UCLA
UCLA Electronic Theses and Dissertations

Title

Gene regulation in the human brain and the biological mechanisms underlying psychiatric disorders

Permalink

https://escholarship.org/uc/item/9v08q5f7

Author

Kim, Minsoo

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Gene regulation in the human brain and the biological mechanisms

underlying psychiatric disorders

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Human Genetics

by

Minsoo Kim

2022

ABSTRACT OF THE DISSERTATION

Gene regulation in the human brain and the biological mechanisms

underlying psychiatric disorders

by

Minsoo Kim

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2022

Professor Michael Jeffrey Gandal, Chair

Large-scale genome-wide association studies (GWAS) have successfully identified hundreds of common variants associated with psychiatric disorders. Meanwhile, large-scale whole exome sequencing (WES) studies that test rare variants in aggregate have identified hundreds of genes for which loss-of-function is associated with increased risk of neurodevelopmental disorders. To understand the exact biological mechanisms in the human brain through which these genetic risk factors impart risk for disease, the first step in post-GWAS analyses is to identify genes that are regulated by disease-associated variants. Thereafter, the functional roles of disease genes from either GWAS or WES studies can be studied in an appropriate experimental system adequately recapitulating the disease-relevant context. In this dissertation, we primarily focus on this two-stage process of identifying genes and understanding their biological functions (herein referred to as the "variant-to-gene-to-function" problem). We approach this problem by integrating existing,

large-scale genotype array and postmortem human brain RNA-seq data from PsychENCODE. The first chapter focuses on proof-of-concept isoform-resolution expression analyses, where we fine-map the most significant GWAS locus for autism spectrum disorder and implicate reduced expression of a specific isoform of the *XRN2* gene as the underlying driver of GWAS signal. In this chapter, we also jointly model constituent isoforms of a given gene with multivariate variance components linear mixed models, enabling systematic dissection of the genetic architecture of isoform-level expression in the human brain, for the first time. The second chapter shifts its focus to characterizing the functional role of the high-confidence schizophrenia risk gene *C4A* in the human brain by investigating the effect of complex structural variation of *C4* genes on gene expression and co-expression. By annotating changes in *C4A* co-expression, we find putative molecular correlates of synaptic pruning and convergence of schizophrenia polygenic effects in synaptic processes, indicating that neuronal and synaptic pathways are the driving force conferring schizophrenia risk. Altogether, this dissertation aims to refine and advance our understanding of mental illnesses by characterizing the neurobiological mechanisms through which known genetic risk factors contribute to psychiatric disorders.

The dissertation of Minsoo Kim is approved.

Daniel H. Geschwind

Bogdan Pasaniuc

Valerie A. Arboleda

Michael Jeffrey Gandal, Committee Chair

University of California, Los Angeles

2022

I dedicate this thesis to my parents and my sister. Without their support,

the completion of this work would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Michael J. Gandal for all the guidance and encouragement he has given me over the years. He has believed in me ever since I joined his lab and provided me with unconditional support and mentorship. I have learned so much from working with him and watching him closely. His intellectual drive and enthusiasm have been inspiring and contagious like no other. I will forever cherish stimulating discussions we have had in our weekly one-on-one meetings, and of course, his sense of humor.

I would like to show my gratitude to the members of the Gandal lab for our time together as well as helpful comments and discussions throughout my Ph.D. training. I feel lucky to have been a part of this capable and generous group of people. I would like to thank the PsychENCODE consortium for the dataset they have generated, which has been pivotal for my doctoral work. I would like to also thank Dr. Hua Zhou for introducing me to the Julia programming language and helping me implement the estimation step in variance components model. Finally, I am grateful to my family for their perpetual support.

Chapter 2 is a version of the following paper:

Kim, M., Jops, C. T., and Gandal, M. J. (2022), "GeneticsMakie.jl: A versatile and scalable toolkit for visualizing locus-level genetic and genomic data," *bioRxiv*. https://doi.org/10.1101/2022.04.18.488573.

Chapter 3 is a manuscript in preparation:

Kim, M., Vo, D. D., Jops, C. T., Wen, C., Patowary, A., Bhattacharya, A., Yap, C., Zhou, H., and Gandal, M. J. (2022), "Multivariate analysis of genetic influences on brain isoform expression uncovers novel psychiatric disease mechanisms," *In preparation*.

Chapter 4 is a version of the following paper:

Kim, M., Haney, J. R., Zhang, P., Hernandez, L. M., Wang, L.-K., Perez-Cano, L., Loohuis, L. M. O., de la Torre-Ubieta, L., and Gandal, M. J. (2021), "Brain gene co-expression networks link complement signaling with convergent synaptic pathology in schizophrenia," *Nature neuroscience*, 24, 799–809.

EDUCATION

2011–2015   B.S., Applied and Computational Mathematics + Bioengineering, Caltech

2016–2024   M.D. Candidate, David Geffen School of Medicine at UCLA

2018–2022   Ph.D. Candidate, Human Genetics, UCLA

PUBLICATIONS

Gandal, M. J., Zhang, P., Hadjimichael, E., Walker, R. L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., Shieh, A. W., Haney, J., Parhami, S., Belmont, J., **Kim, M.**, Moran Losada, P., Khan, Z., Mleczko, J., Xia, Y., Dai, R., Wang, D., Yang, Y. T., Xu, M., Fish, K., Hof, P. R., Warrell, J., Fitzgerald, D., White, K., Jaffe, A. E., PsychENCODE Consortium, Peters, M. A., Gerstein, M., Liu, C., Iakoucheva, L. M., Pinto, D., and Geschwind, D. H. (2018), "Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder," *Science*, 362. https://doi.org/10.1126/science.aat8127.

Hernandez, L. M., **Kim, M.**, Hernandez, C., Thompson, W., Fan, C. C., Galván, A., Dapretto, M., Bookheimer, S. Y., Fuligni, A., and Gandal, M. J. (2022), "Decoupling Sleep and Brain Size in Childhood: An Investigation of Genetic Covariation in the Adolescent Brain Cognitive Development Study," *Biological psychiatry global open science*. https://doi.org/10.1016/j.bpsgos.2021.12.011.

Hernandez, L. M.*, **Kim, M.***, Hoftman, G. D., Haney, J. R., de la Torre-Ubieta, L., Pasaniuc, B., and Gandal, M. J. (2021), "Transcriptomic Insight Into the Polygenic Mechanisms Underlying Psychiatric Disorders," *Biological psychiatry*, 89, 54–64.

Hernandez, L. M., **Kim, M.**, Zhang, P., Bethlehem, R. A.I., Hoftman, G. D., Loughnan, R., Smith, D., Bookheimer, S. Y., Fan, C. C., Bearden, C. E., Thompson, W. K., and Gandal, M. J. (2022), "Multi-ethnic, phenome-wide association of *C4* variation with psychiatric and brain developmental phenotypes in youth," *Under review*.

Kim, M., Haney, J. R., Zhang, P., Hernandez, L. M., Wang, L.-K., Perez-Cano, L., Loohuis, L. M. O., de la Torre-Ubieta, L., and Gandal, M. J. (2021), "Brain gene co-expression networks link complement signaling with convergent synaptic pathology in schizophrenia," *Nature neuroscience*, 24, 799–809.

Kim, M., Jops, C. T., and Gandal, M. J. (2022), "GeneticsMakie.jl: A versatile and scalable toolkit for visualizing locus-level genetic and genomic data," *bioRxiv*. https://doi.org/10.1101/2022.04.18.488573.

Kim, M., Vo, D. D., Jops, C. T., Wen, C., Patowary, A., Bhattacharya, A., Yap, C., Zhou, H., and Gandal, M. J. (2022), "Multivariate analysis of genetic influences on brain isoform expression uncovers novel psychiatric disease mechanisms," *In preparation*.

Mullins, N., Forstner, A. J., O'Connell, K. S., Coombes, B., Coleman, J. R. I., …, **Kim, M.**, …, Stahl, E. A., McQuillin, A., Di Florio, A., Ophoff, R. A., and Andreassen, O. A. (2021), "Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology," *Nature genetics*, 53, 817–829.

Mullins, N., Kang, J., Campos, A. I., Coleman, J. R. I., Edwards, A. C., …, **Kim, M.**, …, Stahl, E. A., Stein, M. B., Streit, F., Willour, V., and Ruderfer, D. M. (2022), "Dissecting the Shared Genetic Architecture of Suicide Attempt, Psychiatric Disorders, and Known Risk Factors," *Biological psychiatry*, 91, 313–327.

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., …, **Kim, M.**, …, Sullivan, P. F., Ripke, S., Walters, J. T. R., O'Donovan, M. C., and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2022), "Mapping genomic loci implicates genes and synaptic biology in schizophrenia," *Nature*, 604, 502–508.

CHAPTER 1

Introduction

Large-scale psychiatric genetics studies have identified hundreds of genetic risk factors associated with increased risk of common, but complex neuropsychiatric disorders such as autism spectrum disorder (ASD) and schizophrenia (SCZ). For example, genome-wide association studies (GWAS) that survey common variants have identified hundreds of genomic regions associated with SCZ (Trubetskoy et al. 2022) as well as dozens for ASD (Grove et al. 2019), bipolar disorder (BD) (Mullins et al. 2021), and major depressive disorder (MDD) (Wray et al. 2018). At the same time, whole-exome sequencing (WES) studies that survey rare variants have identified hundreds of genes for which loss-of-function is implicated in ASD (Satterstrom et al. 2020), SCZ (Singh et al. 2022), and developmental delay disorders (DDD) (Kaplanis et al. 2020). Understanding how these genetic risk factors combined with environmental factors converge at the level of disorder-relevant biological pathways within larger brain circuitry is critical for gaining insights into disease pathophysiology and identifying novel targets for therapeutic development (Gandal et al. 2016). However, the road to translating genetic findings is hampered by the "variant-to-gene-to-function" problem, which we define below and can be decomposed further into two separate problems—the "variant-to-gene" problem and the "gene-to-function" problem.

