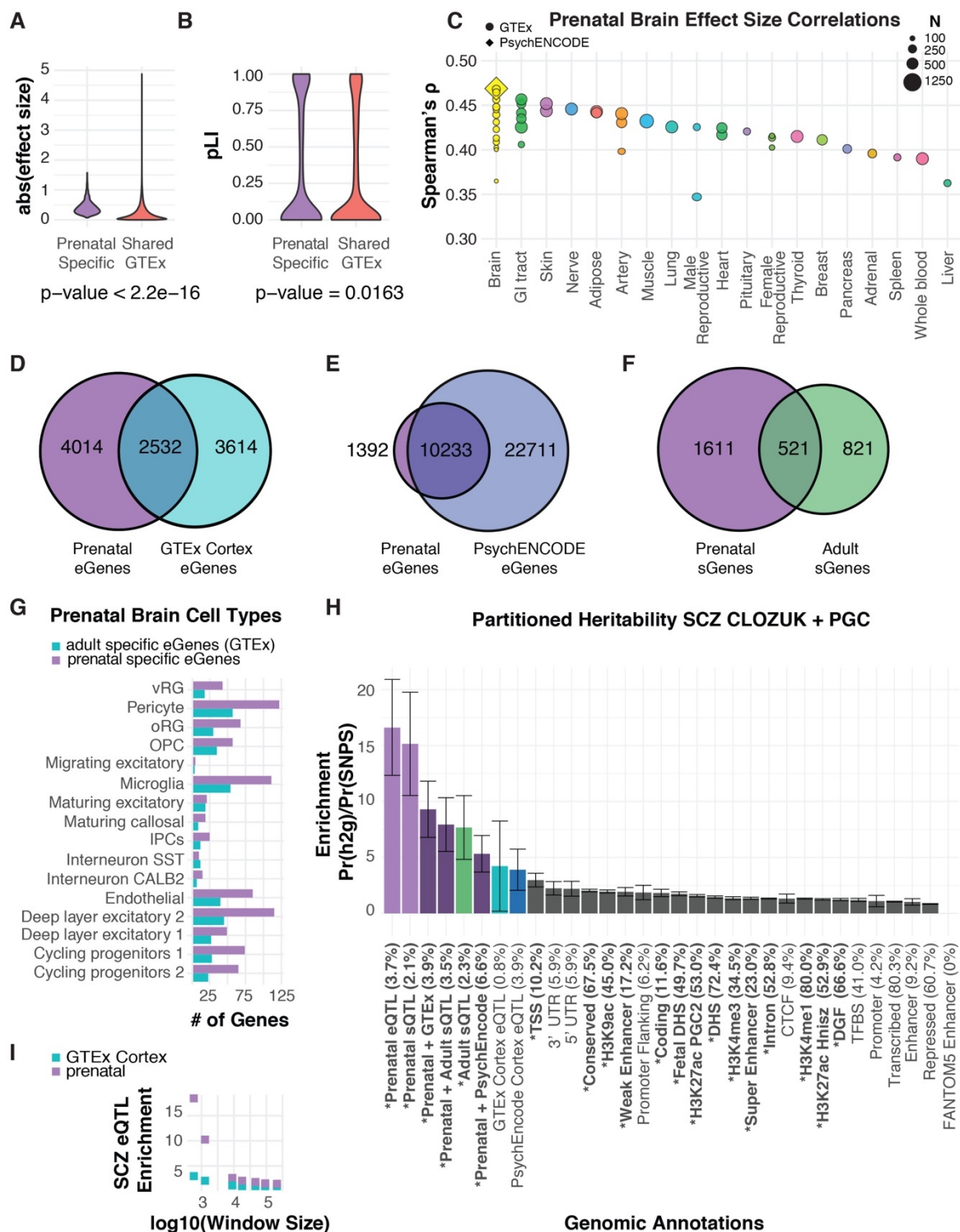


**Figure 4.1: Age Specificity of Brain eQTLs**

(A) Distribution of effect size for prenatal-specific eQTLs vs prenatal eQTLs shared in any GTEx tissue (Wilcoxon Rank Sum  $P < 2.2 \times 10^{-16}$ ). (B) Distribution of pLI scores for prenatal-specific

eQTLs vs prenatal eQTLs shared in any GTEx tissue. (Wilcoxon Rank Sum  $P=0.0163$ ) **(C)** Effect size correlations (Spearman's  $\rho$ ) between significant nominal prenatal eQTLs and corresponding eQTL across tissues grouped by similarity. Tissues from GTEx are denoted as circles, PsychENCODE prefrontal cortex is denoted by a diamond. The size of each point corresponds to the dataset's sample size. **(D)** Venn Diagram comparing eGenes discovered in prenatal vs adult cortex from GTEx. **(E)** Venn Diagram comparing eGenes discovered in prenatal vs adult cortex from PsychENCODE. **(F)** Venn Diagram comparing sGenes discovered in prenatal vs adult cortex (Takata et al., 2017). **(G)** The number of prenatal single cell marker genes containing an eQTL in prenatal-specific eQTLs, GTEx adult brain-specific eQTLs, and those overlapping both datasets. Prenatal = prenatal specific; adult = adult specific; both = overlapping and found in both prenatal and adult. **(H)** LD score regression enrichments where annotations of prenatal eQTLs and adult eQTLs and sQTLs were added to the baseline annotations. A bold annotation and asterisk indicate significance ( $P_{\text{Bonferroni}} < 0.05$ ) for all annotation categories tested, with the proportion of heritability explained in parenthesis. **(I)** LD score regression enrichment of the prenatal cortex and GTEx adult cortex annotations with varying window size around the eSNP.

In contrast to adult eQTL, adult sQTLs do contribute significant SNP heritability in SCZ, explaining 2.3% in SCZ, and when combined with the fetal sQTLs, explain a total of 3.5% of the SNP-based heritability (**Figure 4.1 H**). To test the robustness of this finding, we explored different window sizes around each eGene, removing any window that overlapped between fetal and adult GTEx cortex annotations. We observed consistent significant enrichment over almost an order of magnitude window sizes for fetal annotations over adult for SCZ GWAS loci (**Figure 4.1 I**).

Recent studies have shown heritability for complex disorders is disproportionately enriched in functional categories such as conserved non-coding regions and enhancers (Finucane et al., 2015), as well as regions regulating splicing (Li et al., 2016). However, nothing is known about how allelic variation affects splicing variation and disease risk in developing human brain, and how this compares to effects on transcript expression, so we compared GWAS enrichment patterns in sQTL and eQTL (**Figure 4.2, Methods**). SCZ (Pardinas et al., 2018) associated variants are significantly enriched for both fetal eQTL and sQTL regions (p-

value =0.0004 and p-value =0.002 respectively), whereas educational attainment (Okbay et al., 2016) is only significant for sQTL regions and not eQTL (p-value =0.004; **Figure 4.2**). Genetic variants associated with risk for ADHD (Demontis et al., 2019) show a trend for enrichment in both eQTL and sQTL regions, whereas ASD GWAS (Grove et al., 2019) only shows a trend for enrichment in sQTLs (**Figure 4.2**). As a negative control, I tested variants associated with risk for inflammatory bowel disease (Jostins et al., 2012) and observed no enrichment. Overall, these data are consistent with previous suggestions that sQTL harbor substantial disease risk, perhaps even more so than eQTL (Li et al., 2016), highlighting the relevance of these data to functional characterization of disease-associated risk variants.

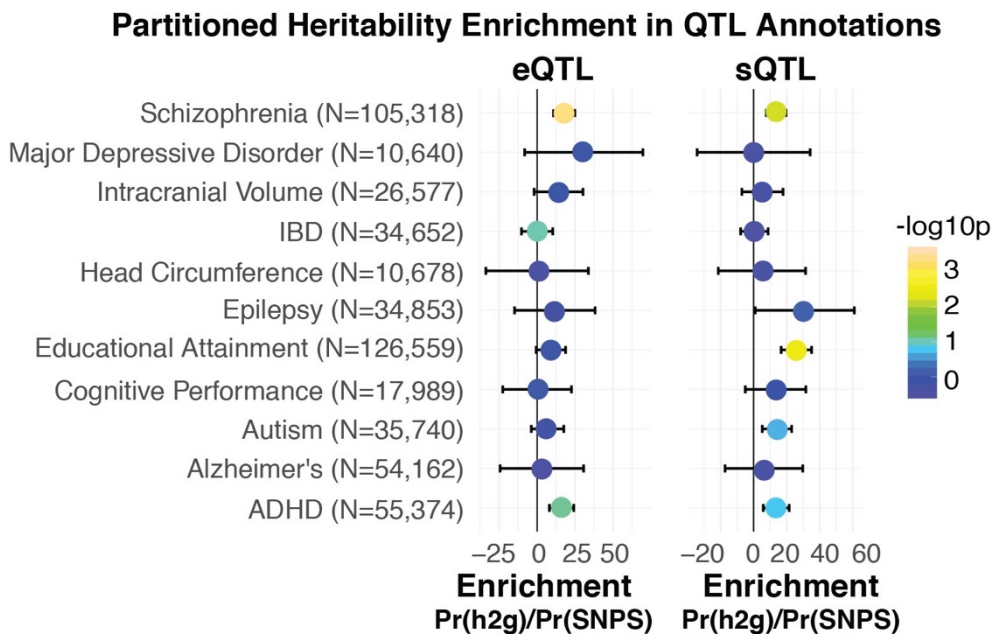


**Figure 4.2: eQTL and sQTL enriched for complex disease heritability**

LD score regression partitioned heritability enrichments for prenatal eQTL and sQTL annotations across multiple GWAS. Enrichments are colored by  $-\log_{10}(P)$  and error bars represent SE.

## **CHAPTER FIVE**

eQTLs in the context of transcriptional networks

## 5.1: Introduction

To explore biological processes under genetic control that differ between fetal and adult, I have built a fetal brain co-expression network using WGCNA to identify modules of co-expressed genes. Detected modules show enrichment of gene ontology terms and cell type specific markers, as well as module preservation between the published network modules of cortical development (Parikshak et al., 2013). Modules were then used to look for enrichment in known disease genes, identified through family-based studies of inherited and *de novo* SNVs and CNVs. Furthermore, modules were integrated with eQTL defined regulatory regions and show disease GWAS loci enriched in regulatory regions of specific biological processes.

## 5.2: Methods

### 5.2.a: WGCNA

After gene counts were filtered by removing genes that were not expressed in 80% of samples with 10 counts or more, expression was conditional quantile normalized, adjusting for gene length and GC content. Sample outliers were removed based on standardized sample network connectivity Z scores  $< 2$ . ComBat batch correction was performed (Johnson et al., 2007). Gestation week, RIN, and the top 4 Picard PCs were regressed from the expression dataset.

Network analysis was performed with robust consensus WGCNA (rWGCNA) (Zhang and Horvath, 2005) assigning genes to specific modules based on biweight midcorrelations among genes. A soft threshold power of 11 was chosen to achieve scale-free topology ( $r^2 > 0.8$ ) (**Figure 5.S1 A,B**). Then, 50 signed co-expression networks were generated on 50 independent bootstraps of the samples; each co-expression network uses the same estimated power parameter. The 50 topological overlap matrices were combined edge-wise by taking the median of each edge across all bootstraps. The topological overlap matrices were clustered hierarchically

using average linkage hierarchical clustering (using  $1 - \text{TOM}$  as a dis-similarity measure). The topological overlap dendrogram was used to define modules using minimum module size of 100, deep split of 4, merge threshold of 0.2, and negative pamStage. Module preservation was run in an independent RNA-seq dataset of cortical development from 8 post conception weeks to 12 months after birth (Parikshak et al., 2013; Sunkin et al., 2013), showing that all of these modules show high preservation in independent data (**Figure S7**).

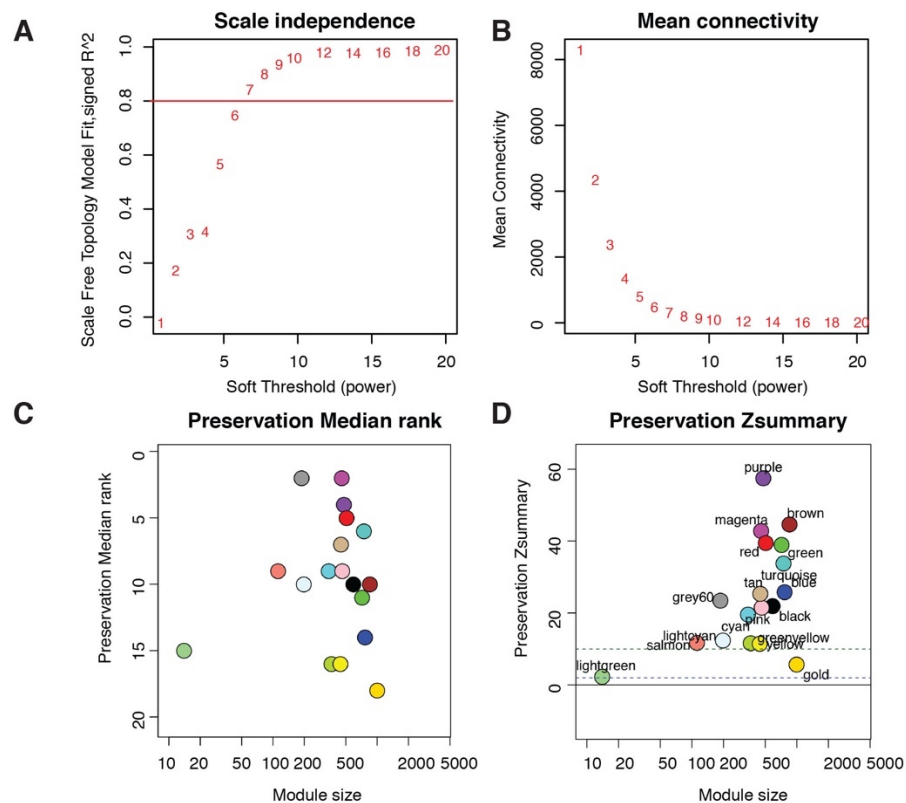


**Figure 5.S1: WGCNA parameters and module preservation**

(A-B) Soft threshold power of 11 was chosen based on scale free topology plateauing at a soft threshold power of 11, as well as mean connectivity. (C-D) Preservation median rand and Z summary of prenatal brain modules in the (Parikshak et al., 2013) network.

#### 5.2.b: Module annotation

### *5.2.b: Module annotation*

#### *GO enrichment*

GO definitions were downloaded from Ensembl release 86. GO terms with a small (<35) or large (>100) number of genes were removed. Logistic regression was performed using the model:  $\text{is.go} \sim \text{is.module} + \text{gene covariates}$  (GC content and gene length) for an indicator-based enrichment, and p-values were Bonferroni FDR corrected. The top two significant terms are reported.

#### *Cell Type enrichment*

Cell type markers from human and mouse brain were downloaded from (Hawrylycz et al., 2015; Lein et al., 2007; Mancarci et al., 2017; Miller et al., 2014; Tasic et al., 2016; Winden et al., 2009; Zhang et al., 2014b; Zhang et al., 2016) as well as the fetal types from (Polioudakis et al., 2019). Logistic regression was performed using the model:  $\text{is.cell type} \sim \text{is.module} + \text{gene covariates}$  (GC content and gene length) for an indicator-based enrichment, and p-values were Bonferroni FDR corrected. The top two significant cell types are reported.

#### *Rare variant enrichment*

Genes containing disease associated *de novo* variants were downloaded from denovo-db v1.5, which is a collection of germline *de novo* single nucleotide variants and small indels consolidated across many studies for many disorders including for Autism Spectrum Disorder, Developmental Disorder, Intellectual Disability, and Schizophrenia (<http://denovo-db.gs.washington.edu>; (Turner et al., 2017)). I compiled a list of all genes containing at least one *de novo* mutation of annotation functional class: frameshift, frameshift near splice, splice acceptor, splice donor, start lost, stop gained, stop gained near splice, and stop lost. We then analyzed the union of these genes with genes implicated by CNVs in (Gandal et al., 2018a).