The variant-to-gene problem applies to GWAS, in which a single genetic variation is tested at a time, resulting in in all variation that is correlated with the causal variant(s) harboring a significant association. As such, within a GWAS locus, it is unclear which of the many associated variants are the true causal variant(s). Furthermore, because most genetic variants are non-coding, even if

the causal variants are identified, it is often unclear which genes are regulated by disease-associated genetic variants. One common bioinformatic approach to address this problem is to first identify variants associated with changes in different molecular phenotypes (e.g. gene expression, local splicing, RNA editing, DNA methylation status, and protein/metabolite levels) (Aygün et al. 2021; Breen et al. 2019; Collado-Torres et al. 2019; Cui et al. 2021; Fromer et al. 2016; Gandal et al. 2018b; Garrido-Martín et al. 2021; GTEx Consortium et al. 2017; GTEx Consortium 2020; Hannon et al. 2016; Jaffe et al. 2018; Li et al. 2019; Lopes et al. 2022; Mittleman et al. 2020; Ng et al. 2017; O'Brien et al. 2018; Panyard et al. 2021; Park et al. 2021; Raj et al. 2018; Sng et al. 2019; Takata et al. 2017; Walker et al. 2019; Wang et al. 2018; Wingo et al. 2021; Xiong et al. 2021; Yang et al. 2021; Zhang et al. 2020a; Zhang et al. 2020b) and then statistically evaluate their relationship to known GWAS signals (Giambartolomei et al. 2014; Gusev et al. 2016; Zhu et al. 2016). Another common approach is to fine-map the GWAS locus first and then investigate the effect of candidate causal variants on molecular phenotypes. We note that this latter approach is susceptible to the issues of unobserved genetic variation. That is, most (if not all) fine-mapping methods (Schaid et al. 2018) start with an underlying assumption that the causal variant is present in association statistics, which may not be the case when unobserved SNP genotypes or structural variations—such as deletion, duplication (Scott et al. 2021), multiallelic copy number variation (Sekar et al. 2016), inversion (Boettger et al. 2012), translocation, and mobile element insertion (Sekar et al. 2016)—reflect the true underlying signals. These issues can be alleviated by cataloguing all common genetic variation including structural variation and using dense genotype array. Additionally, there are high-throughput experimental methods to identify causal variants and their cognate genes (e.g. massively parallel reporter assays), but this approach can also suffer from the above issues related to unobserved genetic variation.

APOE locus

Figure 1.1: A close look at the *APOE* locus using GeneticsMakie.jl. GWAS results for 56 complex phenotypes are shown, which span autoimmune, endocrine, psychiatric, cardiovascular disorders, and cancer. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding linkage disequilibrium (LD) between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *APOE* gene. Note that -$\log_{10} P$ values for Alzheimer disease are clamped to 308, since their P values cannot be represented by the smallest floating-point number. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), RBC (red blood cell), SCZ (schizophrenia).

Compared to GWAS, WES studies provide a list of robust disease genes, which makes them more interpretable and bypass the variant-to-gene problem. However, the prioritized genes from both GWAS and WES studies still need to confront the gene-to-function problem, where the precise functional role of these genes in disease-relevant tissue/cells and developmental time periods must be elucidated. Naturally, this requires an experimental system that adequately recapitulates the disease-relevant context, which is difficult for most human diseases, particularly brain-related disorders. Due to the rapid pace of gene discovery, for example from WES studies (Kaplanis et al. 2020; Satterstrom et al. 2020; Singh et al. 2022), this problem has essentially become the rate-limiting step. A classic example that illustrates the immense difficulty of the gene-to-function problem is *APOE* gene. It has been almost thirty years since the initial discovery of the strong association between *APOE* gene and Alzheimer disease (**Figure 1.1**), where two copies of the e4 allele increase the risk of disease by more than ten-fold. However, how the e4 allele drives development and progression of Alzheimer disease is very much unknown. There are biologically plausible speculations, but the exact mechanisms remain nebulous.

Along these lines, the overarching theme of this dissertation is to tackle the variant-to-gene-to-function problem—discovering high-confidence risk genes and corresponding biological mechanisms that contribute to disease risk—through integration and analysis of large-scale genome-wide genotype array and next-generation RNA-seq data.

In chapter 2, we describe an open source, high-performance Julia package GeneticsMakie.jl for visualizing high-dimensional, multivariate genetic and genomic data (**Figure 1.1**). The goal of this package is to facilitate exploratory data analyses and generation of novel hypotheses. Visual inspection of the underlying data is often more powerful than any type of statistics, even more so when appropriate statistical and computational methods are not available. A suite of functions provided by GeneticsMakie.jl is heavily used in subsequent chapters for data visualization.

In chapter 3, we highlight the potential of leveraging isoform-level expression in prioritizing candidate causal genes in GWAS loci. Conventionally, bulk-level RNA-seq analyses have focused on gene-level expression (Hernandez et al. 2021) and local splicing patterns (Katz et al. 2010; Li et al. 2018b; Shen et al. 2014; Vaquero-Garcia et al. 2016), when isoforms in fact are the fundamental biological units that are expressed in cells. This is mainly because the human transcriptome annotation is incomplete, and as a result probabilistically assigning multi-mapped short-read RNA-seq reads by maximizing the log-likelihood function with the expectation-maximization (EM) algorithm can yield biased estimates (Bray et al. 2016; Li and Dewey 2011; Sterne-Weiler et al. 2018). In the simplest scenario, RNA-seq reads that belong to an "unknown isoform" might get erroneously assigned to different isoforms that are present in the input transcriptome annotation. This means that with more complete transcriptome annotations, we can acquire more accurate estimates of isoform-level expression even with short-read RNA-seq (Sterne-Weiler et al. 2018). In chapter 3, despite this important caveat to accurately assessing isoform abundances, we partake in an intellectual endeavor to quantify the degree of polygenicity and pleiotropy among isoforms by jointly modeling isoform expression of a given gene with multivariate variance components linear mixed models. To fit these multivariate variance

components models, we share another Julia package that implements the minorization-maximization (MM) algorithm (Zhou et al. 2019). We partition genetic variances and covariances between *cis*- and *trans*-SNPs, and hence we estimate heritability attributable to *cis*- and *trans*-SNPs as well as two separate genetic correlation parameters among *cis*- and *trans*-SNPs. This is a proof-of-concept analysis that lays the groundwork for future with improved transcriptome annotations or higher-quality long-read data. Subsequently, we focus on several concrete examples of changes in isoform-level expression that drive GWAS signals, highlighting the importance of isoform-resolution analyses. The most notable example is *XRN2* gene, a specific isoform of which colocalizes strongly with one of top two GWAS loci for ASD (Grove et al. 2019).

In chapter 4, we confront the gene-to-function problem, focusing on the *C4A* gene. The most significant GWAS signal for SCZ lies within the MHC region (Trubetskoy et al. 2022), which has been shown to partly reflect common but complex structural variation of the *C4* locus (Sekar et al. 2016). Specifically, this locus harbors multiallelic copy number variation, where human *C4* encoded by two genes—*C4A* and *C4B*—can exist in different combinations of copy numbers. Previous work demonstrated that increased *C4A* copy number and resulting elevated *C4A* expression is associated with SCZ risk (Sekar et al. 2016). The strength and novelty of the *C4A* association has prompted speculation that *C4A* and the broader complement system may represent a key disease pathway in SCZ pathophysiology. In chapter 4, we directly test this hypothesis by characterizing co-expression partners of *C4A* that are either positively or negatively correlated with *C4A* expression across varying *C4A* genomic copy numbers, annotating their cell-type and pathway contributions as well as their relation to established SCZ genetic risk factors. This type of "seeded" network approach can provide an unbiased functional annotation for a poorly

understood gene by capturing coherent biological processes that covary across samples (Parikshak et al. 2015). We identify a putative transcriptomic signature of *C4A*-mediated synaptic pruning, reinforcing the idea that over-pruning likely contributes to SCZ pathogenesis and/or progression (Feinberg 1982; Sekar et al. 2016). More importantly, we find that negatively correlated genes with *C4A* are over-represented for neuronal and synapse-related pathways, which in turn are enriched for SCZ genetic signals, suggesting that synaptic pathways are the key biological pathways underlying risk for mental illnesses. While both chapters 3 and 4 are broadly concerned with understanding the molecular effects of genetic variation, chapter 3 focuses on molecular effects on gene and isoform expression, whereas chapter 4 takes one step further and focuses on molecular effects of genetic variation on gene co-expression (Lea et al. 2019; van der Wijst et al. 2018).

Of note, each chapter is self-contained in terms of abbreviations and flow of the material, although the references are combined and presented at the end. We also note that no new data have been generated as part of this dissertation, since we leverage existing resources, particularly the results of a mega-analysis of ~2,000 postmortem human brain samples from PsychENCODE. The PsychENCODE consortium has previously aggregated and uniformly processed large numbers of samples across different studies to maximize sample size and hence power to detect biological signals (Gandal et al. 2018b, Wang et al. 2018). Rigorous quality control (QC) has been previously applied to both genetic (SNP array) and transcriptomic (RNA-seq) data to account for batch effects, technical artifacts, and other unwanted sources of variation.

CHAPTER 2

GeneticsMakie.jl: A versatile and scalable toolkit for visualizing locus-level genetic and genomic data

2.1 Abstract

With the continued deluge of results from genome-wide association and functional genomic studies, it has become increasingly imperative to quickly combine and visualize different layers of genetic and genomic data within a given locus to facilitate exploratory and integrative data analyses. While several tools have been developed to visualize locus-level genetic results, the limited speed, scalability, and flexibility of current approaches remains a significant bottleneck. Here, we present a Julia package for high-performance genetics and genomics-related data visualization that enables fast, simultaneous plotting of hundreds of association results along with multiple relevant genomic annotations. Leveraging the powerful plotting and layout utilities from Makie.jl facilitates the customization and extensibility of every component of a plot, enabling generation of publication-ready figures. The GeneticsMakie.jl package is open source and distributed under the MIT license via GitHub (https://github.com/mmkim1210/GeneticsMakie.jl). The GitHub repository contains installation instructions as well as examples and documentation for built-in functions.

2.2 Introduction

The last decade has seen an exponential increase in the volume of large-scale genetic association results such as those from genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS). The rapid advancements in high-throughput sequencing

technologies have further led to an increase in the volume and diversity of molecular genomic readouts, such as 3D genome contacts, ChIP-seq, and ATAC-seq. As these efforts continue to scale, it is becoming increasingly critical to develop efficient ways for simultaneous visualization and integration of multiple such datasets to develop an intuitive understanding of potential underlying biological relationships.

Several tools have been developed to visualize genetic association results within a specific locus along with corresponding genomic annotations, exemplified by the original "LocusZoom" style plots (Pruim et al. 2010). Multiple extensions to these LocusZoom style plots have since been built, spanning a wide array of programming languages, including JavaScript, Python, and R (Boughton et al. 2021; Dadaev et al. 2016; Jorgenson et al. 2009; Juliusdottir et al. 2018; Kramer et al. 2022; Machiela and Chanock 2015, 2018; Schilder et al. 2021). However, efficient customization and extension is limited with these tools, and in general they are not suitable for parallel visualization of large numbers of data points. For example, only a certain genomic range can be shown or a certain number of genes can be plotted (see **Figure 2.1** for more complete comparisons). Certain packages have a lot of dependencies which can further lead to decreased flexibility.

Ideally, one would be able to visualize GWAS loci across multiple phenotypes and multiple ancestries simultaneously, with layered visualization of molecular QTL results across multiple tissues/cells with ease and speed. The Julia programming language is an optimal platform to achieve this goal (Bezanson et al. 2017) by providing performance of a low-level language while retaining the readability and ease-of-use of a high-level language. Makie.jl (Danisch and Krumbiegel 2021) is a Julia plotting package that provides powerful plotting utilities and recipes

that can be easily extended to visualize most (if not all) genomic data. Although Makie.jl is one of many Julia plotting packages, it is particularly performant, and it is distinguished from other plotting packages (and their extensions) in that it comes with advanced layout tools. In fact, using Makie.jl's flexible layout tools, it can be also almost effortless to combine and plot various genetic and genomic data with complex layouts.

	GeneticsMakie.jl	LocusZoom.js	R ecosystem
Munge GWAS summary statistics	Yes	–	Yes (MungeSumstats)
Plot an arbitrary genomic range	Yes	–	Yes
Plot an arbitrary number of genes	Yes	–	–
Color SNPs by LD	Yes	Yes	Yes (bigsnpr + custom Rcpp code)
Minimal number of dependencies	Yes	–	–
Plot various functional genomic data modalities (e.g. Hi-C matrix, enhancer-promoter loops)	–	–	Yes (plotgardener)
Can visualize the entire ~10 Mb extended MHC region	Yes	–	Yes
Programmatically automate drawing many figures	Yes	–	Yes
Drawing figures with a very complex layout	Yes	–	–
Interactivity	–	Yes	–

Figure 2.1: Comparison of functionalities provided by different programming environments for creating LocusZoom plots. Shown in parentheses are the names of packages that are suited for a particular function. We note that plotting functions not provided by the current version of GeneticsMakie.jl can be easily added in the future.

2.3   Plotting phenome-scale LocusZoom plots

Here, we present the Julia package GeneticsMakie.jl, which builds upon and extends Makie.jl's plotting tools to generate publication-quality figures visualizing multiple genetic and genomic data modalities on different layers, as shown for the *GCKR* locus (**Figure 2.2**). To ease this process, we provide functions for munging GWAS (or other association) summary statistics, which can come in various formats (Bulik-Sullivan et al. 2015; Lyon et al. 2021; Murphy et al. 2021) and therefore require harmonization. Once the summary statistics are munged, we recommend storing them as memory friendly Arrow files (using Arrow.jl package), since loading hundreds of genetic

10

GCKR locus

Figure 2.2: A close look at the known pleiotropic *GCKR* locus using GeneticsMakie.jl. GWAS results for 41 complex phenotypes are shown, which span autoimmune, endocrine, psychiatric, cardiovascular disorders, and cancer. *GCKR* locus is one of the most pleiotropic loci in the human genome along with MHC, *FADS1*, and *ABO* regions. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding linkage disequilibrium (LD) between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *GCKR* gene. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), SCZ (schizophrenia).

association results simultaneously is memory intensive and infeasible. Then one can iterate through arbitrary genomic regions of interest. For example, GeneticsMakie.jl conveniently provides functions for identifying GWAS loci and their closest (protein-coding) genes so that one can iterate through either GWAS loci or their cognate genes. To color SNPs by linkage disequilibrium (LD) with a designated SNP of interest (e.g. index SNP), any custom reference panel can be loaded using SnpArrays.jl package (Zhou et al. 2020) and LD is computed on the fly. Additionally, genes and isoforms with constituent exons and introns can be plotted with any custom transcriptome annotation file in GTF format. These functionalities form the backbone of LocusZoom plots and other functional genomic data can be added as separate layers as needed. It is worth nothing that all these functions can be customized or extended very easily with minimal loss of performance, which is an inherent strength of the Julia programming language.

To further showcase the power of GeneticsMakie.jl, we share phenome-wide LocusZoom plots for 239 GWAS loci for schizophrenia (Trubetskoy et al. 2022) which are defined as non-overlapping ±1 Mb windows around the most significantly associated SNPs. We also share such LocusZoom plots for genomic regions known to harbor long-range LD (Anderson et al. 2010), which includes the major histocompatibility complex (MHC) region (**Figure 2.3**). Finally, we share LocusZoom

MHC locus
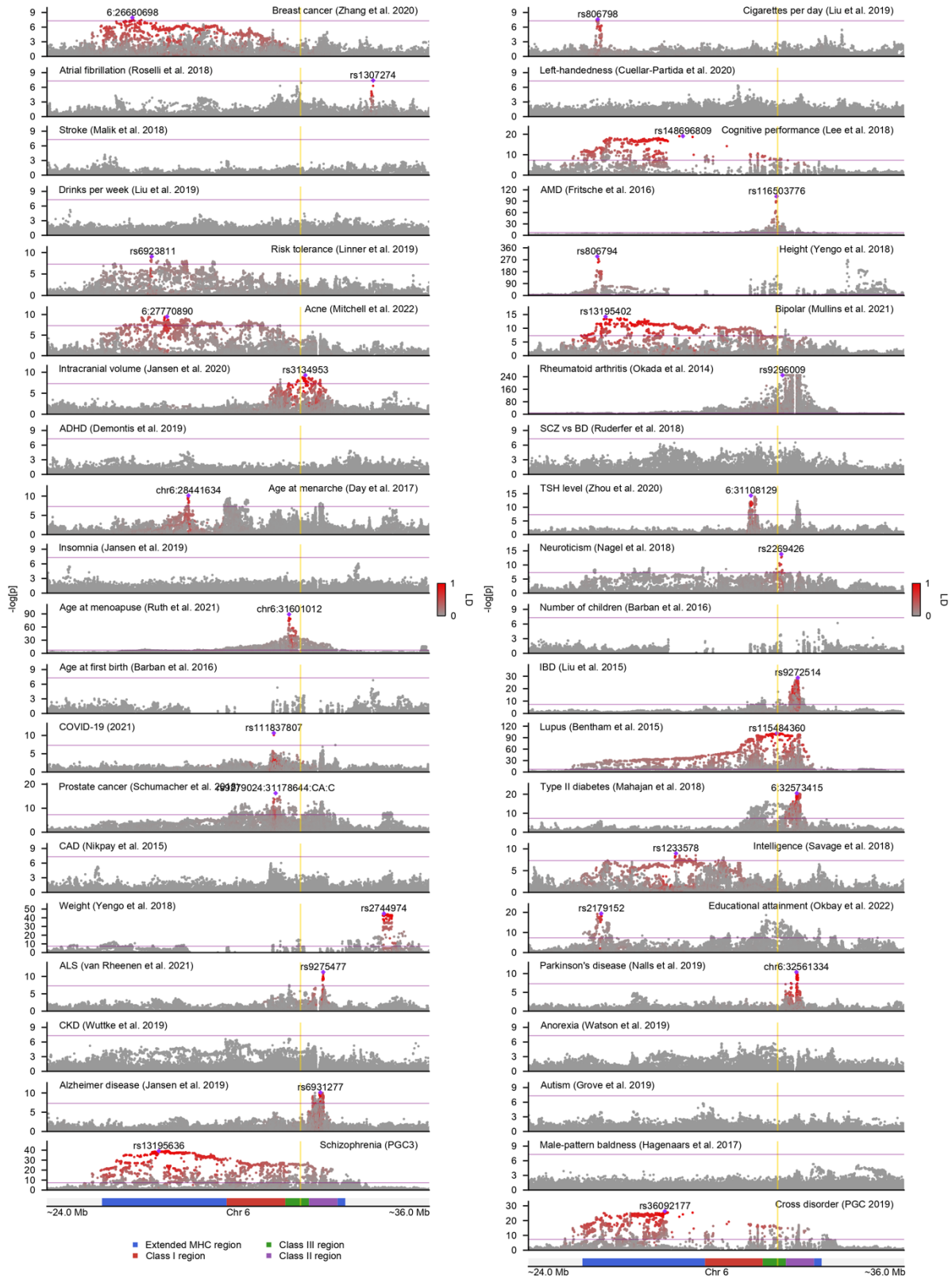
Figure 2.3: A close look at the extended MHC region using GeneticsMakie.jl. GWAS results for 41 complex phenotypes are shown. The extended MHC region is ~10 Mb long with essentially every phenotype harboring a GWAS hit. Although the entire genes that reside within this region can be added as a separate layer/track using GeneticsMakie.jl, that information is omitted due to space constraints. As a sanity check, one can observe that the most strongly associated SNP for age-related macular degeneration lies within the class III region near *CFB*, *C2* genes. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *C4A* gene. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), SCZ (schizophrenia).