Gandal et al included CNVs that were in at least two studies with a p-value < 0.01 or in one study with a p-value <10<sup>-6</sup>, from well powered studies of over 500 subjects per group. (Girirajan et al., 2013; Glessner et al., 2009; International Schizophrenia, 2008; Levinson et al., 2011; Malhotra and Sebat, 2012; Marshall et al., 2008; McCarthy et al., 2009; Moreno-De-Luca et al., 2010; Moreno-De-Luca et al., 2013; Rujescu et al., 2009; Sanders et al., 2011; Stefansson et al., 2005; Vacic et al., 2011; Weiss et al., 2008). Logistic regression was performed using the model: is.disease ~ is.module + gene covariates (GC content and gene length) for an indicator-based enrichment, and p-values were Bonferroni FDR corrected.

**Table 5.S1 : Rare variant disease associated gene count**

| Disorder                | # Genes with de novo variants | # Genes with CNVs | Total # unique genes |
|-------------------------|-------------------------------|-------------------|----------------------|
| Autism                  | 688                           | 288               | 966                  |
| Developmental Disorder  | 516                           | 0                 | 516                  |
| Intellectual disability | 167                           | 0                 | 167                  |
| Schizophrenia           | 81                            | 289               | 365                  |

### 5.2.c: Module eQTL enrichment

To assess module enrichment of GWAS by gene regulatory regions, I first created a map of gene specific regulatory regions. Every eGene was mapped to its eSNP (1st and 2nd if it had one) and extended to the LD block ( $r^2 > 0.8$  PLINK) calculated in our sample. GWAS p-values for ASD (Grove et al., 2019) and SCZ (Schizophrenia Working Group of the Psychiatric Genomics, 2014) were assigned to modules based on overlapping positions with eSNP LD block annotations. To assess significance, 1000 permutations were performed for each module, randomly selecting the number of annotations corresponding to the number of genes in each module from all eSNP LD block annotations (**Figure 5.S2 A,B,C**). Significance was calculated by the proportion of

permuted observed p-values at the expected p-value of 0.001 that are larger than the actual module's observed p-value at the expected p-value of 0.001.

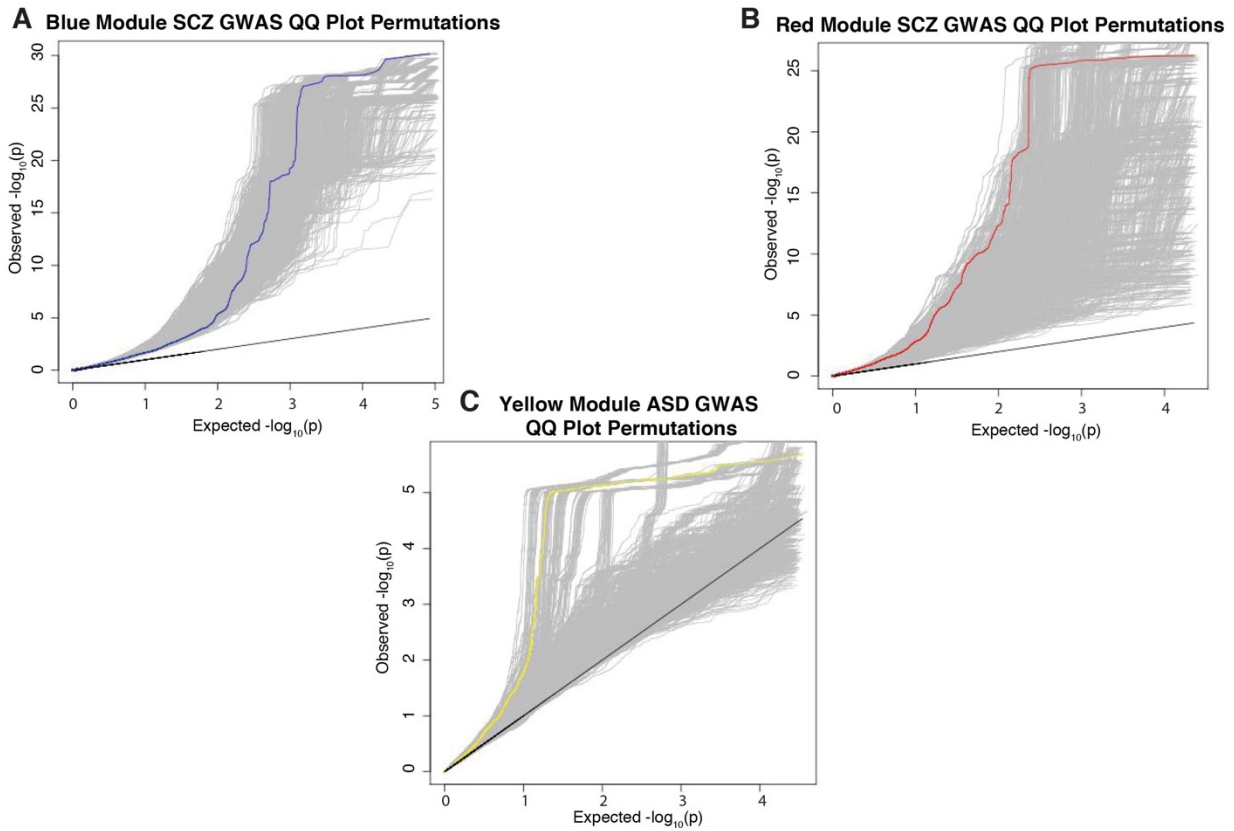


**Figure 5.S2: Permutations of module GWAS enrichment**

**(A-C)** Per module QQ plot of GWAS SNP p-values (SCZ GWAS for blue and red module, ASD GWAS for yellow module) in regulatory regions defined by eQTLs, with 1000 permutations in grey. Expected p-values shown on the X axis and observed p-values shown on the Y axis.

### 5.3: Results

I reasoned that our sample size, which is at least two-fold larger than any previous gene expression study of fetal brain (Gandal et al., 2018b; Kang et al., 2011; O'Brien et al., 2018; Parikshak et al., 2013; Werling et al., 2019), would permit identification of robust sets of co-expressed genes. I applied robust weighted gene co-expression network analysis (Zhang and Horvath, 2005) to construct transcriptional networks, identifying 19 modules (labeled by color) of co-expressed genes during mid-gestation cortical development (**Figure 5.1A**). The modules identified represent genes that correspond to distinct biological functions defined through shared gene ontology (GO) enrichments and cell type markers within the module (**Figure 5.1B and 5.1C**, Methods). Six of these modules are enriched for specific brain cell types or brain-relevant ontological terms: turquoise (fetal mitotic progenitors, cell division), red (fetal mitotic progenitors, outer radial glia, splicing), yellow (superficial layer fetal neurons, splicing), blue (fetal neurons, axon guidance), greenyellow (adult neurons, synaptic transmission, neuron projection development), and brown (adult neurons, CA2+ transport). All of these modules show high preservation in an independent RNA-seq dataset of cortical development from 8 post conception weeks to 12 months after birth (Parikshak et al., 2013; Sunkin et al., 2013). This comparison and replication strongly supports the supposition that this window in mid-gestation, by containing a full range of the major cell types in developing human brain, from proliferating progenitors and post mitotic migrating to post migratory excitatory and inhibitory neurons, captures a substantial portion of the biological processes occurring during prenatal cerebral cortical development (Polioudakis et al., 2019; Pollen et al., 2015).

Fetal brain gene co-expression networks, which define core biological processes occurring during cortical development, are a useful way to identify processes enriched with common risk variants for neurodevelopmental disorders (Parikshak et al., 2015; Parikshak et al., 2013). However, most risk variants lie within noncoding regions, making their assignment to

genes difficult, especially since many regulatory interactions do not involve the closest gene (Won et al., 2016). I used the identified eQTLs to link noncoding variants with target genes and asked whether there were any modules enriched for ASD and SCZ GWAS signal (Methods). I identify significant enrichments for SCZ (blue; p-value = 0.000999) and a marginal trend towards enrichment for ASD (yellow; p-value = 0.061) associated common variants (**Figure 5.1G and 5.1H, Figure 5.S1**). Interestingly, the blue module (**Figure 5.1D**), which shows eQTL GWAS enrichment for SCZ, corresponds to the biological function of neurogenesis as defined by GO analysis (Methods), and includes genes known to play major roles in brain development, such as *DLX1* (essential for GABAergic interneuron production), *FGF2* (anteroposterior neural patterning), *LHX6* (transcriptional regulator of differentiation and development of specific inhibitory neuron classes), and *SMAD1* (progenitor proliferation and differentiation). The human regulatory regions of these developmentally important genes have not been previously defined. The yellow module (**Figure 5.1 F**), which shows suggestive eQTL GWAS enrichment for ASD, corresponds to the biological process of the regulation of gene expression (defined by the GO terms of chromatin organization and mRNA splicing) in fetal neurons, and includes key genes such as *MEF2A* and *MEF2D* (myocyte enhancer factor 2A and 2D transcription factors both involved in neuronal differentiation), *SP2* (transcriptional repressor), *HNRNPH3* (heterogeneous nuclear ribonucleoprotein associated with pre-mRNA processing), and *FOXP4* (transcription regulation in brain development). Remarkably, single cell sequencing data indicates that the genes within this module are also enriched in layer II/III and IV neurons (Polioudakis et al., 2019), consistent with previous data suggesting that ASD associated variation was enriched in superficial cortical layers (Parikshak et al., 2013). This analysis demonstrates the power of eQTLs, when integrated with co-expression modules, to define where common genetic variation associated with a disease acts through regulation of genes with similar biological functions and potentially similar regulatory control (e.g. transcription factors).

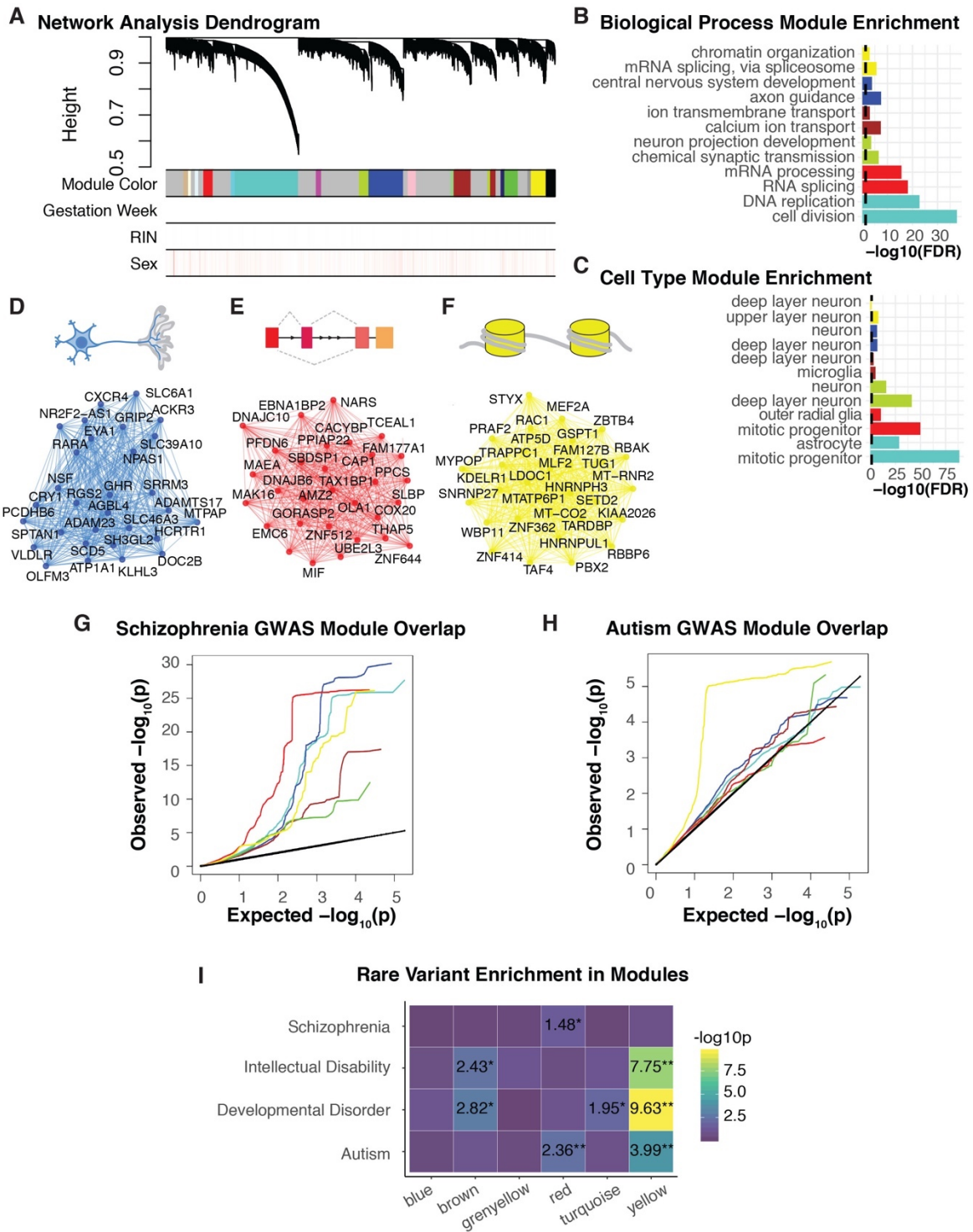


**Figure 5.1: Prenatal Brain Co-expression Networks**

**(A)** Network analysis dendrogram based on hierarchical clustering of genes by their topological overlap, identifies 19 modules. Colored bars below the dendrogram show module membership and expression covariates. Importantly, covariates depicted below the dendrogram are not driving module clustering. **(B)** The top 2 Gene Ontology biological process terms enriched for

each module. The X axis depicts the  $-\log_{10}(\text{FDR})$ , the dotted black line indicates significance  $-\log_{10}(0.05)$  with all listed categories significant. **(C)** The top 2 cell types enriched for genes within each module. The X axis depicts the  $-\log_{10}(\text{FDR})$ , the dotted black line indicates significance  $-\log_{10}(0.05)$ . **(D-F)** Depiction of module GO term biological process summary and top hub genes along with edges supported by co-expression are shown for the blue, red, and yellow module. Hub genes are defined by being the top 30 most connected genes based on kME intermodular connectivity. **(G)** Per module QQ plot of SCZ GWAS SNP p-values in regulatory regions defined by eQTLs. The red module and the blue module show the most inflation for enrichment of SCZ GWAS hits. **(H)** Per module QQ plot of ASD GWAS SNP p-values in regulatory regions defined by eQTLs. The yellow module shows the most inflation for enrichment of ASD GWAS hits. **(I)** Per module rare variant enrichment for SCZ, ASD, ID, Developmental Disorder.

Next, to determine if there was functional overlap in regulatory regions between common GWAS defined annotations above, and rare genetic variation, I examined whether genes harboring rare mutations associated with risk for early onset neurological disease converged on any biological process defined by the fetal co-expression modules (**Figure 5.1 I**). I compiled lists of candidate genes from whole exome sequencing (WES) studies identifying rare and de novo genetic risk variants for ASD, SCZ, intellectual disability, and developmental disorder (Methods; (De Rubeis et al., 2014; Fromer et al., 2014; Girirajan et al., 2013; Glessner et al., 2009; Gulsuner et al., 2013; International Schizophrenia, 2008; Iossifov et al., 2014; Krumm et al., 2015; Levinson et al., 2011; Malhotra and Sebat, 2012; Marshall et al., 2008; McCarthy et al., 2014; McCarthy et al., 2009; Michaelson et al., 2012; Moreno-De-Luca et al., 2010; Moreno-De-Luca et al., 2013; O'Roak et al., 2014; O'Roak et al., 2012; Rujescu et al., 2009; Sanders et al., 2011; Stefansson et al., 2005; Tavassoli et al., 2014; Turner et al., 2016; Turner et al., 2017; Vacic et al., 2011; Weiss et al., 2008; Yuen et al., 2015)). Interestingly, I find only the red module enriched for rare variation for SCZ (nominal p-value 0.037), which is also trending towards enrichment in common variation from SCZ GWAS (**Figure 5.1 E**). The yellow module also exhibits the most significant enrichment for rare variation for ASD (FDR = 0.011), also corresponding to the top module enriched in common variation from the ASD GWAS. These

analyses provide convergent statistical evidence for overlap in the genes and pathways impacted by common and rare genetic variation in these disorders for the first time.