plots for high-confidence neuropsychiatric risk genes, which loss-of-function is implicated in increased risk for schizophrenia (Singh et al. 2022), autism spectrum disorder (Satterstrom et al. 2020), and developmental delay disorder (Kaplanis et al. 2020). This permits qualitative assessment of convergence of common variant and rare variant signals for neurodevelopmental and neuropsychiatric disorders.

2.4 Notable findings from LocusZoom plots

Here, we share some emerging patterns from visual inspection of thousands of genomic regions phenome-wide. We first note that the probability of an arbitrary 2 Mb window harboring at least one genome-wide significant hit across 50 phenotypes is much higher than the probability of none of these phenotypes harboring any significant association. In other words, it is rarely the case that we do not detect any significant GWAS signal when looking at an arbitrary 2 Mb genomic window across 50 phenotypes. This observation suggests that genetic variation occurring throughout the human genome is structured and that it serves a purpose. Second, in a large-scale GWAS that is reaching saturation such as the one for height, it is common to observe allelic heterogeneity, where multiple non-correlated sets of variants are associated with the same phenotype within a given locus. These variants could affect either the same gene or different genes depending on the locus, and when affecting the same gene, they presumably play different biological roles.
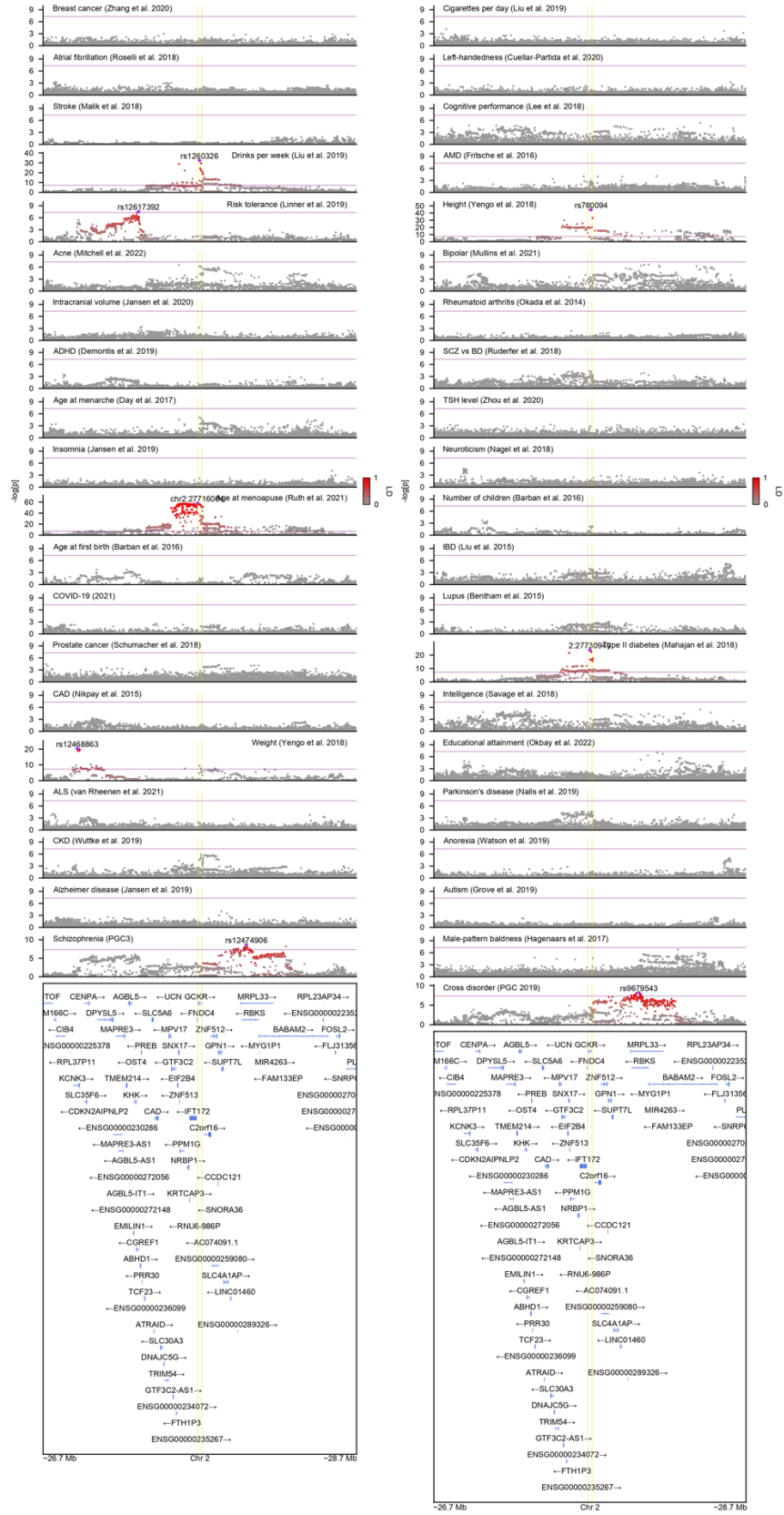
14

GRIN2A locus

Figure 2.4: A close look at the *GRIN2A* locus using GeneticsMakie.jl. GWAS results for 41 complex phenotypes are shown. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *GRIN2A* gene. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), SCZ (schizophrenia).

Third, examples of allelic series abound phenome-wide, where different sets of variants of a given gene are associated with different phenotypes. One example is *GRIN2A* gene, which is a high-confidence schizophrenia risk gene (Singh et al. 2022) that harbors a GWAS signal in the 3' end for schizophrenia and bipolar disorder and other distinct GWAS signals in the 5' end for cognitive performance and educational attainment (**Figure 2.4**). This type of observation is likely missed by PheWAS plots, where association results for only a single variant across multiple phenotypes are visualized, and hence missing other GWAS signals nearby. Another example is *XPO7* gene, which is again a high-confidence schizophrenia risk gene (Singh et al. 2022) that harbors GWAS signals for atrial fibrillation and height (**Figure 2.5**). As can be seen in the *XPO7* locus (**Figure 2.5**), the LD blocks oftentimes tightly hug the gene boundaries (McVean et al. 2004). Note that the LD blocks can hug more than a single gene and hugging multiple genes is quite frequent. Finally, there is huge variability in gene density and LD complexity across different genomic regions, which lends to a conclusion that it might be better to focus our initial gene prioritization efforts on the low complexity regions first, and then start to tackle more complex regions.

2.5   Cautions in interpreting LocusZoom plots

Oftentimes, GWAS loci can harbor extremely small P values, in which cases the P values cannot be represented by a floating-point number and hence set to zero. GeneticsMakie.jl mitigates this issue by clamping P values of such SNPs to the smallest floating-point number, when munging summary statistics. Such cases tend to be more common in phenotypes that are reaching saturation

XPO7 locus

Breast cancer (Zhang et al. 2020)
Atrial fibrillation (Roselli et al. 2018) — rs7846485
Stroke (Malik et al. 2018)
Drinks per week (Liu et al. 2019)
Risk tolerance (Linner et al. 2019)
Acne (Mitchell et al. 2022)
Intracranial volume (Jansen et al. 2020)
ADHD (Demontis et al. 2019)
Age at menarche (Day et al. 2017)
Insomnia (Jansen et al. 2019)
Age at menoapuse (Ruth et al. 2021)
Age at first birth (Barban et al. 2016)
COVID-19 (2021)
Prostate cancer (Schumacher et al. 2018)
CAD (Nikpay et al. 2015)
Weight (Yengo et al. 2018)
ALS (van Rheenen et al. 2021)
CKD (Wuttke et al. 2019)
Alzheimer disease (Jansen et al. 2019)
Schizophrenia (PGC3)