## **CHAPTER SIX**

Transcriptome-wide association studies

## 6.1: Introduction

Transcriptome-wide association studies (TWAS) is a powerful method that integrates gene expression levels with summary statistics from genome-wide association studies to identify genes whose expression, imputed from cis- genetic variants, is associated with the disease of interest. Using fetal brain expression and genotype data described in Chapter 2, I have calculated both gene level and splicing level expression weights for genes with significant cis-SNP heritability by using a linear regression model using all cis-SNPs in a 1MB region around the gene. Additionally, I created adult brain expression weights as a part of the PsychENCODE project following the same method. Weights were used to impute expression into the largest SCZ, ICV, and ASD GWAS for both fetal and adult weights, which were then tested for association with each disease. I have compared the fetal TWAS identified genes with previously published adult brain expression weights TWAS, including PsychENCODE, examined the overlap between mendelian-randomization prioritized genes, as well as characterized significant gene's expression trajectories using BrainSpan.

## 6.2: Methods

I performed SCZ, ASD, and intercranial volume TWASs using the FUSION package (<http://gusevlab.org/projects/fusion/>) with gene and splice expression measured in fetal brain tissue and adult brain tissue.

### 6.2.a: Prenatal brain weights

To identify genes with evidence of genetic control, I used GCTA (Yang et al., 2011) software to estimate cis-SNP heritability  $h^2_g$  (+-1MB window around gene TSS) of all fetal expressed genes. We identified 3,784 genes and 5,738 splicing-events with significant cis-  $h^2_g$  (nominal p-value < 0.05), which were used to calculate the SNP-based predictive weights per gene/intron.

### 6.2.b: PsychEncode adult brain weights

The PsychENCODE consortium dataset consists of post-mortem human brain samples collected as a part of eight studies. These consist of RNA sequencing from BrainGVEX, BrainSpan, CommonMind, Yale-ASD, UCLA-ASD, CMC\_HBCC, BipSeq, and LIBD\_czControl aggregating together 1,321 samples which are well described in (Gandal et al., 2018b).

Using the fusion package, I generated custom SNP-expression weights generated from brain using the PsychENCODE dataset of 1321 unique individuals with imputed genotypes. Using the AI-REML algorithm implemented in GCTA (Yang et al., 2011) by the fusion package, we first identified the subset ( $n=14,750$ ) of total expressed genes found to have significant *cis* SNP-heritability in our dataset (*cis*-  $h^2_g$ ,  $P<0.05$  within 1 Mb window around the gene body). SNP-expression weights were calculated in a 1Mb region around all heritable genes using expression measurements adjusted for diagnosis, study, age, age<sup>2</sup>, RIN, RIN<sup>2</sup>, sex, tissue, PMI, 20 ancestry PCs, and 100 hidden covariates (Mostafavi et al., 2013). Accuracy of five expression prediction models were tested (best *cis*-eQTL, best linear unbiased predictor, Bayesian linear mixed model, Elastic-net regression, LASSO regression) using the most accurate model for final weight calculations as implemented in fusion. TWAS disease-association statistics were computed using these custom weights, LD structure calculated from our PsychENCODE samples' genotypes, and disease GWAS summary statistics for SCZ (Pardinas et al., 2018) and ASD (Grove et al., 2019). For each disease, TWAS association statistics were Bonferroni-corrected for multiple comparisons. At loci ( $\pm 100$  kb) with multiple significant associations, joint and conditional association analyses were further performed as implemented in the FUSION.post\_process.R script.

### 6.2.c: *Published adult weights*

Published weights were downloaded from (<http://gusevlab.org/projects/fusion/>) for CommonMind adult brain expression and splicing, as well as GTEx adult cortex expression and whole blood expression.

### 6.2d: *TWAS*

Using the FUSION package, five-fold cross-validation of five models of expression prediction (best cis-eQTL, best linear unbiased predictor, Bayesian sparse linear mixed model (BSLMM) (Zhou et al., 2013), Elastic-net regression, LASSO regression) were calculated and evaluated for accuracy. The model with largest cross-validation  $R^2$  was chosen for downstream association analyses.

TWAS statistics were calculated using fetal weights, as well as published adult expression weights (Gusev et al., 2018) calculated from the CommonMind Consortium eQTL dataset for gene and intron level weights (Fromer et al., 2016) and GTEx Cortex expression gene weights (<http://gusevlab.org/projects/fusion/>), LD SNP correlations from the 1000 Genomes European Phase 3 reference panel (as the GWAS used are from European populations) (The 1000 Genomes Project Consortium et al., 2015), and GWAS summary statistics from CLOZUK +PGC2 SCZ GWAS (105,318 individuals) (Pardinas et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics, 2014) and CHARGE ENIGMA Meta-Analysis ICV GWAS (26,577 individuals) (Adams et al., 2016). TWAS association statistics were Bonferroni corrected per GWAS, gene and intron, separately. Additional Fusion parameters include running colocalization by COLOC (--coloc\_P 0.05) and increasing the maximum fraction of missing SNPs to be imputed due to ancestry differences between the fetal eQTLs and the GWAS (--max\_impute 0.06).

#### *6.2.e: High confidence adult brain implicated genes*

For the SCZ TWAS, we created a set of high confident adult brain TWAS gene associations, designated by being significant in 2 of 3 adult brain datasets (CommonMind Consortium, GTEx Cortex, and PsychENCODE). The CommonMind Consortium, GTEx Cortex weights were run with the CLOZUK+ PCG SCZ GWAS and PsychENCODE SCZ TWAS results were downloaded from Gandal et al. (Gandal et al., 2018b)

#### *6.2.f: TWAS overlap with GWAS loci*

Overlap of TWAS hits ( $\pm 500\text{kb}$ ) with GWAS significant loci (LD block reported in paper) was assessed for both fetal TWAS hits and CommonMind Consortium, GTEx cortex, and PsychENCODE adult prefrontal cortex TWAS hits. Novel regions were identified by genes and introns ( $\pm 500\text{k}$ ) that did not overlap a GWAS significant loci, and then regions were merged if the gene/intron  $\pm 500\text{kb}$  overlapped.

#### *6.2.g: Summary-based mendelian randomization*

I performed summary based mendelian randomization (SMR) (Zhu et al., 2016) to identify gene-trait associations by integrating eQTL and GWAS data, as a complementary approach to TWAS. All nominal snp-gene pair fetal cis-eQTLs and CLOZUK+ PGC SCZ GWAS (Pardinas et al., 2018) were used as input to SMP. The set based SMR test was run (`--smr-multi`) using SNPs in the cis window with a  $P_{\text{eqtl}} < 10^{-4}$  (`--peqtl-smr 1e-4`) and the associated HEIDI test. For replication of significant TWAS hits, we require genes to have a significant SMR association ( $P_{\text{SMR}} < 0.05$ ) and HEIDI test ( $P_{\text{HEIDI}} < 0.05$ ).

#### *6.2.h: TWAS fine-mapping*

TWAS association statistics at genomic risk regions tend to be correlated as a function of linkage and eQTL overlap between predictive models (Mancuso et al., 2019; Wainberg et al.,

2017). To prioritize candidate susceptibility genes in our TWAS we performed statistical fine-mapping using FOCUS (Mancuso et al., 2019). FOCUS models the correlation structure induced by LD and overlapping eQTL weights across predictive models and computes posterior probabilities for a gene/intron to explain all observed TWAS association signal at a region. To account for missing gene predictions models, I include gene expression prediction models for genes for genes not predictable from the fetal brain dataset across 45 tissues measured in 47 reference panels and include the model with the best accuracy among the other tissues. I next computed 90% credible gene-sets by taking genes with largest posterior probability until 90% density was explained.

## **6.3: Results**

### *6.3.a: Schizophrenia*

As a part of PsychENCODE, we performed a formal transcriptome-wide association study (TWAS;(Gusev et al., 2016)) to directly identify those genes whose *cis*-regulated expression is associated with disease. TWAS and related methods have the advantage of aggregating the effects of multiple SNPs onto specific genes, reducing multiple comparisons and increasing power for association testing, although results can still be influenced by LD and pleiotropy (Gusev et al., 2016). Further, by imputing the *cis*-regulated heritable component of brain gene expression into the association cohort, TWAS enables direct prediction of the transcriptomic effects of disease-associated genetic variation, identifying potential mechanisms through which variants may impart risk. However, the limited size of brain eQTL datasets to date has necessitated use of non-CNS tissues to define TWAS weights. Given the enrichment of psychiatric GWAS signal within CNS expressed regulatory elements (Finucane et al., 2015), we reasoned that the *PsychENCODE* dataset would provide substantial power and specificity. Indeed, we identify 14,750 genes with heritable *cis*-regulated expression in brain in the

PsychENCODE cohort, enabling increased transcriptomic coverage for detection of association signal.

By integrating PsychENCODE expression weights with SCZ GWAS, TWAS identifies 193 genes of which 107 remain significant following conditional analysis at each gene within multi-hit loci. Excluding the MHC region, there remained 164 significant genes representing 78 genome-wide significant GWAS loci (**Fig 6.1**). Of the 107 conditionally significant genes prioritized by TWAS, 62 were further supported by SMR ( $P_{\text{SMR}} < 0.05$ ,  $P_{\text{HEIDI}} > 0.05$ ) and 11 were also concordantly DE in SCZ brain in the same direction as predicted by TWAS. Altogether, 64 genes were consistently prioritized across multiple methods. These included a number of intriguing novel candidates for SCZ: two downregulated lysine methyltransferases (*SETD6*, *SETD8*); *RERE*, a downregulated, mutationally intolerant nuclear receptor co-regulator of retinoic acid signaling associated with a rare neurodevelopmental genetic syndrome; *LINC00634*, a downregulated poorly annotated brain-enriched lincRNA; and *SLC12A5*, encoding a mitochondrial  $\text{Ca}^{2+}$  binding aspartate/glutamate carrier protein, associated with a recessive epileptic encephalopathy. Overall, this analysis provides a core set of strong candidate genes implicated by risk loci, and provides a mechanistic basis for the composite activity of disease risk variants.

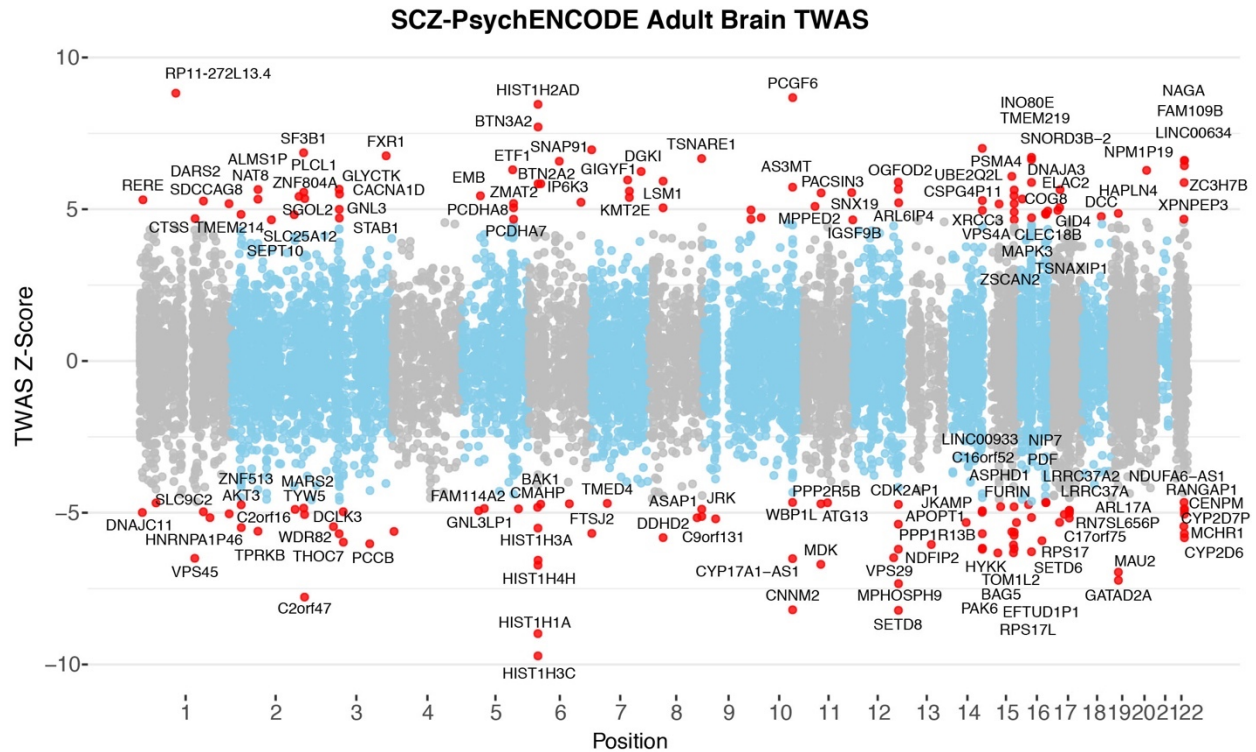


**Figure 6.1: SCZ-PsychENCODE Adult Brain TWAS**

Results from TWAS prioritize genes whose *cis*-regulated expression in brain is associated with disease. The plot shows 193 genes (164 outside of MHC) are prioritized by TWAS at Bonferroni-corrected  $P < 0.05$ , including 107 genes with conditionally independent signals. The sign of TWAS Z-scores indicates predicted direction of effect.

To further leverage these data to refine and characterize neuropsychiatric disease loci with developmental origins, I used an imputation-based transcription-wide association study (TWAS; (Gusev et al., 2016), to integrate *cis*-eQTLs and GWAS loci to identify genes whose fetal brain expression is correlated with schizophrenia. Previous TWAS studies in SCZ both relied on adult brain or non-brain tissues and did not have access to fetal brain eQTL (Gandal et al., 2018b; Gusev et al., 2018). Given the evidence for a neurodevelopmental origin of these disorders (de la Torre-Ubieta et al., 2018; Gulsuner et al., 2013), I reasoned that these fetal data

would provide new perspectives on SCZ risk, especially given the sensitivity of TWAS to tissue source (Wainberg et al., 2017).

I used the summary statistics from CLOZUK + Psychiatric Genomics Consortium (PGC) schizophrenia (SCZ) GWAS meta-analysis (Pardinas et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics, 2014) consisting of 105,318 individuals and our fetal brain eQTL data set to identify genes and splicing-events whose imputed cis-regulated expression is associated with SCZ. Given the expression weights, CLOZUK+PGC SCZ GWAS Z-scores, and linkage-disequilibrium (LD) reference panel, I computed TWAS statistics (Methods). Of the 3,145 genes and 4,796 introns with significant cis-heritability, I identified 62 gene-SCZ and 91 intron-SCZ (corresponding to 65 unique genes) significant transcriptome-wide associations at a Bonferroni corrected p-value <0.05 (**Figure 6.2 A**). Five of these genes, *SF3B1*, *GNL3*, *NT5DC2*, *SMDT1*, and *RALGAPA2* harbor both a transcription-level and splicing-level association.

For orthogonal validation, I ran Summary-data-based Mendelian Randomization (SMR) and the associated HEIDI test (Zhu et al., 2016) to test for pleiotropic association in the cis window of eQTL associations to distinguish pleiotropy from linkage. Six SCZ TWAS significant genes and five TWAS significant introns are significant across both methods with a SMR p-value <0.05 and HEIDI p-value <0.05 (**Table 6.S1**). With support from two independent methods, TWAS and SMR, these genes and introns provide a prioritized set of targets for further investigation.

**Table 6.S1: Fetal brain-SCZ genes significant across TWAS and SMR**

| <b>Dataset</b> | <b>Chr</b> | <b>Gene</b>   | <b>GWAS<br/>p-value</b> | <b>eQTL<br/>p-value</b> | <b>SMR<br/>p-value</b> | <b>HEIDI<br/>p-value</b> |
|----------------|------------|---------------|-------------------------|-------------------------|------------------------|--------------------------|
| Gene           | 10         | C10orf32      | 3.68E-02                | 6.98E-25                | 4.07E-02               | 7.19E-06                 |
| Gene           | 11         | ARL14EP       | 4.11E-08                | 2.22E-12                | 1.54E-05               | 3.74E-02                 |
| Gene           | 16         | YPEL3         | 2.29E-05                | 5.50E-12                | 0.00030881             | 0.026379                 |
| Gene           | 22         | RP1-257I20.14 | 9.99E-01                | 2.72E-08                | 0.99935                | 0.00184825               |
| Gene           | 2          | TYW5          | 2.91E-08                | 2.44E-21                | 1.69E-06               | 0.00118295               |
| Gene           | 2          | C2orf47       | 9.29E-17                | 3.37E-07                | 1.37E-05               | 0.0409299                |
| Intron         | 12         | MPHOSPH9      | 7.84E-11                | 6.56E-17                | 2.87E-07               | 0.01078895               |
| Intron         | 1          | SDCCAG8       | 4.73E-07                | 9.28E-11                | 6.99E-05               | 0.001823                 |
| Intron         | 22         | NDUFA6-AS1    | 9.99E-01                | 4.49E-09                | 0.99935                | 0.00223602               |
| Intron         | 5          | CDC25C        | 3.99E-01                | 2.77E-08                | 4.04E-01               | 0.00138174               |
| Intron         | 5          | FAM53C        | 8.21E-04                | 7.60E-06                | 0.00736905             | 0.0413288                |

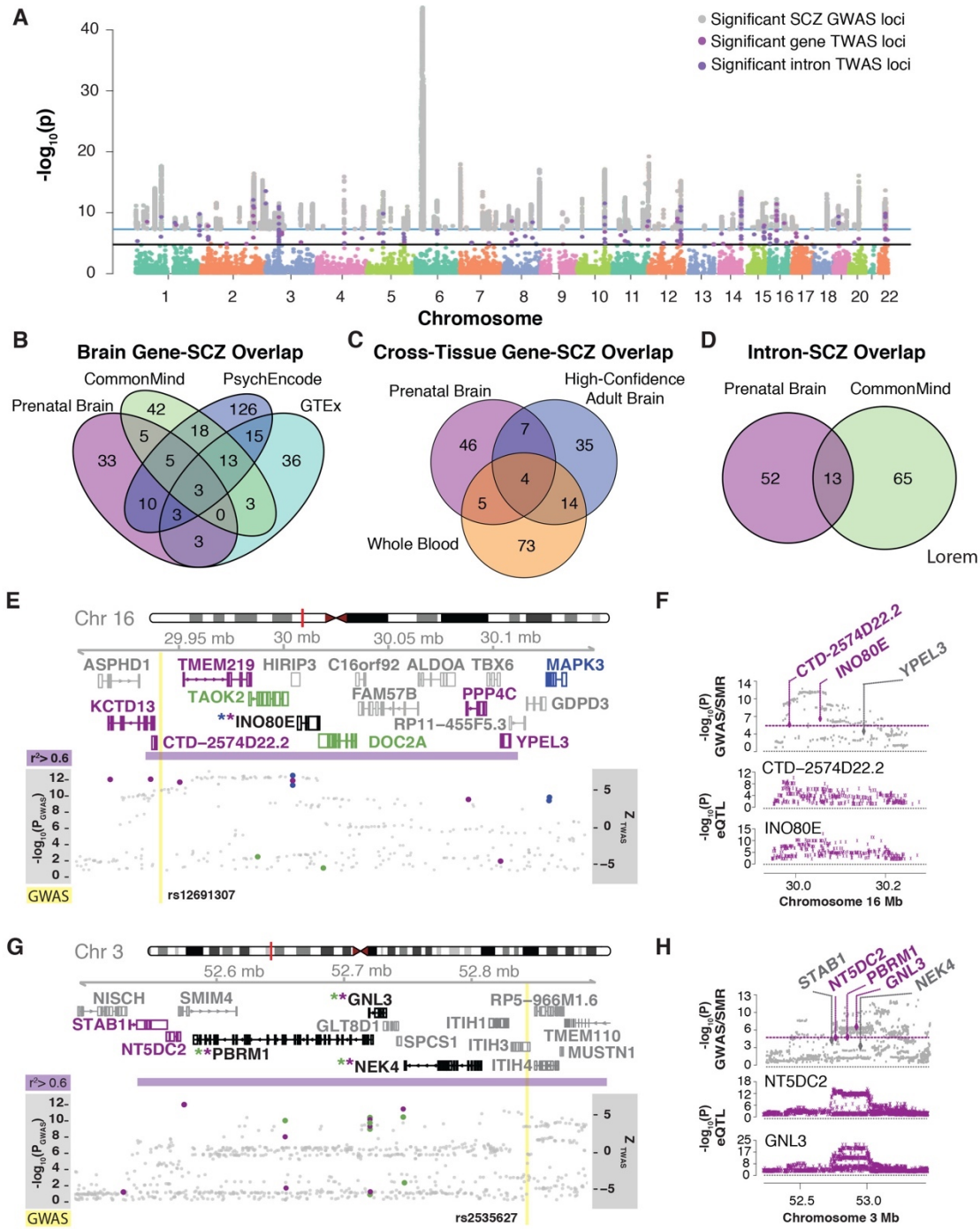


**Figure 6.2: SCZ TWAS**

(A) Manhattan plot of TWAS results from genes and intron SCZ associations, highlighting the 62 genes and 91 introns significant at  $P_{\text{bonferroni}} < 0.05$ , overlaid with the SCZ GWAS significant loci. Key depicts significant GWAS loci (grey), TWAS genes (magenta), and introns (purple). (B) Overlap of TWAS-significant genes from prenatal brain, adult brain GTEx, adult brain CMC, and adult brain PsychENCODE. (C) Overlap of TWAS genes from prenatal brain, high-confident adult brain and whole-blood. (D) Overlap of TWAS introns from prenatal brain and adult brain. (E, G) Illustration of two genomic regions harboring at least one prenatal TWAS loci. The Y axis

shows Ensembl gene ID, LD, TWAS results overlapped with GWAS loci  $-\log_{10}(P)$ ; prenatal brain TWAS genes and introns are magenta dots, high-confident adult brain TWAS genes are blue dots, adult brain TWAS introns are green dots, and GWAS SNPs are grey dots) and SCZ GWAS loci (yellow line). Ensembl gene names and gene models are colored by significance in datasets; prenatal brain gene and intron associations (magenta), adult brain intron associations (green), high-confident adult brain gene associations (blue). If a gene is significant in both prenatal and adult it is black with an asterisk colored by the dataset. SCZ GWAS loci are shown in yellow, with the LD block surrounding the SCZ locus in purple defined by  $r^2 > 0.6$ . (**F, H**) Summary-data based mendelian randomization of prenatal brain genes with SCZ GWAS. The Y axis shows  $-\log_{10}(P)$  of SMR and GWAS loci and individual gene eQTL  $-\log_{10}(P)$ . Genes are colored by passing the 5% FDR threshold for  $P_{\text{SMR}}$ .

I next was interested in comparing the overlap in SCZ candidate risk genes implicated by either fetal or adult brain TWAS. I first identified 60 high-confidence adult brain SCZ TWAS implicated genes, designated by significance in at least 2 of 3 adult brain datasets (GTEx Cortex, CommonMind, and PsychENCODE; **Figure 6.2 B**, Methods). Eleven of the high confidence adult associations overlapped with fetal brain gene-SCZ associations (binomial test  $p\text{-value} < 2.2\text{e-}16$ ; *SNX19*, *VSP29*, *XRCC3*, *TSNAXIP1*, *DNAJA3*, *INO80E*, *NAGA*, *SF3B1*, *TYW5*, *C2orf47*, and *DDHD2*). These TWAS associations significant across the multiple adult brain data sets and fetal brain implicate genes that are expressed throughout development and may impute risk across developmental stages.

I highlight three of these genes: INO80 complex subunit E (*INO80E*), splicing factor 3b subunit 1 (*SF3B1*) and matrix AAA peptidase interacting protein 1 (*MAIP1*; *C2orf47*). *INO80E* is a component of a chromatin remodeling complex involved nucleosome spacing and modulates transcriptional regulation during corticogenesis (Ayala et al., 2018; Sokpor et al., 2018). *SF3B1* is within a SCZ GWAS locus, and is supported by an animal model of psychosis (Ingason et al., 2015). *MAIP1* (*C2orf47*) plays a role in mitochondrial  $\text{Ca}^{2+}$  handling and cell survival and was previously associated with SCZ in a large Swedish GWAS (Ripke et al., 2013). As a comparison, we performed a TWAS incorporating GTEx whole blood expression weights with the SCZ GWAS. I find that only 4 genes overlap between prenatal brain, high confidence adult

brain and whole blood SCZ transcription-level associations (binomial test p-value =  $2.548e-14$ ; **(Figure 6.2 C)**).

At the level of splicing, I identify 13 genes with intron-SCZ associations that overlapped between the 91 fetal brain intron-SCZ associations and the 120 adult brain intron-SCZ associations (**Figure 6.2 D**). Interestingly, the intron level TWAS highlights more splicing differences between developmental time points associated with SCZ than the gene level, with about 10% overlap, consistent with previous indications that splicing may play a larger role in disease biology than gene expression alone (Gandal et al., 2018b). I highlight two genes whose splicing is implicated by both fetal and adult brain: adaptor related protein complex 3 subunit beta 2 (*AP3B2*) and M-phase Phosphoprotein 9 (*MPHOSPH9*). *AP3B2* is a component of the neuron-specific AP-3 complex and has been shown to interact with dysbindin-1, a schizophrenia- related protein, through direct binding to the AP-3 complex through *AP3B2* (Hashimoto et al., 2009; Oyama et al., 2009). Additionally, mutations in *AP3B2* have been linked to early-onset epileptic encephalopathy (Assoum et al., 2016). *MPHOSPH9* has been associated with SCZ through differential expression at the exon level in postmortem brain in both Brodmann Area 10 and the caudate nucleus, implicating splicing abnormalities of *MPHOSPH9* to be associated with SCZ (Cohen et al., 2012).

A major advantage of TWAS is the identification of candidate genes that did not pass genome-wide significance based on GWAS alone. To identify new risk regions for SCZ, we next examined the overlap between the 145-independent genome-wide significant SCZ GWAS loci (Pardinas et al., 2018) and the above significant fetal brain TWAS associations. We find 50 GWAS loci harboring fetal TWAS associations and 23 additional SCZ risk loci where significant TWAS hits do not overlap a GWAS significant loci (Methods). Of these 23 newly identified loci, 18 regions were also identified by a high confident adult brain TWAS association, providing

additional support for their putative roles in SCZ (Methods). Five of these new loci were unique to fetal brain.

To further dissect causality of the TWAS associations that overlap with previously identified GWAS regions, I applied FOCUS, a fine-mapping approach that refines TWAS gene-trait associations by accounting for the correlation of linkage disequilibrium and SNP weights (Mancuso et al., 2019). FOCUS was run using a tissue-prioritized approach which includes gene weights from a multi-tissue reference panel for genes not predictive in the fetal brain dataset (Methods). I identified a credible set of 155 genes, with an average of 3 genes fine-mapped at each GWAS locus, largely reflecting local patterns of LD, and indicative of the need for further experimental dissection at many of these complex loci.

A salient example of fetal and adult TWAS overlap within a known GWAS region is on chromosome 16. This region is complex, containing 6 fetal associations and 4 adult associations, with only one gene in common to both stages, *INO80E* (**Figure 6.2 E**). *INO80E* shows both significant co-localization of its fetal brain eQTL and SCZ GWAS in this region by SMR ( $p_{\text{SRM } INO80E} = 3.7 \times 10^{-6}$ , **Figure 6.2 F**) and is identified as a casual gene by FOCUS fine-mapping (Methods; Table S8), making it a high confidence locus identified in both tissues by both methods. *TMEM219* is the only other gene in the credible set fine-mapped to this region and is specific to the fetal time point. Similarly, another SCZ GWAS locus on chromosome 3 harbors three splicing associations that are implicated by both fetal and adult TWAS (*PBRM1*, *GLN3*, and *NEK4*). This locus also contains two other genes that are implicated by fetal brain expression only, *STAB1* and *NT5DC2* (**Figure 6.2 G**). Co-localization of fetal brain eQTLs by SMR supports *NT5DC2*, *GLN3* and *PBRM1* at this locus ( $p_{\text{SRM } NT5DC2} = 1.4 \times 10^{-6}$ ,  $p_{\text{SRM } GLN3 \text{ intron}} = 2.9 \times 10^{-6}$ ,  $p_{\text{SRM } PBRM1 \text{ intron}} = 3.0 \times 10^{-6}$ , **Figure 6.2 H**), providing multiple lines of evidence linking fetal expression changes in *NT5DC2*, and splicing of *GLN3* and *PBRM1* at both

developmental time points to SCZ risk. In summary, I find convergent evidence supporting *INO80E*, *GLN3*, *PBRM1*, which are implicated by data from both fetal and adult stages. I additionally uncover fetal specific associations, such as *TMEM219* and *NT5DC2*, which highlights the utility of using disease relevant tissues, which for SCZ, includes developing brain (de la Torre-Ubieta et al., 2018; Won et al., 2016) in addition to adult tissue.

### 6.3.b: Intracranial volume

Given the importance of cortical neurogenesis, which peaks during mid-gestation (Kang et al., 2011; Miller et al., 2014), as a major driver of brain evolution and brain size (Bae et al., 2015; Geschwind and Rakic, 2013; Kostovic and Jovanov-Milosevic, 2006; Rakic, 1995, 2009), I reasoned that these developing cortex eQTL and sQTL would be particularly valuable in defining loci involved in intracranial volume in humans. This is also supported by analysis of open chromatin, which supports the strong genome-wide predominance of fetal over adult effects on intracranial volume (de la Torre Ubieta et al. 2018). By using fetal brain expression weights, I could specifically highlight genetic association signals acting during a time frame crucial to the expansion of the cerebral cortex (de la Torre-Ubieta et al., 2018).

I leveraged the GWAS summary statistics of intracranial volume, a meta-analysis of Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) and Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortia consisting of 26,577 individuals, to perform a transcription-wide association study, as intracranial volume is directly related to brain growth early in brain development (Adams et al., 2016). Of the 3,215 genes and 4,932 introns with significant cis-heritability, we identified 7 genes whose expression (*NSF*, *LRRC37A*, *LRRC37A2*, *LRRC37A17P*, *RNF123*, *RP11-156P1.3*, *MAPT-AS1*) and 8 introns whose splicing (*NT5C2*, *CRHR1*, *LRRC37A*, *LRRC37A2*, *USMG5*, *KANSL1*, *USP4*) was significantly associated with intracranial volume (Bonferroni corrected p-value <0.05/3,215 for

genes and p-value  $<0.05/4,932$  for introns; **Figure 6.3 A**, Methods). For comparison, I conducted TWAS using the published CMC adult prefrontal cortex and GTEx adult cortex (Gusev et al., 2018; The GTEx Consortium et al., 2017) with the ENIGMA-CHARGE ICV GWAS (Adams et al., 2016). *LRRC37A2* is the only gene overlaps between the fetal implicated eGenes and the 16 significant GTEx adult cortex implicated genes (**Figure 6.3 B**). Three genes (*CRHR1*, *LRRC37A2* and *USMG5*) with intron-ICV associations overlapped between the 8 fetal brain intron-ICV associations and 10 adult brain intron-ICV associations (**Figure 6.3 C**).