Cigarettes per day (Liu et al. 2019)
Left-handedness (Cuellar-Partida et al. 2020)
Cognitive performance (Lee et al. 2018) — rs1473634
AMD (Fritsche et al. 2016)
Height (Yengo et al. 2018) — rs7823327
Bipolar (Mullins et al. 2021)
Rheumatoid arthritis (Okada et al. 2014)
SCZ vs BD (Ruderfer et al. 2018)
TSH level (Zhou et al. 2020)
Neuroticism (Nagel et al. 2018)
Number of children (Barban et al. 2016)
IBD (Liu et al. 2015)
Lupus (Bentham et al. 2015)
Type II diabetes (Mahajan et al. 2018)
Intelligence (Savage et al. 2018) — rs1473634
Educational attainment (Okbay et al. 2022)
Parkinson's disease (Nalls et al. 2019)
Anorexia (Watson et al. 2019)
Autism (Grove et al. 2019)
Male-pattern baldness (Hagenaars et al. 2017)
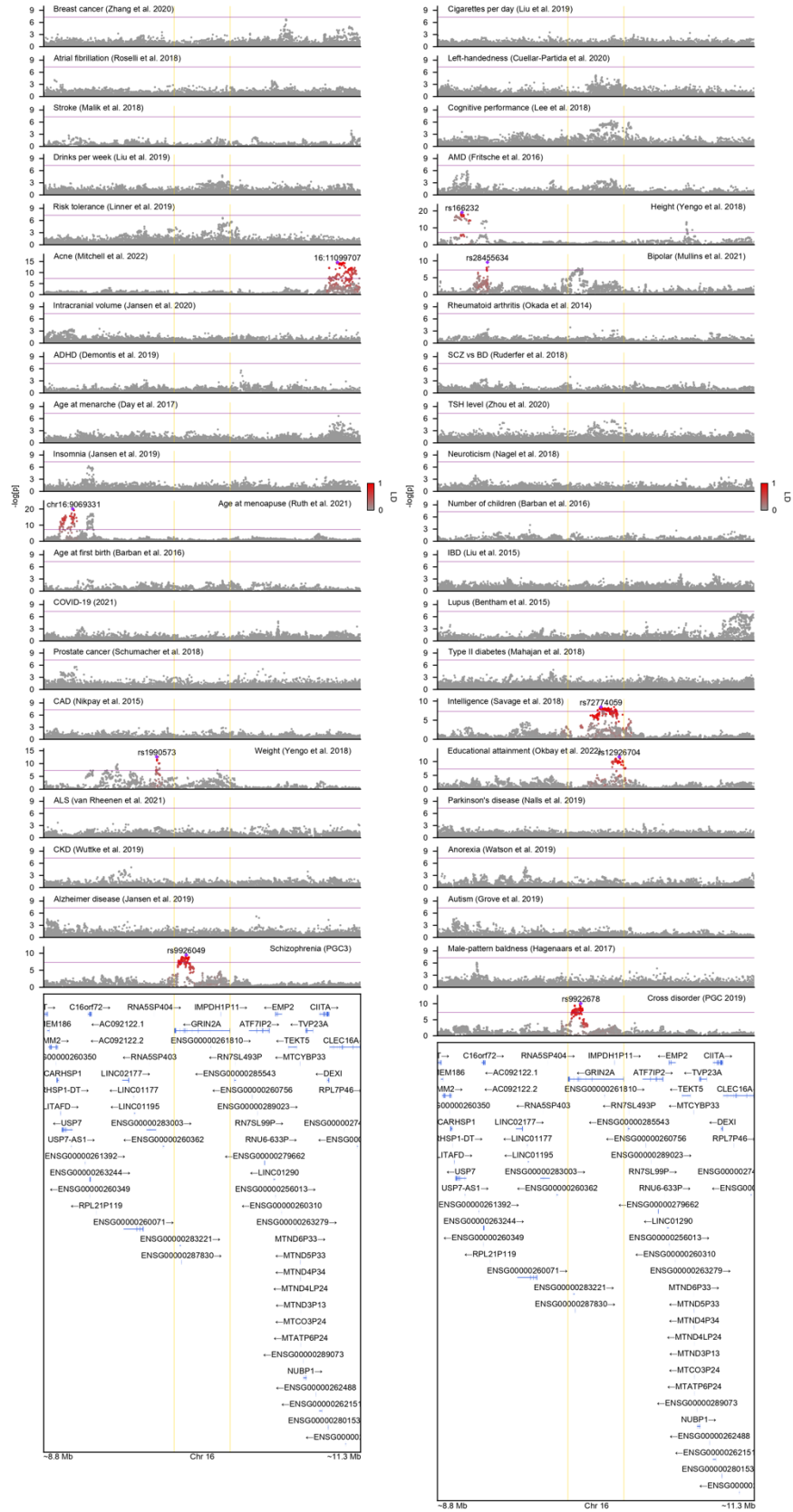Cross disorder (PGC 2019)

Figure 2.5: A close look at the *XPO7* locus using GeneticsMakie.jl. GWAS results for 41 complex phenotypes are shown. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *XPO7* gene. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), SCZ (schizophrenia).

in terms of GWAS discovery such as height and weight. In **Figure 2.3**, rheumatoid arthritis suffers from this case. To fundamentally address this issue, the *P* values in GWAS summary statistics need to be shared in a -$\log_{10}$ scale or it might be more appropriate to plot alternative measures of strength of association such as Z scores. Next, the gene names and SNP rsIDs frequently change over time, so one needs to be cognizant about using the appropriate gene names or rsIDs when drawing LocusZoom plots. For example, within the MHC region, the index SNP in a previous iteration of schizophrenia GWAS no longer has the same rsID in the most recent GWAS (**Figure 2.6**). As another example, *KIZ* gene has a different gene name of *PLK1S1* in previous versions of Gencode annotation (e.g. Gencode v19). To circumvent this issue, SNPs can be queried based on genomic positions with or without matching alleles, and the past gene names can be looked up in databases such as GeneCards (genecards.org). Lastly, GWAS summary statistics are extremely heterogeneous in the number of SNPs that they contain. In the *KANSL1* locus, several phenotypes such as stroke, height, and weight are missing a substantial proportion of SNPs (**Figure 2.7**). This has important consequences in that the direct comparison of GWAS results across phenotypes may be difficult for certain loci. Moreover, in the context of fine-mapping candidate causal SNPs, the assumption that the causal variant is present in the summary statistics (whether be it typed or imputed) is unrealistic and highly likely to be violated. One way of addressing this issue is to utilize and visualize GWAS results from biobank data, which are much more homogeneous in its SNP content. However, leveraging biobank data comes with its own limitations such as not having enough ascertained cases, in particular for psychiatric disorders.
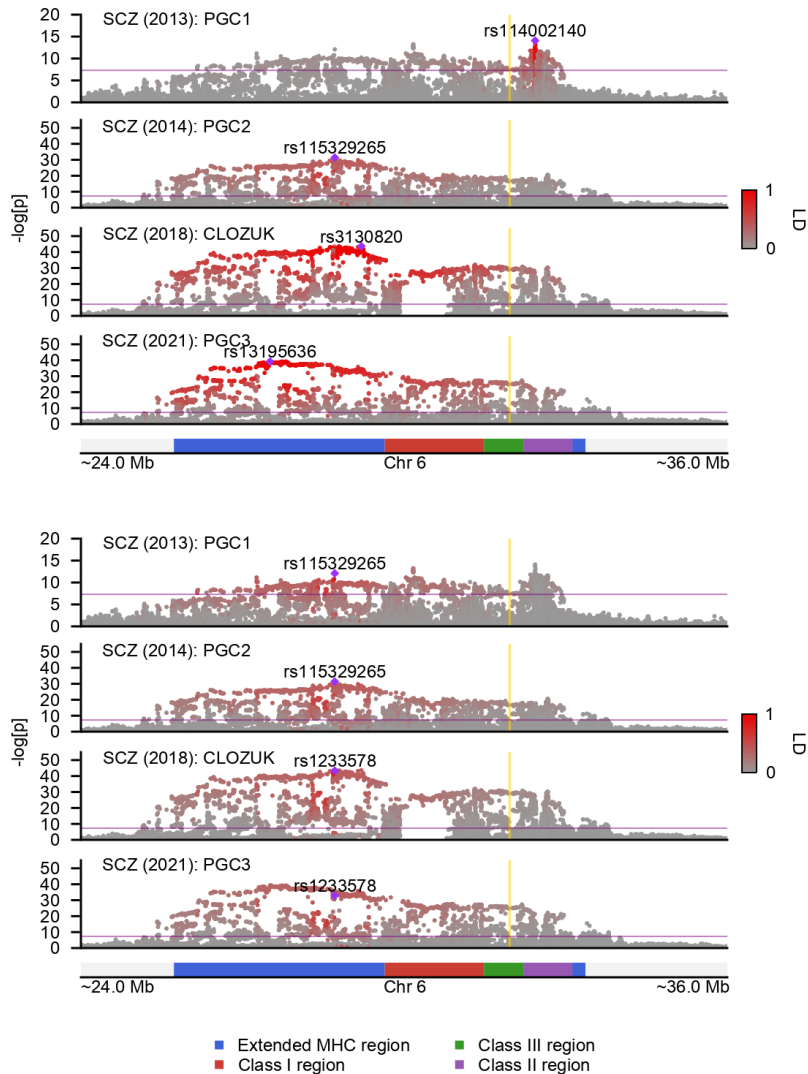
Figure 2.6: MHC association for schizophrenia with increasing sample size. Top, index SNPs for each GWAS are labeled and corresponding LD between other SNPs are displayed. Bottom, rs115329265 is used as a reference to calculate LD for all GWAS results. Note that rsID for rs115329265 is switched to rs1233578 in later versions of dbSNP and hence in subsequent GWAS results as well. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *C4A* gene.