I compared the overlap of significant GWAS regions with significant TWAS associations to identify any new regions harboring genetic modifiers of ICV (Methods). I find 13 out of the 15 fetal significant associations fall within 2 of the 17 significant ENIGMA-CHARGE ICV GWAS loci ( $\pm 500\text{kb}$ ). The remaining two fetal associations (*RNF123* and *USP4*) lie within close proximity to each other on chromosome 3, revealing a new locus associated with ICV. We identify 33 adult brain significant genes and introns (GTEx and CMC), 30 of which overlap with known ICV loci, and 3 of which represent novel loci. One of these novel loci, *GPX1*, overlaps the new locus identified from the fetal brain dataset on chromosome 3, while *IGFBP2* and *CTD-2292M16.8* reveal adult specific regions on chromosomes 2 and 14, respectively.

One particularly interesting GWAS region lies on chromosome 17 where several significant TWAS associations are found (6 fetal brain genes, 3 fetal brain introns, 13 adult brain genes, and 6 adult brain introns) (**Figure 6.3 D**). This complex region contains a large structural chromosomal rearrangement on chromosome 17q21 that includes *NSF*, *LRRC37A*, *LRRC37A2*, and has been associated with several brain related diseases, including schizophrenia, dyslexia, and Parkinson's Disease (Chen et al., 2016a; Latourelle et al., 2012; Veerappa et al., 2014). Fine-mapping of the TWAS associations at this region prioritizes *LRRC37A*, which is the only gene that is significant at both developmental time points (Methods, **Figure 6.3 E**)(BrainSpan,

2013). The only other ENIGMA-CHARGE ICV GWAS locus overlapping a fetal brain TWAS association is on chromosome 10 (**Figure 6.3 F**), which contains significant splicing associations in *NT5C2* and *USMG5*, the latter also harboring an adult brain splicing association. *USMG5 (ATP5MD)* is a subunit of the mitochondrial ATP synthase (Ohsakaya et al., 2011), in which mutations cause Leigh syndrome (Barca et al., 2018). *NT5C2* is a hydrolase that plays a role in cellular purine metabolism that has been linked to intellectual disability and spastic paraplegia (Darvish et al., 2017), and shows greater prenatal than postnatal expression (**Figure 6.3 G**)(BrainSpan, 2013).

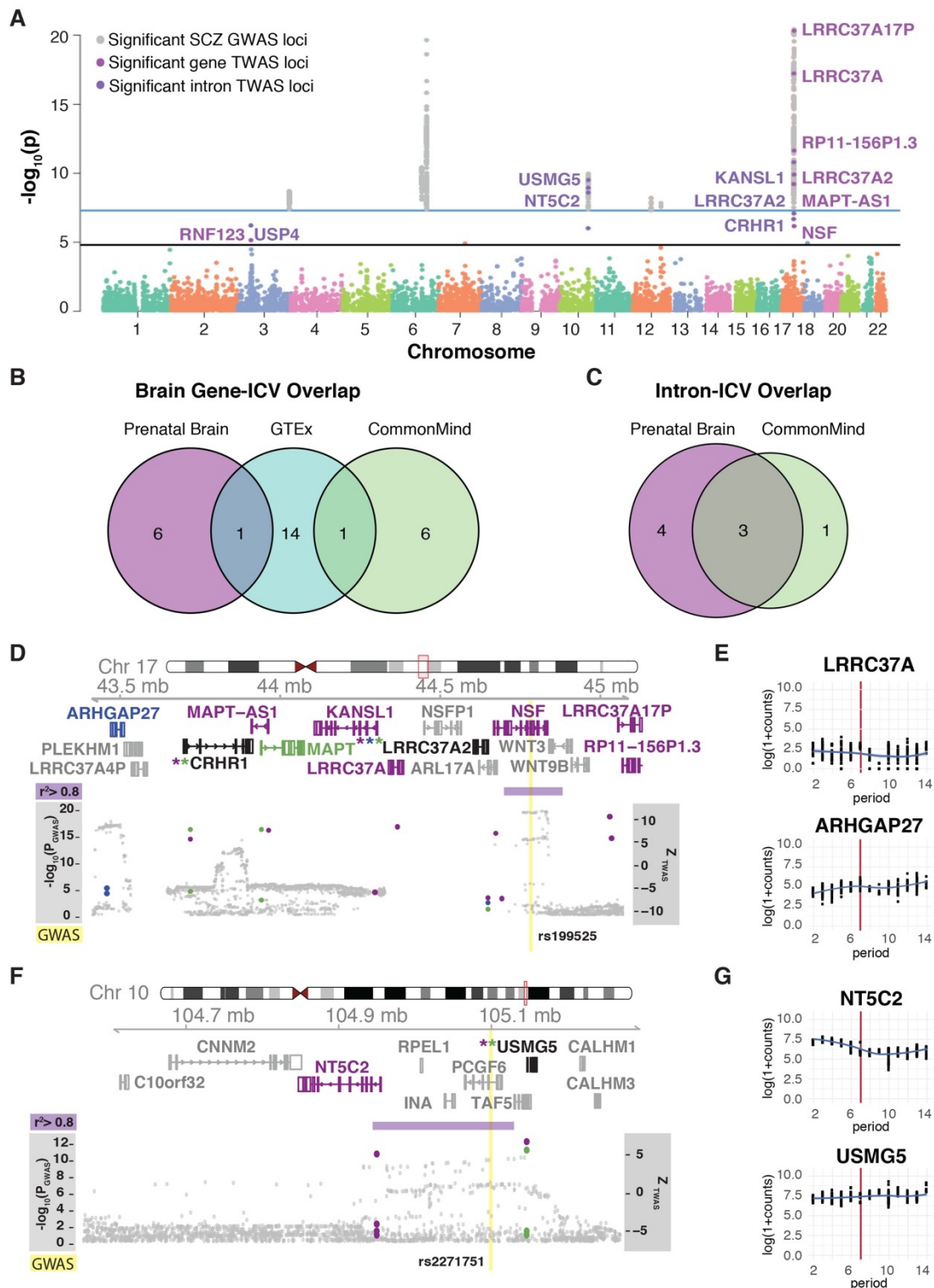


**Figure 6.3: ICV TWAS**

**(A)** Manhattan plot of TWAS results from genes and intron ICV associations, highlighting the 7 genes and 8 introns significant at  $P_{\text{bonferroni}} < 0.05$ , overlaid with ICV GWAS significant loci. Key depicts significant GWAS loci (grey), prenatal brain TWAS genes (magenta) and introns

(purple). **(B)** Overlap of TWAS genes from prenatal brain, adult brain from GTEx, and adult brain from CMC. **(C)** Overlap of TWAS introns from prenatal brain and adult brain from CMC (Gusev et al., 2018). **(D, F)** Illustration of two genomic regions harboring at least one prenatal TWAS loci. The Y axis shows Ensembl gene ID, LD, TWAS results overlapped with GWAS loci -  $\log_{10}(P)$ ; prenatal brain TWAS genes and introns are magenta dots, high-confident adult brain TWAS genes are blue dots, adult brain TWAS introns are green dots, and GWAS SNPs are grey dots) and ICV GWAS loci (yellow line). Ensembl gene names and gene models are colored by significance in datasets; prenatal brain gene and intron associations (magenta), adult brain intron associations (green), high-confident adult brain gene associations (blue). If the gene is implicated by more than one dataset there is an asterisk colored by the datasets. The ICV GWAS locus is shown in yellow with the LD block in purple defined by  $r^2 > 0.8$ . **(E, G)** Corresponding BrainSpan(BrainSpan, 2013) gene expression trajectories throughout human life span of highlighted TWAS genes. X-axis corresponds to developmental periods defined in (Kang et al., 2011), where the red line represents period 8, corresponding to birth.

### 6.3.c: Autism Spectrum Disorder

By integrating PsychENCODE expression weights with ASD GWAS, TWAS prioritizes 12 genes across 3 genomic loci (Bonferroni-corrected  $P < 0.05$ ; **Figure 6.4 A**). This includes the 17q21.31 region, which did not reach genome-wide significance in the largest GWAS to date (Grove et al., 2019). Of the seven TWAS-significant genes at this locus, conditional analysis prioritizes one – *LRRC37A*, which is further supported by SMR (Zhu et al., 2016) . *LRRC37A* is intriguing due to its primate-specific evolutionary expansion, loss-of-function intolerance, and expression patterns in brain and testis (Zody et al., 2008). However, common variants in GWAS are also likely tagging the common inversion and other recurrent structural variants present at this locus (Zody et al., 2008). TWAS additionally prioritizes genes on chromosomes 8 and 20 (**Figure 6.4 A**).

I also integrated the fetal brain expression weights with ASD GWAS (Grove et al., 2019) and identify only one significant gene, *XRN2*, at Bonferroni corrected p-value  $< 0.05$  (**Figure 6.4 B**). *XRN2* lies on chromosome 20 and overlaps the locus identified by PsychENCODE significant genes *NKX2-2* and *PLK1S1*. However, *XRN2* was implicated by a recently published ASD TWAS using multiple expression data sets both brain and non-brain, as one of five

significant TWAS implicated genes (*PDIA6*, *RP11-10A14.3*, *CTSB*, *RP11-259G18.1*, and *XRN2*)(Pain et al., 2019), though it was not identified by the independent fetal brain dataset included in their study (O'Brien et al., 2018).

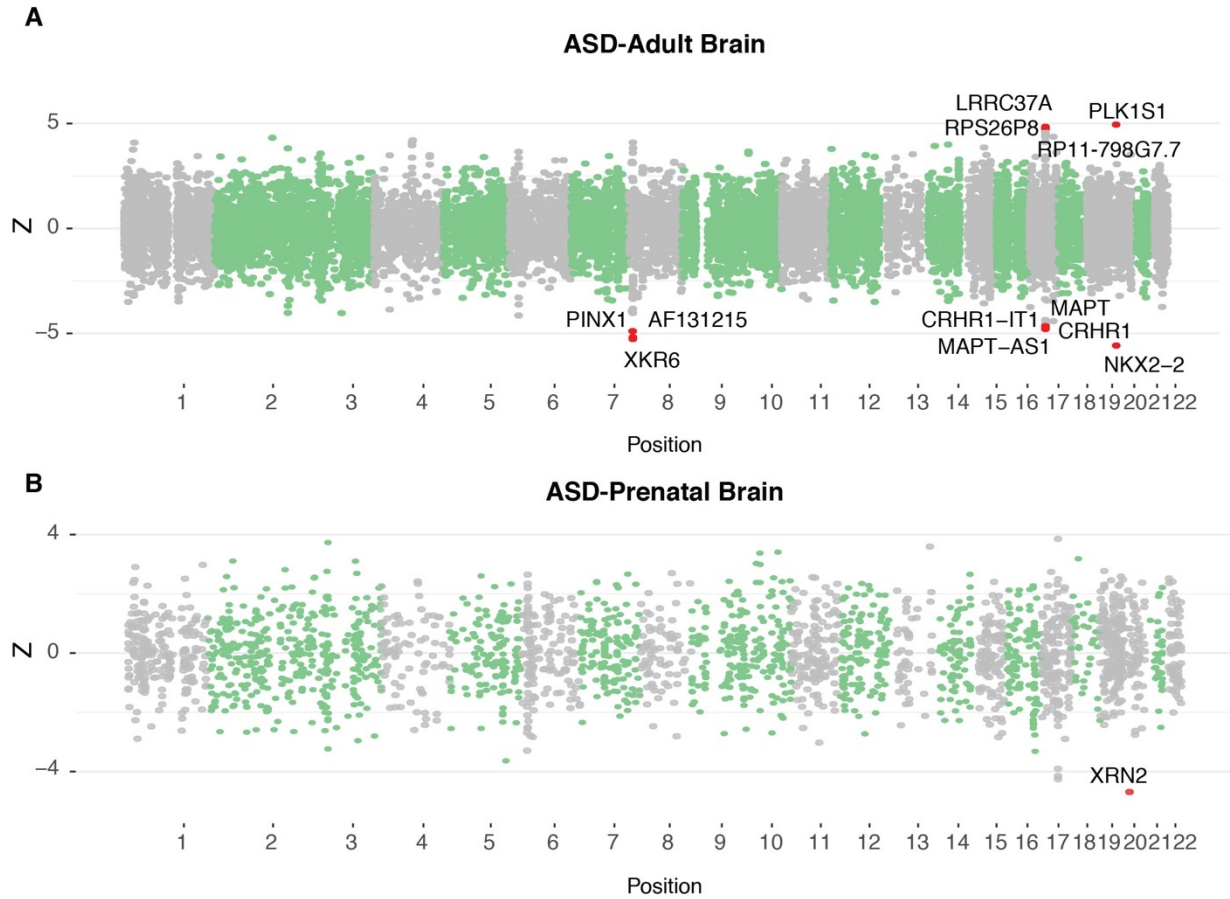


**Figure 6.4: ASD TWAS**

Results from TWAS prioritize genes whose *cis*-regulated expression in brain is associated with ASD. **(A)** The plot shows 12 genes are prioritized by TWAS at Bonferroni-corrected  $P < 0.05$  by PsychENCODE adult brain weights and **(B)** 1 gene prioritized by TWAS at Bonferroni-corrected  $P < 0.05$  by fetal brain weights. The sign of TWAS Z-scores indicates predicted direction of effect.

## **CHAPTER SEVEN**

### Conclusion and Future Directions

## Summary of Novel Findings

My analysis provides the largest genome wide map of human eQTLs and the first map of sQTLs during cerebral corticogenesis, expanding our understanding of complex gene and splicing regulation during this critical epoch of brain development. I have comprehensively described the implicated regulatory regions defined by eSNPs and sSNPs, enabling the integration of developmental diversity to previous adult brain functional genomic analysis and genetic variant interpretation. This allows us to identify known major regulators of both expression (transcription factors) and splicing (splicing factors/RNA binding proteins) whose activity is predicted to be affected by common human allelic variation. Several of these regulators of gene expression and splicing are known to be disrupted in neurodevelopmental disease via harboring rare variation (Carvill et al., 2013; Irimia et al., 2014; Lai et al., 2001; Suls et al., 2013; Zhang et al., 2014a), providing a link between the activities of common and rare variation on disease risk during brain development. I also show that much of the genetic variation controlling regulation of expression and splicing in the human brain is distinct, and highly sensitive to developmental stage. In the context of early onset neurological and psychiatric disorders, this provides a new window into genetic control of gene expression and splicing regulation during a critical time point for disease development. Also, to enable easier access to these fetal brain eQTLs and sQTL, we have built the DevEloPing Cortex Transcriptome viewer (DEPICT, <https://labs.dgsom.ucla.edu/geschwind/pages/eqtl-browser>), a web portal that allows browsing by gene or SNP summary-level data.

By way of comparison, a smaller fetal brain eQTL dataset consisting of 120 different samples (O'Brien et al., 2018), shows high concordance among significant eQTLs found in both studies (effect size correlation,  $r=0.67$ , Methods) (**Figure 4.S1 F**). Moreover, the effect size correlation of the overlapping eQTLs from this fetal dataset is significantly higher than all GTEx tissues, including the adult brain and PsychENCODE prefrontal cortex (**Figure 4.1 C**). However,

given our sample size, we identified nearly five times the number of eQTLs as that study (6,546 vs. 1,329), consistent with the well described relationship between eQTL discovery and sample size (Nica and Dermitzakis, 2013). Additionally, both studies identify a common inversion polymorphism on chromosome 17q21 (Stefansson et al., 2005) which harbors many eGenes. O'Brien et al. (O'Brien et al., 2018) find associations between 7 eGenes in this region with variants associated with neuroticism, including *LRRC37A*, a gene we identify as a fetal specific TWAS hit for intracranial volume and have fine mapped to this locus. Together, our studies emphasize the importance of understanding transcriptional regulation during prenatal periods.

I find that integrating eQTLs through regulatory region annotation with WGCNA defined modules from fetal brain expression shows distinct GWAS disease enrichment for ASD and SCZ within specific biological processes, indicating disease-specific convergent biology. This substantially advances previous work based on rare, gene disrupting *de novo* mutations (Iossifov et al., 2012; Ruzzo et al., 2019; Sanders et al., 2012) by showing convergence in genetic risk factors, even at the level of common variants that lie in regulatory regions. I show that genes involved in chromatin organization and splicing, as well as cell type markers for superficial layer fetal neurons are enriched for ASD GWAS loci in their eQTL defined regulatory regions, similar to rare variants. For SCZ, we find enrichment of the biological pathways of neurogenesis and central nervous system development, parallel to previous observations based on analysis of chromatin confirmation in fetal brain (Won et al., 2016), providing independent support for neurogenesis as a key process disrupted in SCZ risk. Furthermore, I show that SCZ GWAS enrichment from partitioned heritability significantly differs when using fetal eQTL defined genome annotations versus adult eQTL defined genome annotations, consistently indicating the fetal annotations to be the highest enriched among all functional annotations even when comparing to the PsychENCODE dataset with a sample size nearly 10 times larger than this cohort. Furthermore, the heritability explained by fetal and adult eQTLs shows an additive

effect, indicating the regulatory regions identified at these distinct time points indeed differ and are capturing different aspects of disease risk.

Similar significant enrichment of SCZ partitioned heritability has been shown for fetal brain ATAC-seq peaks (de la Torre-Ubieta et al., 2018), demonstrating the multiple lines of evidence that regulatory regions active in early cortical development harbor significant SCZ risk and are likely to be a crucial site of action for disease risk. Additionally, genetic risk for ADHD implicates fetal eQTL and sQTL, while ASD risk shows a trending enrichment in sQTL. These trending enrichments are likely to gain significance as GWAS sample sizes increase. The identification of splicing QTL from fetal brain will further allow mechanistic exploration of the effects of disease risk on gene isoform expression. Particularly interesting in this regard is the set of known splicing factors whose targets are enriched for disease-associated variation. This includes *SRRM4*, which is a major regulator of neuronal splicing including micro-exon inclusion, a process that seems to be particularly important for nervous system development and function (Irimia et al., 2014; Quesnel-Vallieres et al., 2015; Raj et al., 2014).

Finally, by integrating our map of fetal gene expression and splicing regulation with the SCZ, intracranial volume, and ASD GWAS studies via TWAS, we are able to identify new candidate genes and candidate molecular mechanisms through which these disease-associated common variants may be acting. We discover several genes that lie in significant GWAS loci and several new regions of the genome associated with SCZ through aggregation of GWAS signal among from the significant genetic cis-predictors of genes expression. Moreover, when comparing the implicated genes and introns from fetal brain expression weights with genes and introns implicated via adult brain expression weights, there is very little overlap. To some degree this is expected and reflects TWAS's known sensitivity to the specific expression panel (Wainberg et al., 2017), as well as the differences in eGenes and effect sizes between fetal and

adult eQTLs shown in our analysis. It does, however, highlight the importance of picking appropriate tissues and gene expression studies for a given phenotype. Here, the SCZ risk genes predicted by TWAS based on fetal brain expression contributes to the growing list of candidate genes discovered from adult expression for a more complete view of potentially casual genes and splicing events. These results and others (de la Torre-Ubieta et al., 2018; Won et al., 2016) are very consistent with the original seminal framing of SCZ as a neurodevelopmental disorder (Weinberger, 1987). These data strongly indicate that the developmental sensitivity for SCZ lies not only in adolescent development, but in the earliest periods of cortical neurogenesis and cortical patterning. This epoch is also particularly important for brain volume and patterning (de la Torre-Ubieta et al., 2018; Geschwind and Rakic, 2013), and the integration of eQTL data with the ENIGMA-CHARGE GWAS identifies many new genes influencing human brain size.

GWAS regions that yield discordant significant TWAS gene associations depending on the developmental time period of reference panel are interesting to consider. A particularly salient example is the region containing the 16p11.2 copy number variant, which has been associated with multiple neurodevelopmental conditions, including SCZ, ASD, ID, and BIP, (Bernier et al., 2017; Hanson et al., 2015; Shinawi et al., 2010; Weiss et al., 2008; Zhou et al., 2018). Fine mapping of this locus implicates *INO80E* and *TMEM219* using fetal brain data, with *INO80E* being the only gene in this locus implicated by both fetal and adult data. Fetal brain TWAS also implicated *KCTD13*, *CTD-2574D22.2*, *PPP4C*, and *YPEL3*, whereas *MAPK3*, *DOC2A* and *TAOK2*, all located in the copy variant region, are implicated by adult brain expression data, but not found in the credible gene list from FOCUS fine-mapping. This CNV region contains 29 genes, so the relationship of this CNV to phenotype is unlikely to be simple or be driven by only one gene (Escamilla et al., 2017; Golzio et al., 2012). Given the complexity

of these loci, additional fine mapping and functional analyses will be necessary to more conclusively identify SCZ risk genes.

In this regard, *INO80E* is an interesting candidate, due to its similar role as the chromatin-helicase-DNA binding protein family, including *CHD8*, in chromatin remodeling during corticoneurogenesis. The INO80 family of chromatin remodelers differ by having editing functions, along with their role in nucleosome spacing (Clapier et al., 2017; Sokpor et al., 2018). Of the fetal specific genes in this locus, *KCTD13* has been studied extensively and has been posited as a major driver of patient phenotypes based on experimental manipulation in animal models (Golzio et al., 2012; Luo et al., 2012). While more work is necessary to understand the role of *INO80E*, *TMEM219*, *KCTD13*, *MAPK3*, and the 16p11.2 CNV, our TWAS results suggest these genes may play a role in disease biology through action at different developmental time periods.

These results further demonstrate the importance of considering developmental stage of brain expression and regulation when using eQTL and sQTL data to interpret disease associated variants. They also highlight the value of using these data to further characterize developmental time points during which genetic disruption acts for neurodevelopmental and early onset neuropsychiatric diseases, suggesting that more detailed maps of the effects of common genetic variation effects throughout the lifespan will be of value.

## **Future Directions**

The first extension of this work is to combine this data with other recently published fetal brain eQTL datasets, including O'Brien et al (O'Brien et al., 2018), to increase power for eQTL detection. As the PsychENCODE project demonstrated (Wang et al., 2018c), combining multiple eQTL datasets, and therefore increasing sample size, greatly aids in eGene discovery, which

would provide a more complete picture of gene regulation during this critical time point in brain development. Furthermore, advancements in single-cell sequencing has allowed for the detection and profiling of complex cell populations. Single-cell RNA sequencing has been used to characterize both fetal and adult brain, revealing rare and novel cell types and cellular states (Lake et al., 2018; Polioudakis et al., 2019), and further implicating the need to assess cell-type specific effect on gene regulation. While experimental design for single-cell RNA-sequencing has largely focused on high-coverage sequencing of many cells from only a few individuals due to high costs (Gao et al., 2018; Joost et al., 2016), eQTL studies require many individuals to achieve sufficient power to detect population level variation in gene expression. However, a recent study has shown low-coverage single cell sequencing, followed by aggregation of reads within cell type is adequate for cell-type-specific QTL discovery when performed on enough individuals (Mandric et al., 2019). Such studies are inevitable in the near future and will provide a new, more refined context for GWAS follow-up.

Lastly, as described in Chapter 1.6, implicated TWAS genes and casual variants require further experimental validation. Massively parallel reporter assays (MPRA) allow for directly testing the effects of thousands of putative regulatory sequences in a high-throughput manner *in vitro* (Melnikov et al., 2012). By systematically creating a library of enhancer variation, with all known common variants, it is possible to gain single-nucleotide resolution of causal variants that control gene expression (Arnold et al., 2013; Kheradpour et al., 2013). To begin to validate identified fetal brain eQTLs, we can use MPRA's to test eSNP identified regulatory regions, starting with those identified in lying in close proximity to transcription factor binding sites of TFs with known roles in neurodevelopment as well as hub genes for the red and yellow modules, which showed significant enrichment in disease risk. While MPRA's readouts provide strong evidence of functional regulatory elements, they are performed in an artificial system out of the native genomic context and may not be able to detect the full effects of cell type specificity.

However, genome editing technology by CRISPR-Cas9 (Doudna and Charpentier, 2014), has revolutionized the ability to manipulate the human genome in *in vitro* systems of human cell lines. Introducing genetic variants or regions implicated to control expression of TWAS implicated genes via CRISPR in human fetal neuronal progenitors or neuronal organoids would allow for validation of eQTLs (Won et al. 2016; De La Torre Ubieta, 2018), as well as potential downstream consequences of altered expression in the proper genomic context (Lappalainen, 2015).

## REFERENCES

- Adams, H.H., Hibar, D.P., Chouraki, V., Stein, J.L., Nyquist, P.A., Renteria, M.E., Trompet, S., Arias-Vasquez, A., Seshadri, S., Desrivieres, S., *et al.* (2016). Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat Neurosci* 19, 1569-1582.
- Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., *et al.* (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, 787903.
- Almaguer-Mederos, L.E., Mesa, J.M.L., Gonzalez-Zaldivar, Y., Almaguer-Gotay, D., Cuello-Almarales, D., Aguilera-Rodriguez, R., Falcon, N.S., Gispert, S., Auburger, G., and Velazquez-Perez, L. (2018). Factors associated with ATXN2 CAG/CAA repeat intergenerational instability in Spinocerebellar ataxia type 2. *Clin Genet* 94, 346-350.
- 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. *Biol Psychiatry* 58, 85-96.
- Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166-169.
- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455-461.
- 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.
- Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074-1077.
- Assoum, M., Philippe, C., Isidor, B., Perrin, L., Makrythanasis, P., Sondheimer, N., Paris, C., Douglas, J., Lesca, G., Antonarakis, S., *et al.* (2016). Autosomal-Recessive Mutations in AP3B2, Adaptor-Related Protein Complex 3 Beta 2 Subunit, Cause an Early-Onset Epileptic Encephalopathy with Optic Atrophy. *Am J Hum Genet* 99, 1368-1376.
- Autism Spectrum Disorders Working Group of The Psychiatric Genomics, C. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism* 8, 21.
- Ayala, R., Willhoft, O., Aramayo, R.J., Wilkinson, M., McCormack, E.A., Ocloo, L., Wigley, D.B., and Zhang, X. (2018). Structure and regulation of the human INO80-nucleosome complex. *Nature* 556, 391-395.
- Bae, B.I., Jayaraman, D., and Walsh, C.A. (2015). Genetic changes shaping the human brain. *Dev Cell* 32, 423-434.

Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., and Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet* 10, e1004663.

Barca, E., Ganetzky, R.D., Potluri, P., Juanola-Falgarona, M., Gai, X., Li, D., Jalas, C., Hirsch, Y., Emmanuele, V., Tadesse, S., *et al.* (2018). USMG5 Ashkenazi Jewish founder mutation impairs mitochondrial complex V dimerization and ATP synthesis. *Hum Mol Genet* 27, 3305-3312.

Barrett, J.C., and Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies. *Nat Genet* 38, 659-662.

Basu, A.C., Tsai, G.E., Ma, C.L., Ehmsen, J.T., Mustafa, A.K., Han, L., Jiang, Z.I., Benneyworth, M.A., Froimowitz, M.P., Lange, N., *et al.* (2009). Targeted disruption of serine racemase affects glutamatergic neurotransmission and behavior. *Mol Psychiatry* 14, 719-727.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. *Science* 347, 664-667.

Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschield, C.D., Beckman, K.B., Shi, J., Mei, R., *et al.* (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24, 14-24.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321-1325.

Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., and Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 12, R10.

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. *J Roy Stat Soc B Met* 57, 289-300.

Benyamin, B., Pourcain, B., Davis, O.S., Davies, G., Hansell, N.K., Brion, M.J., Kirkpatrick, R.M., Cents, R.A., Franic, S., Miller, M.B., *et al.* (2014). Childhood intelligence is heritable, highly polygenic and associated with FBNP1L. *Mol Psychiatry* 19, 253-258.

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.

Bernier, R., Hudac, C.M., Chen, Q., Zeng, C., Wallace, A.S., Gerdts, J., Earl, R., Peterson, J., Wolken, A., Peters, A., *et al.* (2017). Developmental trajectories for young children with 16p11.2 copy number variation. *Am J Med Genet B Neuropsychiatr Genet* 174, 367-380.

Bestman, J.E., Huang, L.C., Lee-Osbourne, J., Cheung, P., and Cline, H.T. (2015). An in vivo screen to identify candidate neurogenic genes in the developing *Xenopus* visual system. *Dev Biol* 408, 269-291.

- Biemont, C., and Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. *Nature* 443, 521-524.
- Billingslea, E.N., Tatard-Leitman, V.M., Anguiano, J., Jutzeler, C.R., Suh, J., Saunders, J.A., Morita, S., Featherstone, R.E., Ortinski, P.I., Gandal, M.J., *et al.* (2014). Parvalbumin cell ablation of NMDA-R1 causes increased resting network excitability with associated social and self-care deficits. *Neuropsychopharmacology* 39, 1603-1613.
- Bolton, P., Macdonald, H., Pickles, A., Rios, P., Goode, S., Crowson, M., Bailey, A., and Rutter, M. (1994). A case-control family history study of autism. *J Child Psychol Psychiatry* 35, 877-900.
- BrainSpan (2013). BrainSpan: Atlas of the Developing Human Brain.
- Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34, 525-527.
- Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752-755.
- Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* 98, 116-126.
- Buizer-Voskamp, J.E., Muntjewerff, J.W., Genetic, R., Outcome in Psychosis Consortium, M., Strengman, E., Sabatti, C., Stefansson, H., Vorstman, J.A., and Ophoff, R.A. (2011). Genome-wide analysis shows increased frequency of copy number variation deletions in Dutch schizophrenia patients. *Biol Psychiatry* 70, 655-662.
- Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291-295.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36, 411-420.
- Cannon, M.E., and Mohlke, K.L. (2018). Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *Am J Hum Genet* 103, 637-653.
- Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22, 231-238.
- Carvill, G.L., Heavin, S.B., Yendle, S.C., McMahon, J.M., O'Roak, B.J., Cook, J., Khan, A., Dorschner, M.O., Weaver, M., Calvert, S., *et al.* (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat Genet* 45, 825-830.
- Chakravarti, A. (1999). Population genetics--making sense out of sequence. *Nat Genet* 21, 56-60.
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7.

- Chen, C., Cheng, L., Grennan, K., Pibiri, F., Zhang, C., Badner, J.A., Members of the Bipolar Disorder Genome Study, C., Gershon, E.S., and Liu, C. (2013). Two gene co-expression modules differentiate psychotics and controls. *Mol Psychiatry* 18, 1308-1314.
- Chen, J., Calhoun, V.D., Perrone-Bizzozero, N.I., Pearlson, G.D., Sui, J., Du, Y., and Liu, J. (2016a). A pilot study on commonality and specificity of copy number variants in schizophrenia and bipolar disorder. *Transl Psychiatry* 6, e824.
- Chen, W., McDonnell, S.K., Thibodeau, S.N., Tillmans, L.S., and Schaid, D.J. (2016b). Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. *Genetics* 204, 933-958.
- Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M., and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437, 1365-1369.
- Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet* 10, e1004787.
- Clapier, C.R., Iwasa, J., Cairns, B.R., and Peterson, C.L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat Rev Mol Cell Biol* 18, 407-422.
- Cockerill, P.N. (2011). Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J* 278, 2182-2210.
- Cohen, O.S., McCoy, S.Y., Middleton, F.A., Bialosuknia, S., Zhang-James, Y., Liu, L., Tsuang, M.T., Faraone, S.V., and Glatt, S.J. (2012). Transcriptomic analysis of postmortem brain identifies dysregulated splicing events in novel candidate genes for schizophrenia. *Schizophr Res* 142, 188-199.
- Colantuoni, C., Lipska, B.K., Ye, T., Hyde, T.M., Tao, R., Leek, J.T., Colantuoni, E.A., Elkahoul, 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.
- Constantino, J.N., Zhang, Y., Frazier, T., Abbacchi, A.M., and Law, P. (2010). Sibling recurrence and the genetic epidemiology of autism. *Am J Psychiatry* 167, 1349-1356.
- CONVERGE Consortium (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 523, 588-591.
- Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., *et al.* (2011). A copy number variation morbidity map of developmental delay. *Nat Genet* 43, 838-846.
- Cotney, J., Muhle, R.A., Sanders, S.J., Liu, L., Willsey, A.J., Niu, W., Liu, W., Klei, L., Lei, J., Yin, J., *et al.* (2015). The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat Commun* 6, 6404.
- Cross-Disorder Group of the Psychiatric Genomics, C. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 381, 1371-1379.

- Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PLoS Genet* 10, e1004226.
- Da Cruz, S., and Cleveland, D.W. (2011). Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. *Curr Opin Neurobiol* 21, 904-919.
- Darnell, R.B. (2013). RNA protein interaction in neurons. *Annu Rev Neurosci* 36, 243-270.
- Darvish, H., Azcona, L.J., Tafakhori, A., Ahmadi, M., Ahmadifard, A., and Paisan-Ruiz, C. (2017). Whole genome sequencing identifies a novel homozygous exon deletion in the NT5C2 gene in a family with intellectual disability and spastic paraplegia. *NPJ Genom Med* 2, 20.
- de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* 172, 289-304 e218.
- de la Torre-Ubieta, L., Won, H., Stein, J.L., and Geschwind, D.H. (2016). Advancing the understanding of autism disease mechanisms through genetics. *Nat Med* 22, 345-361.
- De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., *et al.* (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209-215.
- Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., *et al.* (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390-394.
- Demontis, D., Walters, R.K., Martin, J., Mattheisen, M., Als, T.D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Baekvad-Hansen, M., *et al.* (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet* 51, 63-75.
- Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., *et al.* (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246-1250.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21.
- Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346, 1258096.
- Durak, O., Gao, F., Kaeser-Woo, Y.J., Rueda, R., Martorell, A.J., Nott, A., Liu, C.Y., Watson, L.A., and Tsai, L.H. (2016). Chd8 mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. *Nat Neurosci* 19, 1477-1488.
- Eising, E., Carrion-Castillo, A., Vino, A., Strand, E.A., Jakielski, K.J., Scerri, T.S., Hildebrand, M.S., Webster, R., Ma, A., Mazoyer, B., *et al.* (2019). A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Molecular Psychiatry* 24, 1065-1078.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., *et al.* (2008). Genetics of gene expression and its effect on disease. *Nature* 452, 423-428.

Encode Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636-640.

Encode Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 9, e1001046.

Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74.

Encode Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799-816.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 9, 215-216.

Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* 33, 364-376.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* 12, 2478-2492.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43-49.

Escamilla, C.O., Filonova, I., Walker, A.K., Xuan, Z.X., Holehonnur, R., Espinosa, F., Liu, S., Thyme, S.B., Lopez-Garcia, I.A., Mendoza, D.B., *et al.* (2017). Kctd13 deletion reduces synaptic transmission via increased RhoA. *Nature* 551, 227-231.

Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shores, N., Whitton, H., Ryan, R.J., Shishkin, A.A., *et al.* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337-343.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., *et al.* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 47, 1228-1235.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* 9, e1003486.

Fraser, H.B. (2013). Gene expression drives local adaptation in humans. *Genome Res* 23, 1089-1096.

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.

- Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., *et al.* (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* 19, 1442-1453.
- Gaffney, D.J. (2013). Global properties and functional complexity of human gene regulatory variation. *PLoS Genet* 9, e1003501.
- Gaffney, D.J., Veyrieras, J.B., Degner, J.F., Pique-Regi, R., Pai, A.A., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol* 13, R7.
- Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to Function. *Am J Hum Genet* 102, 717-730.
- Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Consortium, G.T., Nicolae, D.L., *et al.* (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 47, 1091-1098.
- Gamazon, E.R., Zwinderman, A.H., Cox, N.J., Denys, D., and Derks, E.M. (2019). Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. *Nat Genet* 51, 933-940.
- Gandal, M.J., Haney, J.R., Parikshak, N.N., Leppa, V., Ramaswami, G., Hartl, C., Schork, A.J., Appadurai, V., Buil, A., Werge, T.M., *et al.* (2018a). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* 359, 693-697.
- Gandal, M.J., Leppa, V., Won, H., Parikshak, N.N., and Geschwind, D.H. (2016). The road to precision psychiatry: translating genetics into disease mechanisms. *Nat Neurosci* 19, 1397-1407.
- Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., *et al.* (2018b). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 362, eaat8127.
- Gao, S., Yan, L., Wang, R., Li, J., Yong, J., Zhou, X., Wei, Y., Wu, X., Wang, X., Fan, X., *et al.* (2018). Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat Cell Biol* 20, 721-734.
- 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. *Neurobiol Dis* 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.
- Gerrits, A., Li, Y., Tesson, B.M., Bystrykh, L.V., Weersing, E., Ausema, A., Dontje, B., Wang, X., Breitling, R., Jansen, R.C., *et al.* (2009). Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet* 5, e1000692.
- Geschwind, D.H. (2011). Genetics of autism spectrum disorders. *Trends Cogn Sci* 15, 409-416.

- Geschwind, D.H., and Flint, J. (2015). Genetics and genomics of psychiatric disease. *Science* 349, 1489-1494.
- 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 Rakic, P. (2013). Cortical evolution: judge the brain by its cover. *Neuron* 80, 633-647.
- Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. *Nat Rev Genet* 10, 94-108.
- Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 10, e1004383.
- 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.
- 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.
- Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., *et al.* (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* 43, 860-863.
- Girirajan, S., Dennis, M.Y., Baker, C., Malig, M., Coe, B.P., Campbell, C.D., Mark, K., Vu, T.H., Alkan, C., Cheng, Z., *et al.* (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am J Hum Genet* 92, 221-237.
- Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., *et al.* (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459, 569-573.
- Goldstein, D.B. (2009). Common genetic variation and human traits. *N Engl J Med* 360, 1696-1698.
- Golzio, C., Willer, J., Talkowski, M.E., Oh, E.C., Taniguchi, Y., Jacquemont, S., Reymond, A., Sun, M., Sawa, A., Gusella, J.F., *et al.* (2012). KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* 485, 363-367.
- Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 82, 100-112.
- Grammatikakis, I., Zhang, P., Panda, A.C., Kim, J., Maudsley, S., Abdelmohsen, K., Yang, X., Martindale, J.L., Motino, O., Hutchison, E.R., *et al.* (2016). Alternative Splicing of Neuronal Differentiation Factor TRF2 Regulated by HNRNPH1/H2. *Cell Rep* 15, 926-934.
- 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.

- Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., *et al.* (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**, 431-444.
- Grozeva, D., Kirov, G., Ivanov, D., Jones, I.R., Jones, L., Green, E.K., St Clair, D.M., Young, A.H., Ferrier, N., Farmer, A.E., *et al.* (2010). Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* **67**, 318-327.
- Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A., *et al.* (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-1089.
- Gulsuner, S., Walsh, T., Watts, A.C., Lee, M.K., Thornton, A.M., Casadei, S., Rippey, C., Shahin, H., Consortium on the Genetics of, S., Group, P.S., *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 Commun* **5**, 5748.
- Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., *et al.* (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-252.
- Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Schizophrenia Working Group of the Psychiatric Genomics, C., McCarroll, S., *et al.* (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat Genet* **50**, 538-548.
- Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., *et al.* (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523.
- Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S.B., Buil, A., Yurovsky, A., Bryois, J., Padiou, I., Romano, L., Planchon, A., *et al.* (2015). Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet* **11**, e1004958.
- 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.
- Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T.M., Troakes, C., Turecki, G., O'Donovan, M.C., Schalkwyk, L.C., *et al.* (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci* **19**, 48-54.
- Hanson, E., Bernier, R., Porche, K., Jackson, F.I., Goin-Kochel, R.P., Snyder, L.G., Snow, A.V., Wallace, A.S., Campe, K.L., Zhang, Y., *et al.* (2015). The cognitive and behavioral phenotype of the 16p11.2 deletion in a clinically ascertained population. *Biol Psychiatry* **77**, 785-793.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., *et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22, 1760-1774.

Hashimoto, R., Ohi, K., Okada, T., Yasuda, Y., Yamamori, H., Hori, H., Hikita, T., Taya, S., Saitoh, O., Kosuga, A., *et al.* (2009). Association analysis between schizophrenia and the AP-3 complex genes. *Neurosci Res* 65, 113-115.

Hashimoto, T., Arion, D., Unger, T., Maldonado-Aviles, J.G., Morris, H.M., Volk, D.W., Mirnics, K., and Lewis, D.A. (2008). Alterations in GABA-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia. *Mol Psychiatry* 13, 147-161.

Hawrylycz, M., Miller, J.A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A.L., Jegga, A.G., Aronow, B.J., Lee, C.K., Bernard, A., *et al.* (2015). Canonical genetic signatures of the adult human brain. *Nat Neurosci* 18, 1832-1844.

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.

Hernandez, D.G., Nalls, M.A., Moore, M., Chong, S., Dillman, A., Trabzuni, D., Gibbs, J.R., Ryten, M., Arepalli, S., Weale, M.E., *et al.* (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol Dis* 47, 20-28.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., *et al.* (2013). DNA methylation contributes to natural human variation. *Genome Res* 23, 1363-1372.

Hibar, D.P., Stein, J.L., Renteria, M.E., Arias-Vasquez, A., Desrivieres, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., *et al.* (2015). Common genetic variants influence human subcortical brain structures. *Nature* 520, 224-229.

Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J., Loh, P.R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., *et al.* (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat Genet* 50, 1041-1047.

Hormozdiari, F., van de Bunt, M., Segre, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* 99, 1245-1260.

Ingason, A., Giegling, I., Hartmann, A.M., Genius, J., Konte, B., Friedl, M., Schizophrenia Working Group of the Psychiatric Genomics, C., Ripke, S., Sullivan, P.F., St Clair, D., *et al.* (2015). Expression analysis in a rat psychosis model identifies novel candidate genes validated in a large case-control sample of schizophrenia. *Transl Psychiatry* 5, e656.

Ingason, A., Kirov, G., Giegling, I., Hansen, T., Isles, A.R., Jakobsen, K.D., Kristinsson, K.T., le Roux, L., Gustafsson, O., Craddock, N., *et al.* (2011). Maternally derived microduplications at 15q11-q13: implication of imprinted genes in psychotic illness. *Am J Psychiatry* 168, 408-417.

International HapMap, C. (2003). The International HapMap Project. *Nature* 426, 789-796.

International League Against Epilepsy Consortium on Complex Epilepsies (2014). Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol* 13, 893-903.

International Schizophrenia, C. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455, 237-241.

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-Vallieres, 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.

Jacquemont, M.L., Sanlaville, D., Redon, R., Raoul, O., Cormier-Daire, V., Lyonnet, S., Amiel, J., Le Merrer, M., Heron, D., de Blois, M.C., *et al.* (2006). Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *J Med Genet* 43, 843-849.

Jaffe, A.E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T.M., Weinberger, D.R., and Kleinman, J.E. (2016). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* 19, 40-47.

Jaffe, A.E., Straub, R.E., Shin, J.H., Tao, R., Gao, Y., Collado-Torres, L., Kam-Thong, T., Xi, H.S., Quan, J., Chen, Q., *et al.* (2018). Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* 21, 1117-1125.

Johnson, M.B., Kawasawa, Y.I., Mason, C.E., Krsnik, Z., Coppola, G., Bogdanovic, D., Geschwind, D.H., Mane, S.M., State, M.W., and Sestan, N. (2009). Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 62, 494-509.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118-127.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 13, 484-492.

Joost, S., Zeisel, A., Jacob, T., Sun, X., La Manno, G., Lönnerberg, P., Linnarsson, S., and Kasper, M. (2016). Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity. *Cell Syst* 3, 221-237 e229.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., *et al.* (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119-124.

- Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 91, 839-848.
- Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M., Pletikos, M., Meyer, K.A., Sedmak, G., *et al.* (2011). Spatio-temporal transcriptome of the human brain. *Nature* 478, 483-489.
- Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709-1723.
- Karayiorgou, M., Morris, M.A., Morrow, B., Shprintzen, R.J., Goldberg, R., Borrow, J., Gos, A., Nestadt, G., Wolyniec, P.S., Lasseter, V.K., *et al.* (1995). Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc Natl Acad Sci U S A* 92, 7612-7616.
- Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* 23, 800-811.
- Kichaev, G., and Pasaniuc, B. (2015). Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am J Hum Genet* 97, 260-271.
- Kichaev, G., Yang, W.Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* 10, e1004722.
- Kim, Y., Xia, K., Tao, R., Giusti-Rodriguez, P., Vladimirov, V., van den Oord, E., and Sullivan, P.F. (2014). A meta-analysis of gene expression quantitative trait loci in brain. *Transl Psychiatry* 4, e459.
- Kim, Y.E., Oh, K.W., Noh, M.Y., Park, J., Kim, H.J., Park, J.E., Ki, C.S., and Kim, S.H. (2018). Analysis of ATXN2 trinucleotide repeats in Korean patients with amyotrophic lateral sclerosis. *Neurobiol Aging* 67, 201 e205-201 e208.
- Kirov, G., Grozeva, D., Norton, N., Ivanov, D., Mantripragada, K.K., Holmans, P., International Schizophrenia, C., Wellcome Trust Case Control, C., Craddock, N., Owen, M.J., *et al.* (2009). Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* 18, 1497-1503.
- Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., 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.
- Kostovic, I., and Jovanov-Milosevic, N. (2006). The development of cerebral connections during the first 20-45 weeks' gestation. *Semin Fetal Neonatal Med* 11, 415-422.
- Kremen, W.S., Seidman, L.J., Pepple, J.R., Lyons, M.J., Tsuang, M.T., and Faraone, S.V. (1994). Neuropsychological risk indicators for schizophrenia: a review of family studies. *Schizophr Bull* 20, 103-119.

- Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., *et al.* (2015). Excess of rare, inherited truncating mutations in autism. *Nat Genet* 47, 582-588.
- Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* 413, 519-523.
- Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V., *et al.* (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* 36, 70-80.
- Lalonde, E., Ha, K.C., Wang, Z., Bemmo, A., Kleinman, C.L., Kwan, T., Pastinen, T., and Majewski, J. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res* 21, 545-554.
- Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., *et al.* (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45, 1452-1458.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559.
- Lappalainen, T. (2015). Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res* 25, 1427-1431.
- Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., *et al.* (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506-511.
- Latourelle, J.C., Dumitriu, A., Hadzi, T.C., Beach, T.G., and Myers, R.H. (2012). Evaluation of Parkinson disease risk variants as expression-QTLs. *PLoS One* 7, e46199.
- Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3, 1724-1735.
- 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.* (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168-176.
- Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291.
- Levinson, D.F., Duan, J., Oh, S., Wang, K., Sanders, A.R., Shi, J., Zhang, N., Mowry, B.J., Olincy, A., Amin, F., *et al.* (2011). Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* 168, 302-316.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.

Li, Q., Lee, J.A., and Black, D.L. (2007). Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci* 8, 819-831.

Li, Y., and Kellis, M. (2016). Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res* 44, e144.

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* 50, 151-158.

Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600-604.

Ling, S.C., Polymenidou, M., and Cleveland, D.W. (2013). Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron* 79, 416-438.

Liu, B., Pjanic, M., Wang, T., Nguyen, T., Gloudemans, M., Rao, A., Castano, V.G., Nurnberg, S., Rader, D.J., Elwyn, S., *et al.* (2018). Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci. *Am J Hum Genet* 103, 377-388.

Luo, R., Sanders, S.J., Tian, Y., Voineagu, I., Huang, N., Chu, S.H., Klei, L., Cai, C., Ou, J., Lowe, J.K., *et al.* (2012). Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet* 91, 38-55.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., *et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202-1214.

Malhotra, D., McCarthy, S., Michaelson, J.J., Vacic, V., Burdick, K.E., Yoon, S., Cichon, S., Corvin, A., Gary, S., Gershon, E.S., *et al.* (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* 72, 951-963.

Malhotra, D., and Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148, 1223-1241.

Mancarci, B.O., Toker, L., Tripathy, S.J., Li, B., Rocco, B., Sibille, E., and Pavlidis, P. (2017). Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. *eNeuro* 4.

- Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet* 51, 675-682.
- Mandric, I., Schwarz, T., Majumdar, A., Perez, R., Subramaniam, M., Ye, C.J., Pasaniuc, B., and Halperin, E. (2019). Optimal design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *bioRxiv*, 766972.
- Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747-753.
- Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., *et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 82, 477-488.
- Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190-1195.
- McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9, 356-369.
- McCarthy, S.E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., Mistry, M., Pavlidis, P., Solomon, R., Ghiban, E., *et al.* (2014). De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol Psychiatry* 19, 652-658.
- McCarthy, S.E., Makarov, V., Kirov, G., Addington, A.M., McClellan, J., Yoon, S., Perkins, D.O., Dickel, D.E., Kusenda, M., Krastoshevsky, O., *et al.* (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 41, 1223-1227.
- McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122.
- Mefford, H.C., Sharp, A.J., Baker, C., Itsara, A., Jiang, Z., Buysse, K., Huang, S., Maloney, V.K., Crolla, J.A., Baralle, D., *et al.* (2008). Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* 359, 1685-1699.
- Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B., *et al.* (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30, 271-277.
- 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.

- Mohammadi, P., Castel, S.E., Brown, A.A., and Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res* 27, 1872-1884.
- Monlong, J., Calvo, M., Ferreira, P.G., and Guigo, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseeker. *Nat Commun* 5, 4698.
- Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773-777.
- Moreno-De-Luca, D., Consortium, S., Mulle, J.G., Simons Simplex Collection Genetics, C., Kaminsky, E.B., Sanders, S.J., GeneStar, Myers, S.M., Adam, M.P., Pakula, A.T., *et al.* (2010). Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* 87, 618-630.
- Moreno-De-Luca, D., Sanders, S.J., Willsey, A.J., Mulle, J.G., Lowe, J.K., Geschwind, D.H., State, M.W., Martin, C.L., and Ledbetter, D.H. (2013). Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Mol Psychiatry* 18, 1090-1095.
- Mostafavi, S., Battle, A., Zhu, X., Urban, A.E., Levinson, D., Montgomery, S.B., and Koller, D. (2013). Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One* 8, e68141.
- Mulle, J.G., Dodd, A.F., McGrath, J.A., Wolyniec, P.S., Mitchell, A.A., Shetty, A.C., Sobreira, N.L., Valle, D., Rudd, M.K., Satten, G., *et al.* (2010). Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* 87, 229-236.
- 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.
- Nica, A.C., and Dermitzakis, E.T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Hum Mol Genet* 17, R129-134.
- Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* 368, 20120362.
- Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., *et al.* (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 7, e1002003.
- Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 6, e1000888.
- Nord, A.S., Pattabiraman, K., Visel, A., and Rubenstein, J.L.R. (2015). Genomic perspectives of transcriptional regulation in forebrain development. *Neuron* 85, 27-47.
- O'Brien, H.E., Hannon, E., Hill, M.J., Toste, C.C., Robertson, M.J., Morgan, J.E., McLaughlin, G., Lewis, C.M., Schalkwyk, L.C., Hall, L.S., *et al.* (2018). Expression quantitative trait loci in the

developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol* 19, 194.

O'Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J.B., Nickerson, D.A., *et al.* (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat Commun* 5, 5595.

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.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symp Biol* 23, 366-370.

Ohsakaya, S., Fujikawa, M., Hisabori, T., and Yoshida, M. (2011). Knockdown of DAPIT (diabetes-associated protein in insulin-sensitive tissue) results in loss of ATP synthase in mitochondria. *J Biol Chem* 286, 20292-20296.

Ojeda, S.R., Hill, J., Hill, D.F., Costa, M.E., Tapia, V., Cornea, A., and Ma, Y.J. (1999). The Oct-2 POU domain gene in the neuroendocrine brain: a transcriptional regulator of mammalian puberty. *Endocrinology* 140, 3774-3789.

Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.B., Emilsson, V., Meddens, S.F., *et al.* (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533, 539-542.

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.

Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32, 1479-1485.

Oyama, S., Yamakawa, H., Sasagawa, N., Hosoi, Y., Futai, E., and Ishiura, S. (2009). Dysbindin-1, a schizophrenia-related protein, functionally interacts with the DNA- dependent protein kinase complex in an isoform-dependent manner. *PLoS One* 4, e4199.

Pai, A.A., Cain, C.E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.B., Degner, J.F., Gaffney, D.J., Pickrell, J.K., Stephens, M., *et al.* (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet* 8, e1003000.

Pai, A.A., Pritchard, J.K., and Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet* 11, e1004857.

Pain, O., Pocklington, A.J., Holmans, P.A., Bray, N.J., O'Brien, H.E., Hall, L.S., Pardini, A.F., O'Donovan, M.C., Owen, M.J., and Anney, R. (2019). Novel Insight Into the Etiology of Autism Spectrum Disorder Gained by Integrating Expression Data With Genome-wide Association Statistics. *Biol Psychiatry* 86, 265-273.

Pardini, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S.E., Bishop, S., Cameron, D., Hamshere, M.L., *et al.* (2018). Common schizophrenia alleles

are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* 50, 381-389.

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

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.

Parikshak, N.N., Swarup, V., Belgard, T.G., Irimia, M., Ramaswami, G., Gandal, M.J., Hartl, C., Leppa, V., Ubieta, L.T., Huang, J., *et al.* (2016). Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* 540, 423-427.

Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* 42, 570-575.

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

Pavlidis, J.M., Zhu, Z., Gratten, J., McRae, A.F., Wray, N.R., and Yang, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med* 8, 84.

Pe'er, I., de Bakker, P.I., Maller, J., Yelensky, R., Altshuler, D., and Daly, M.J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* 38, 663-667.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., *et al.* (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499-502.

Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* 94, 559-573.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010a). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768-772.

Pickrell, J.K., Pai, A.A., Gilad, Y., and Pritchard, J.K. (2010b). Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* 6, e1001236.

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.

Polderman, T.J., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* 47, 702-709.

Polioudakis, D., de la Torre-Ubieta, L., Langerman, J., Elkins, A.G., Shi, X., Stein, J.L., Vuong, C.K., Nichterwitz, S., Gevorgian, M., Opland, C.K., *et al.* (2019). A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron*.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., *et al.* (2015). Molecular identity of human outer radial glia during cortical development. *Cell* 163, 55-67.

Preciados, M., Yoo, C., and Roy, D. (2016). Estrogenic Endocrine Disrupting Chemicals Influencing NRF1 Regulated Gene Networks in the Development of Complex Human Brain Diseases. *Int J Mol Sci* 17, 2086.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69, 124-137.

Psychiatric, G.C.B.D.W.G. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 43, 977-983.

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., Kahler, A., *et al.* (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185-190.

Quesnel-Vallieres, M., Irimia, M., Cordes, S.P., and Blencowe, B.J. (2015). Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes Dev* 29, 746-759.

Raj, B., and Blencowe, B.J. (2015). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron* 87, 14-27.

Raj, B., Irimia, M., Braunschweig, U., Sterne-Weiler, T., O'Hanlon, D., Lin, Z.Y., Chen, G.I., Easton, L.E., Ule, J., Gingras, A.C., *et al.* (2014). A global regulatory mechanism for activating an exon network required for neurogenesis. *Mol Cell* 56, 90-103.

Rakic, P. (1995). A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. *Trends Neurosci* 18, 383-388.

Rakic, P. (2009). Evolution of the neocortex: a perspective from developmental biology. *Nat Rev Neurosci* 10, 724-735.

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Consortium, U.K.B.E., North American Brain Expression, C., Coin, L., *et al.* (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat Neurosci* 17, 1418-1428.

Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* 10, e1004525.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends Genet* 17, 502-510.

Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., *et al.* (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340, 1467-1471.

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

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kahler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., *et al.* (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 45, 1150-1159.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* 273, 1516-1517.

Ritvo, E.R., Jorde, L.B., Mason-Brothers, A., Freeman, B.J., Pingree, C., Jones, M.B., McMahon, W.M., Petersen, P.B., Jenson, W.R., and Mo, A. (1989). The UCLA-University of Utah epidemiologic survey of autism: recurrence risk estimates and genetic counseling. *Am J Psychiatry* 146, 1032-1036.

Rivera, C.M., and Ren, B. (2013). Mapping human epigenomes. *Cell* 155, 39-55.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., *et al.* (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317-330.

Robinson, E.B., St Pourcain, B., Anttila, V., Kosmicki, J.A., Bulik-Sullivan, B., Grove, J., Maller, J., Samocha, K.E., Sanders, S.J., Ripke, S., *et al.* (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat Genet* 48, 552-555.

Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. *Nat Rev Genet* 7, 862-872.

Rujescu, D., Ingason, A., Cichon, S., Pietilainen, O.P., Barnes, M.R., Touloupoulou, T., Picchioni, M., Vassos, E., Ettinger, U., Bramon, E., *et al.* (2009). Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* 18, 988-996.

Ruzzo, E.K., Perez-Cano, L., Jung, J.Y., Wang, L.K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J.N., Leventhal, O., *et al.* (2019). Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* 178, 850-866 e826.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, 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.

Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., *et al.* (2019). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *bioRxiv*, 484113.

Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19, 491-504.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res* 22, 1748-1759.

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

Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* 31, 2601-2606.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19, 212-219.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., *et al.* (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445-449.

Sebat, J., Levy, D.L., and McCarthy, S.E. (2009). Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet* 25, 528-535.

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

Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* 111, E5593-5601.

Shinawi, M., Liu, P., Kang, S.H., Shen, J., Belmont, J.W., Scott, D.A., Probst, F.J., Craigen, W.J., Graham, B.H., Pursley, A., *et al.* (2010). Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet* 47, 332-341.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15, 1034-1050.

Silbereis, J.C., Pochareddy, S., Zhu, Y., Li, M., and Sestan, N. (2016). The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. *Neuron* 89, 248-268.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat Rev Genet* 14, 204-220.

Sokpor, G., Castro-Hernandez, R., Rosenbusch, J., Staiger, J.F., and Tuoc, T. (2018). ATP-Dependent Chromatin Remodeling During Cortical Neurogenesis. *Front Neurosci* 12, 226.

Spain, S.L., and Barrett, J.C. (2015). Strategies for fine-mapping complex traits. *Hum Mol Genet* 24, R111-119.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., *et al.* (2005). A common inversion under selection in Europeans. *Nat Genet* 37, 129-137.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100, 9440-9445.

Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., *et al.* (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* 8, e1002639.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., *et al.* (2007). Population genomics of human gene expression. *Nat Genet* 39, 1217-1224.

Strunz, T., Grassmann, F., Gayan, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B.H.F. (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci Rep* 8, 5865.

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. *Proc Natl Acad Sci U S A* 111, E4468-4477.

Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* 13, 537-551.

Suls, A., Jaehn, J.A., Kecskes, A., Weber, Y., Weckhuysen, S., Craiu, D.C., Siekierska, A., Djemie, T., Afrikanova, T., Gormley, P., *et al.* (2013). De novo loss-of-function mutations in CHD2 cause a fever-sensitive myoclonic epileptic encephalopathy sharing features with Dravet syndrome. *Am J Hum Genet* 93, 967-975.

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.

Taal, H.R., Pourcain, B.S., Thiering, E., Das, S., Mook-Kanamori, D.O., Warrington, N.M., Kaakinen, M., Kreiner-Moller, E., Bradfield, J.P., Freathy, R.M., *et al.* (2012). Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat Genet* 44, 532-538.

Takata, A., Matsumoto, N., and Kato, T. (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun* 8, 14519.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., *et al.* (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* 19, 335-346.

Tavassoli, T., Kolevzon, A., Wang, A.T., Curchack-Lichtin, J., Halpern, D., Schwartz, L., Soffes, S., Bush, L., Grodberg, D., Cai, G., *et al.* (2014). De novo SCN2A splice site mutation in a boy with Autism spectrum disorder. *BMC Med Genet* 15, 35.

The 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., *et al.* (2015). A global reference for human genetic variation. *Nature* 526, 68-74.

The FANTOM Consortium and the RIKEN PMI and CLST, Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., *et al.* (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462-470.

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

The GTEx Consortium, Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, *et al.* (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204-213.

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.

Torres, J.M., Gamazon, E.R., Parra, E.J., Below, J.E., Valladares-Salgado, A., Wachter, N., Cruz, M., Hanis, C.L., and Cox, N.J. (2014). Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *Am J Hum Genet* 95, 521-534.

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

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45, 124-130.

Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A., *et al.* (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* 98, 58-74.

Turner, T.N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., HA, F.S., Doebley, A.L., Bernier, R.A., Nickerson, D.A., and Eichler, E.E. (2017). denovo-db: a compendium of human de novo variants. *Nucleic Acids Res* 45, D804-D811.

Vacic, V., McCarthy, S., Malhotra, D., Murray, F., Chou, H.H., Peoples, A., Makarov, V., Yoon, S., Bhandari, A., Corominas, R., *et al.* (2011). Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* 471, 499-503.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., *et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43, 11 10 11-33.

Veerappa, A.M., Saldanha, M., Padakannaya, P., and Ramachandra, N.B. (2014). Family based genome-wide copy number scan identifies complex rearrangements at 17q21.31 in dyslexics. *Am J Med Genet B Neuropsychiatr Genet* 165B, 572-580.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The sequence of the human genome. *Science* 291, 1304-1351.

Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 4, e1000214.

Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. *Nature* 461, 199-205.

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.

Wainberg, M., Sinnott-Armstrong, N., Knowles, D., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, J.L., Rivas, M.A., *et al.* (2017). Vulnerabilities of transcriptome-wide association studies. *bioRxiv*, 206961.

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., *et al.* (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320, 539-543.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., *et al.* (2018a). Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362, eaat8464.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., *et al.* (2018b). Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., *et al.* (2018c). Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362, eaat8464.

Wang, E., Dimova, N., and Cambi, F. (2007). PLP/DM20 ratio is regulated by hnRNPH and F and a novel G-rich enhancer in oligodendrocytes. *Nucleic Acids Res* 35, 4164-4178.

Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 30, 1095-1106.

Weinberger, D.R. (1987). Implications of normal brain development for the pathogenesis of schizophrenia. *Arch Gen Psychiatry* 44, 660-669.

Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., *et al.* (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 358, 667-675.

Werling, D.M., Pochareddy, S., Choi, J., An, J.-Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A.M.M., *et al.* (2019). Whole-genome and RNA sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *bioRxiv*, 585430.

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

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.

Windén, 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.

Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., *et al.* (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 538, 523-527.

Xu, B., Ionita-Laza, I., Roos, J.L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J.A., and Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* 44, 1365-1369.

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

Yang, Y.C., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., and Lu, Z.J. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* 16, 51.

Yuen, R.K., 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.

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

Zhang, F., Wang, G., Shugart, Y.Y., Xu, Y., Liu, C., Wang, L., Lu, T., Yan, H., Ruan, Y., Cheng, Z., *et al.* (2014a). Association analysis of a functional variant in ATXN2 with schizophrenia. *Neurosci Lett* 562, 24-27.

Zhang, X., Joehanes, R., Chen, B.H., Huan, T., Ying, S., Munson, P.J., Johnson, A.D., Levy, D., and O'Donnell, C.J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet* 47, 345-352.

Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keefe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., *et al.* (2014b). An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* 34, 11929-11947.

Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S., Li, G., *et al.* (2016). Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* 89, 37-53.

Zheng, S., and Black, D.L. (2013). Alternative pre-mRNA splicing in neurons: growing up and extending its reach. *Trends Genet* 29, 442-448.

Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.J., *et al.* (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* 49, 139-145.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 12, 7-18.

Zhou, W., Shi, Y., Li, F., Wu, X., Huai, C., Shen, L., Yi, Z., He, L., Liu, C., and Qin, S. (2018). Study of the association between Schizophrenia and microduplication at the 16p11.2 locus in the Han Chinese population. *Psychiatry Res* 265, 198-199.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* 9, e1003264.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., *et al.* (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 48, 481-487.

Zody, M.C., Jiang, Z., Fung, H.C., Antonacci, F., Hillier, L.W., Cardone, M.F., Graves, T.A., Kidd, J.M., Cheng, Z., Abouelleil, A., *et al.* (2008). Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* 40, 1076-1083.

Zwack, M.E., Cutler, D.J., and Chakravarti, A. (2000). Patterns of genetic variation in Mendelian and complex traits. *Annu Rev Genomics Hum Genet* 1, 387-407.