2.6 Other usage

GeneticsMakie.jl further supports efficient generation of Manhattan plots and corresponding QQ

plots for GWAS summary statistics. In addition, it can be used to visualize gene-level association

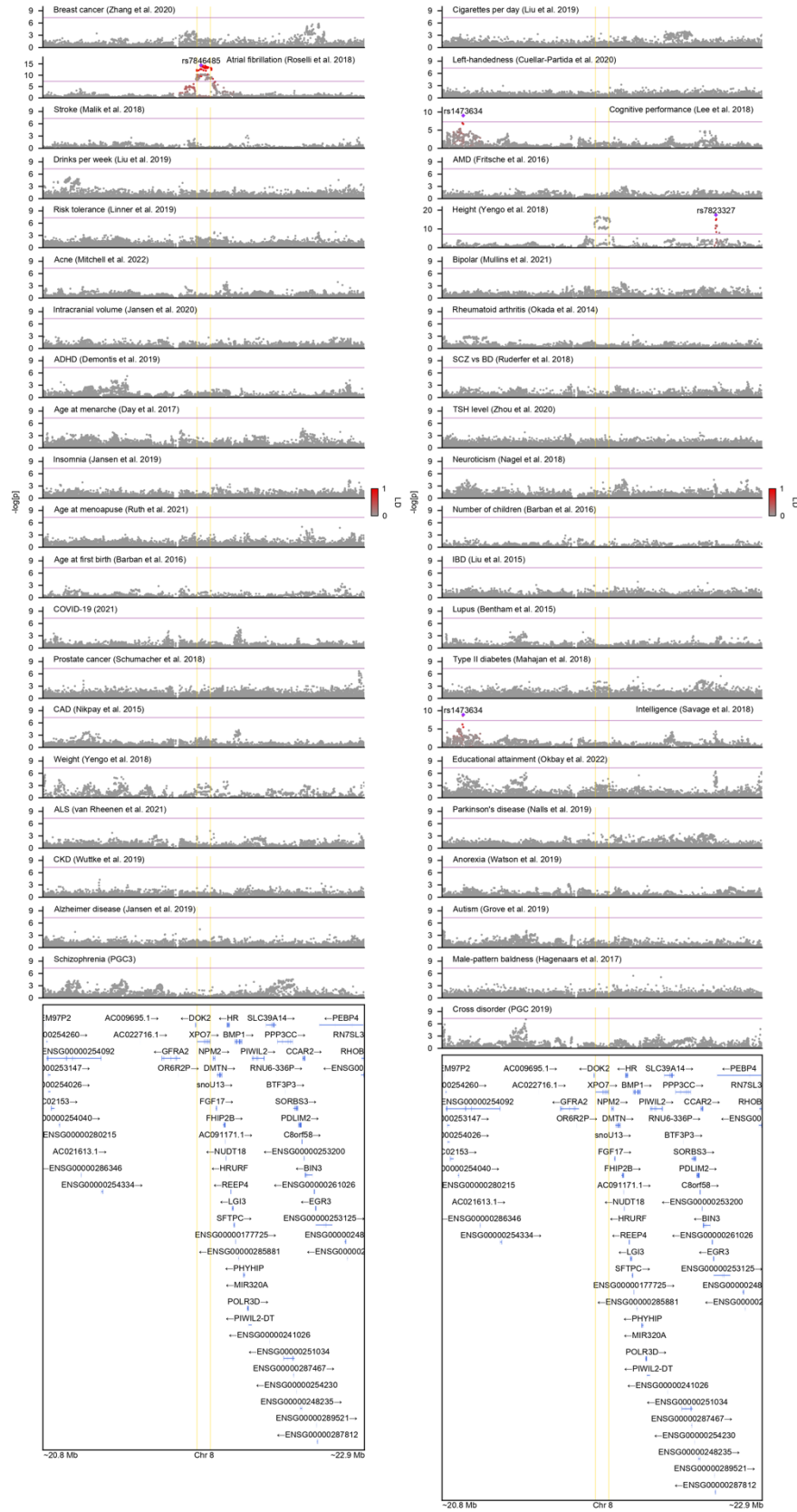results such as transcriptome-wide association studies (TWAS). GeneticsMakie.jl was written with

KANSL1 locus
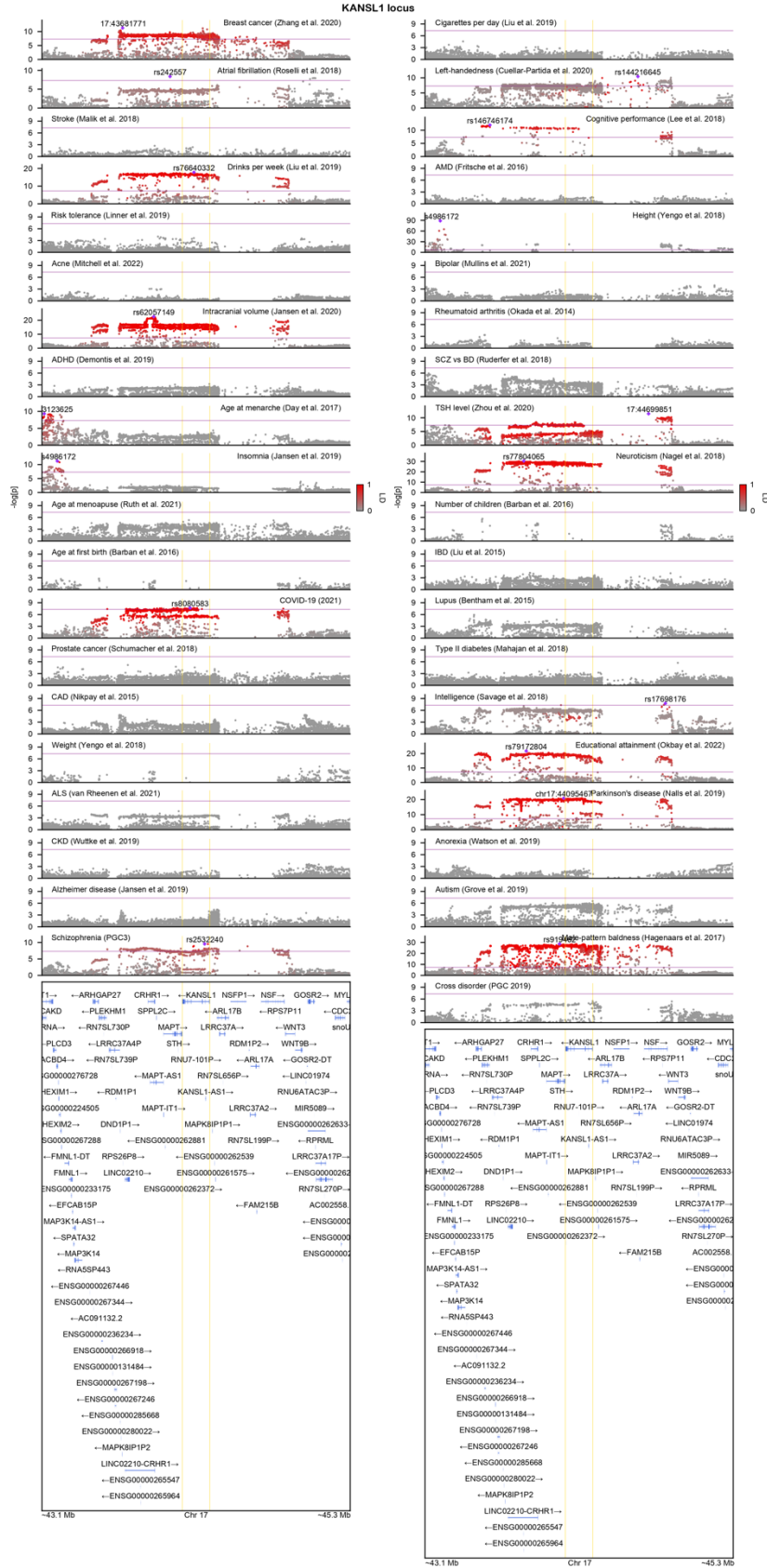
Figure 2.7: A close look at the *KANSL1* locus using GeneticsMakie.jl. GWAS results for 41 complex phenotypes are shown. This genomic region is characterized by 1 Mb inversion and other complex structural variation that affects multiple genes (Boettger et al. 2012). Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *KANSL1* gene. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), SCZ (schizophrenia).

plotting multiple phenotypes in mind such that direct comparison of genetic results across different phenotypes is possible. Visualizing LD blocks as well as plotting correlation results as typically done for reporting genetic correlation is also possible (Bulik-Sullivan et al. 2015). Other ways of reporting genetic correlation results such as forest plots can be easily drawn using existing plotting functions provided by Makie.jl. Along the same line, colocalization plots (Liu et al. 2019) or circos plots can also be drawn using Makie.jl, which we leave as an exercise to the readers.

2.7 Discussion

In summary, GeneticsMakie.jl allows scalable and flexible visual display of genetic and genomic data within the Julia ecosystem, taking LocusZoom plots to the next level. It produces high-quality, publication-ready figures by default. In the future, we envision other data modalities being plotted on top of what we already have implemented in GeneticsMakie.jl to provide better interpretation of underlying genetic association, hence facilitating the exploratory data analyses (EDA) and generation of novel hypotheses. This includes visualizing PheWAS results (Carroll et al. 2014; Gagliano Taliun et al. 2020) as well as other molecular readouts (Boix et al. 2021; Granja et al. 2021). It would be also worthwhile to implement interactive plots as usually provided by web browser-based tools (Geihs et al. 2015; Kwong et al. 2021).

CHAPTER 3

Multivariate analysis of genetic influences on brain isoform expression uncovers novel psychiatric disease mechanisms

3.1 Abstract

The intricate molecular mechanisms occurring within the human brain are under tight genetic control. Existing work has examined genetic influences mainly at the gene-level and thus the extent to which the potentially large number of distinct isoforms derived from each gene is under (shared) genetic influences remains unexplored. This is important as isoforms are particularly diverse in the human brain and the disruption in their gene regulation may contribute to the pathogenesis of psychiatric disorders. Here, we investigate the genetic architecture of brain isoform expression by jointly modeling them with multivariate variance components linear mixed models. We find a significant proportion of isoforms to be under genetic control with substantial shared genetic influences among local (or *cis*-) genetic variants. Importantly, a significant proportion of brain-expressed genes are found heritable only at the isoform-level. By integrating these isoform-specific genetic signals with psychiatric GWAS signals, we uncover previously hidden psychiatric disease mechanisms. In particular, we implicate reduced expression of a specific *XRN2* isoform as the underlying driver of the strongest GWAS signal for autism spectrum disorder.

3.2 Introduction

A single gene can give rise to multiple distinct mRNA molecules or isoforms through the process of alternative splicing. This mechanism has been recognized as an important factor in generating the molecular and functional complexity and diversity in human tissues with ~95% of multi-exonic

22

genes in humans subject to alternative splicing (Lieberman 2018). Aberrant splicing can have substantial effects on human health and is the pathogenic mechanism underlying a number of highly debilitating diseases, including spinal muscular atrophy (SMA) and Duchenne muscular dystrophy (DMD). Furthermore, alternative splicing and resulting isoform expression is often tissue-specific, and the human brain particularly exhibits a high degree of alternative splicing (GTEx Consortium 2020; Raj and Blencowe 2015; Südhof 2018) with brain-expressed genes tending to being longer and harboring more exons and isoforms than other genes (Choi and An 2021). Such ubiquitous and intricate nature of alternative splicing in the human brain hints at the integral role it plays in normal brain function as well as its potential to cause brain-related disorders.

Genome-wide association studies (GWAS) have identified hundreds of genomic regions associated with increased risk of neuropsychiatric disorders (Grove et al. 2019; Mullins et al. 2021; Trubetskoy et al. 2022), but prioritizing candidate causal genes that are regulated by disease-associated variants remains a challenge. One common bioinformatic approach is to first identify variants associated with changes in different molecular phenotypes such as gene-level expression and local splicing patterns (Katz et al. 2010; Li et al. 2018b; Shen et al. 2014; Vaquero-Garcia et al. 2016) and then statistically evaluate their relationship to known GWAS signals (Giambartolomei et al. 2014; Gusev et al. 2016; Zhu et al. 2016). For example, previous studies have identified expression or splicing quantitative trait loci (i.e. eQTL or sQTL) and compared them with GWAS signals to identify putative disease genes.

Even though isoforms are the fundamental biological units expressed in cells, to date, no large-scale isoform-level genetic analyses have been undertaken. This critical gap likely reflects the difficulty in accurately quantifying isoform-level expression from short-read RNA-seq data. The state-of-the-art isoform quantification methods (Bray et al. 2016; Li and Dewey 2011) address this issue by maximizing the log-likelihood function of the underlying statistical model with the expectation-maximization (EM) algorithm, thereby probabilistically assigning multi-mapped RNA-seq reads. Unfortunately, this approach holds valid only when all isoforms of a given transcriptome are known (Sterne-Weiler et al. 2018). For example, when the transcriptome annotation is incomplete, multi-mapped RNA-seq reads that belong to an "unknown isoform" might get erroneously assigned to different isoforms that are present in the input transcriptome annotation, yielding biased estimates. However, this also means that with more complete transcriptome annotations, the accuracy of isoform-level expression and subsequently gene-level expression will likely increase even with short-read RNA-seq data.

Despite this important limitation in accurately quantifying isoform-level abundances, to begin to understand how isoform expression changes in the human brain could mediate psychiatric GWAS signals, we leverage the PsychENCODE dataset (Gandal et al. 2018b; Wang et al. 2018) composed of 855 unrelated European individuals to assess the genetic architecture of brain gene and isoform expression. Although genetic influences on gene expression have been investigated in peripheral blood samples (Lloyd-Jones et al. 2017; Ouwens et al. 2020; Wheeler et al. 2016; Wright et al. 2014), similar estimates are missing for isoform expression, let alone in brain samples. Additionally, while several studies have jointly estimated shared genetic influences pairs of gene expression (Liu et al. 2017; Lukowski et al. 2017; Price et al. 2011), similar bivariate or

multivariate analyses are missing for isoform expression. Therefore, we quantify the degree of polygenicity and pleiotropy for 24,905 gene and 93,293 isoform expression in the human brain. We model isoform expression jointly for the first time by using multivariate variance components linear mixed models. This contrasts with pairwise bivariate models that are commonly implemented in human genetics settings (Bulik-Sullivan et al. 2015; Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013; Lee et al. 2012; Yang et al. 2011). We find genetic variants local to the gene or isoform in question (herein referred to as *cis*-SNPs) have a large effect on expression, whereas distal genetic variants (i.e. *trans*-SNPs) have individually small effects but collectively substantial effect on expression. By partitioning expression covariances among *cis*- and *trans*-SNPs, and residual effects, we find that there are substantial shared genetic influenced among *cis*-SNPs for isoform expression. Importantly, we find isoform-level analyses to lead to discovery of many more genetic signals that were not present at the gene-level. Some of these isoform-specific genetic signals were significantly associated with increased risk for brain-related disorders such as autism spectrum disorder (ASD). These findings indicate that isoform-resolution analyses have the potential to uncover novel disease genes and aid in interpreting GWAS results. Overall, we present a comprehensive dissection of genetic influences on brain gene and isoform expression and their relation to brain-related disorders.

3.3 Results

3.3.1 Overview of h^2_{SNP} and r_{g} analyses for brain gene and isoform expression

The genetic contribution to phenotypes of interest (i.e. SNP-based heritability or h^2_{SNP}) and the extent to which they are shared (i.e. genetic correlation or r_{g}) can be quantified through variance components linear mixed models. Under the variance components model, when there are no mean effects or there is only the intercept term, phenotypic (or expression) variance and covariance can be partitioned into the sum of variance components parameters (Methods). The simplest variance components model assumes two variance components, one of which captures aggregate genome-wide genetic effects. However, this model is likely misspecified for either gene or isoform expression, since SNPs in the vicinity of a gene (i.e. *cis*-SNPs) tend to exert stronger effects on its expression than distal SNPs (i.e. *trans*-SNPs). Similarly, the simplest multivariate variance components model looks at phenotypes pairwise and assumes two variance components, one of which captures the aggregate degree of pleiotropy among genome-wide SNPs. This model is likely inadequate for isoform-level expression, since the degree of genetic correlation can differ among different sets of SNPs (i.e. *cis*- and *trans*-SNPs).

Hence, in the present study, for all variance components models, whether be it univariate, pairwise bivariate, or multivariate, we specified three variance components that include *cis*- and *trans*-SNP effects, and residual effects. This model specification corresponds to an assumption that *cis*- and *trans*-SNP effects are realized from different distributions of effect sizes. Hereafter, we refer to the variance components parameters corresponding to either *cis*- or *trans*-SNP effects as genetic variances and genetic covariances. We defined *cis*-SNPs as those within ±1 Mb window of gene

start and gene end sites and *trans*-SNPs as all other SNPs (**Figure 3.1**). Indeed, other definitions are possible, and we tested varying windows as part of sensitivity analyses.

Unfortunately, fitting the most general form of multivariate variance components models with more than two phenotypes and more than two variance components still remains a significant computational challenge. To address this critical gap, we implemented the minorization-maximization (MM) algorithm for estimation in variance components linear mixed models (Zhou et al. 2019), using the Julia programming language (Bezanson et al. 2017). We refer to the Methods for a comprehensive summary of optimization methods for tackling variance components models, but we note here that the major advantages of the MM algorithm include numerical stability, fast convergence, and graceful adaptation to the positive semidefinite constraint of variance components parameters (Zhou et al. 2019).

We present estimates from the MM algorithm and restricted (or residual) maximum likelihood (REML) estimation, unless otherwise stated. The standard errors for variance components estimates were calculated from the Fisher information matrix (Methods). Inference on variance components parameters in the univariate case was done using a variation of the likelihood ratio test (LRT) (Molenberghs and Verbeke 2007; Yang et al. 2011). In the multivariate setting, inference on the off-diagonal elements of the variance components parameters was done using the Wald test (Methods).

We first estimated h^2_{SNP} and assessed its significance by fitting the univariate variance components model. Since we specified two separate variance components parameters for *cis*- and *trans*-SNPs,

27

h^2_{SNP} could be decomposed into h^2_{cis} and h^2_{trans} (i.e. $h^2_{\text{SNP}} = h^2_{\text{cis}} + h^2_{\text{trans}}$). To avoid estimating too many parameters in multivariate models, for each gene, we focused on isoforms with significant h^2_{SNP} at $P < 0.05$ in the univariate model. Since we partitioned expression covariances among *cis*- and *trans*-SNPs, and residual effects, there were two genetic correlation parameters (r_g's) that correspond to aggregate degree of pleiotropy among *cis*-SNPs ($r_{g,\text{cis}}$) and aggregate degree of pleiotropy among *trans*-SNPs ($r_{g,\text{trans}}$) (**Figure 3.1**). There was also residual correlation (r_e) which captures correlation due to biological and technical factors such as shared gene regulation and measurement error (Methods). It is important to note that r_g is not the sum of $r_{g,\text{cis}}$ and $r_{g,\text{trans}}$, since these parameters are on a normalized scale. Finally, we conducted *cis*-eQTL analyses in parallel with the same set of *cis*-SNPs (Delaneau et al. 2017) to validate h^2_{SNP} results as a positive control and to link isoform-specific signals with GWAS signals (**Figure 3.1**).
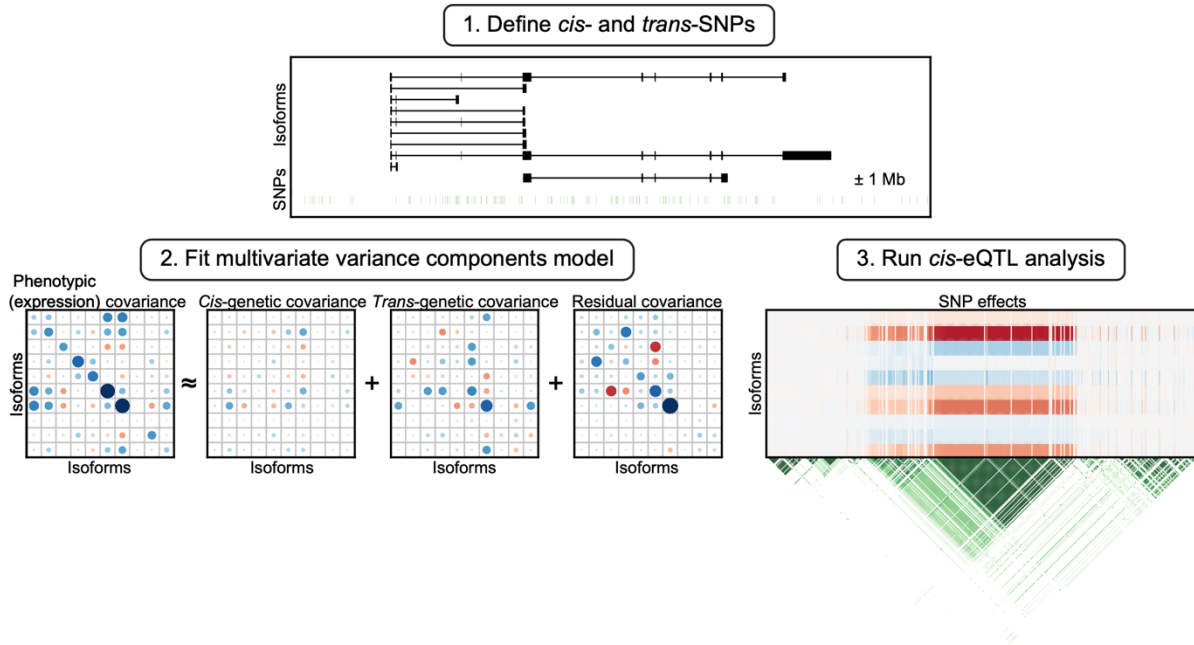
Figure 3.1: Overview of isoform-centric h^2_{SNP}, r_g, and *cis*-eQTL analyses. For a given gene and its constituent isoforms, *cis*-SNPs are defined as SNPs within ±1 Mb window of (collapsed) gene start and end sites, while *trans*-SNPs are defined as all other SNPs. Expression variances and covariances are partitioned among *cis*- and *trans*-SNPs for each gene and its constituent isoforms by fitting univariate, pairwise bivariate, and multivariate variance components linear mixed models (Zhou et al. 2019). The same sets of *cis*-SNPs are used for *cis*-eQTL analyses (Delaneau et al. 2017), results of which are compared to heritability and genetic correlation analyses. This schematic figure is based on the real data for ten isoforms belonging to the *KLHL24* gene.

3.3.2 Polygenicity of brain gene and isoform expression

We started with 24,905 genes and 93,293 isoforms in PsychENCODE. On average, each gene had four isoforms, but some genes had as many as 64 brain-expressed isoforms (**Supplementary Figure 3.1**). Of these, 22,965 genes and 89,926 isoforms had converged h^2_{SNP} estimates in the univariate model. Median h^2_{SNP} estimates were 0.31 and 0.35 for gene and isoform expression, respectively (**Figure 3.2a**), and median h^2_{cis} estimates were 0.01 and 0.01 for gene and isoform expression, respectively. 2,822 genes had significant h^2_{SNP} at Bonferroni-adjusted P value < 0.05, while 3,557 genes had at least one isoform which h^2_{SNP} was significantly different from zero at Bonferroni-adjusted P value < 0.05. With more lenient threshold at $P < 0.05$, 7,239 and 10,139 genes were heritable at the gene- and isoform-level, respectively. This increase in the number of heritable genes reflects added granularity with isoform-level analyses. It is standard practice when fitting linear mixed models to compare REML and maximum likelihood (ML) estimates, since REML estimates tend to be less biased, while ML estimates can have lower mean squared error (MSE). For the univariate variance components model, we found strong concordance between REML and ML estimates (**Supplementary Figure 3.2**). We note that h^2_{trans} and h^2_{SNP} estimates exhibit larger standard errors than h^2_{cis} estimates (**Supplementary Figure 3.2**), which is in line with expectation (Visscher et al. 2014), so we caution the readers in interpreting these h^2_{trans} and h^2_{SNP} estimates.

To make sure that our estimates were robust, we compared gene- and isoform-level h^2_{SNP} estimates for genes with a single isoform (n = 7,246 genes). We found that they were highly concordant (**Supplementary Figure 3.3**) despite gene- and isoform-level expression data being processed separately. When we fit a univariate variance components model with two variance components

that only specifies *cis*-genetic effects, we observed concordant estimates for h^2_{cis} (data not shown), although the estimates for the two variance components model were slightly inflated over the three variance components model, indicating the importance of correct model specification. Finally, we ran *cis*-eQTL analyses for gene- and isoform-level expression using QTLtools and found that most (if not all) heritable genes and isoforms harbor a significant eQTL (**Supplementary Figure 3.4**).

Using QTLtools, we identified 8,981 genes and 17,174 isoforms that harbor *cis*-eQTL at FDR < 0.05. The top associated SNP or index eQTL explained about 70% of variance from h^2_{cis} (**Supplementary Figure 3.5**). Conditional analyses using QTLtools found that most genes and isoforms have a single *cis*-eQTL signal, suggesting sparse and large genetic effects among *cis*-SNPs. We note that the number of independent *cis*-eQTL will likely increase with an increase in sample size, but in the present study, only 25% of genes and 15% of isoforms harbored more than one significant eQTL.

Next, we sought to investigate the polygenicity of *trans*-SNPs by partitioning genetic variances among 22 autosomal chromosomes. Since these estimates were noisy due to the limited sample size of PsychENCODE, we fit a penalized model with lasso penalty (Kim et al. 2021; Methods), which had an effect of shrinking variance components estimates and hence prioritizing the most important SNP effects. We found that the chromosome a given gene or isoform belongs to almost always had non-zero estimates, indicating substantial *cis*-SNP effects, while the rest of the chromosomes had non-zero estimates in proportion to their number of SNPs, suggesting polygenic effects. Altogether, *cis*-SNP effects were sparse but large, and *trans*-SNP effects were polygenic.

3.3.3 Pleiotropy among brain isoform expression

Given their genomic proximity and co-regulation, we hypothesized that isoforms are under shared genetic influences. To address this question, for genes with multiple heritable isoforms, we jointly modeled isoform expression using the multivariate variance components models, which allowed us to estimate h^2_{SNP} and r_g together. Of note, 3,801 genes had at least two isoforms heritable, and 1,743 genes had at least three isoforms heritable at $P < 0.05$. This led to modeling from two to as many as 23 isoforms jointly. We observed concordant variance components (**Figure 3.2b**) and h^2_{SNP} estimates (**Figure 3.2c**) between the univariate and multivariate models, although we note that h^2_{trans} and h^2_{SNP} estimates were slightly deflated for multivariate models compared to univariate models. The distribution of $r_{g,cis}$ was bimodal with the two extremes (**Figure 3.2d**) and a median estimate of 0.31, suggesting that *cis*-SNPs tend to affect or co-regulate nearby isoforms together. The negative $r_{g,cis}$ estimates close to -1 are characteristic of isoform switching events. Meanwhile, $r_{g,trans}$ and r_e estimates followed unimodal distributions with their medians shifted right from zero (**Figure 3.2d**). This makes biological sense in that the distal genetic regulators and other biological factors tend to influence transcription of nearby isoforms together.

Next, we sought to understand if fitting pairwise bivariate or multivariate models lead to any meaningful differences. This question is relevant for not only molecular readouts such as isoform expression, but also for complex traits and diseases in general, because most of the genetic correlation estimates are based on fitting pairwise bivariate models (Bulik-Sullivan et al. 2015; Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013; Lee et al. 2012; Yang et al. 2011). As for isoform expression, covariance terms in variance components estimates were concordant (**Figure 3.2e**), while r_g estimates were slightly inflated for pairwise bivariate models

31

compared to multivariate models (**Figure 3.2f**). Further, we observed numerous cases where the signs of r_g estimates flipped between these two models. Based on these findings and previous observations that pairwise bivariate models tend to yield estimates that do not respect the positive semidefinite constraint of variance components parameters (de Vlaming et al. 2021), we construe that multivariate estimates are generally more accurate than pairwise bivariate estimates.

Previous work demonstrated that genetic correlation mimics phenotypic correlation in the direction of effect (Bitner-Mathé and Klaczko 1999; Cheverud 1988; van Rheenen et al. 2019; Searle 1961; Sodini et al. 2018; Vattikuti et al. 2012; Waitt and Levin 1998), and thus motivating the use of phenotypic correlation as a proxy for genetic correlation. We sought to test this hypothesis for isoform expression and we observed that genetic correlation (both $r_{g,cis}$ and $r_{g,trans}$) generally recapitulates phenotypic correlation (or co-expression) in the direction of effect with $r_{g,cis}$ being overall larger in magnitude than phenotypic correlation (**Figure 3.2g**). This is consistent with previous studies and our own observation that the isoforms of a given gene are under substantial shared genetic influences, particularly among *cis*-SNPs (**Figure 3.2d**).
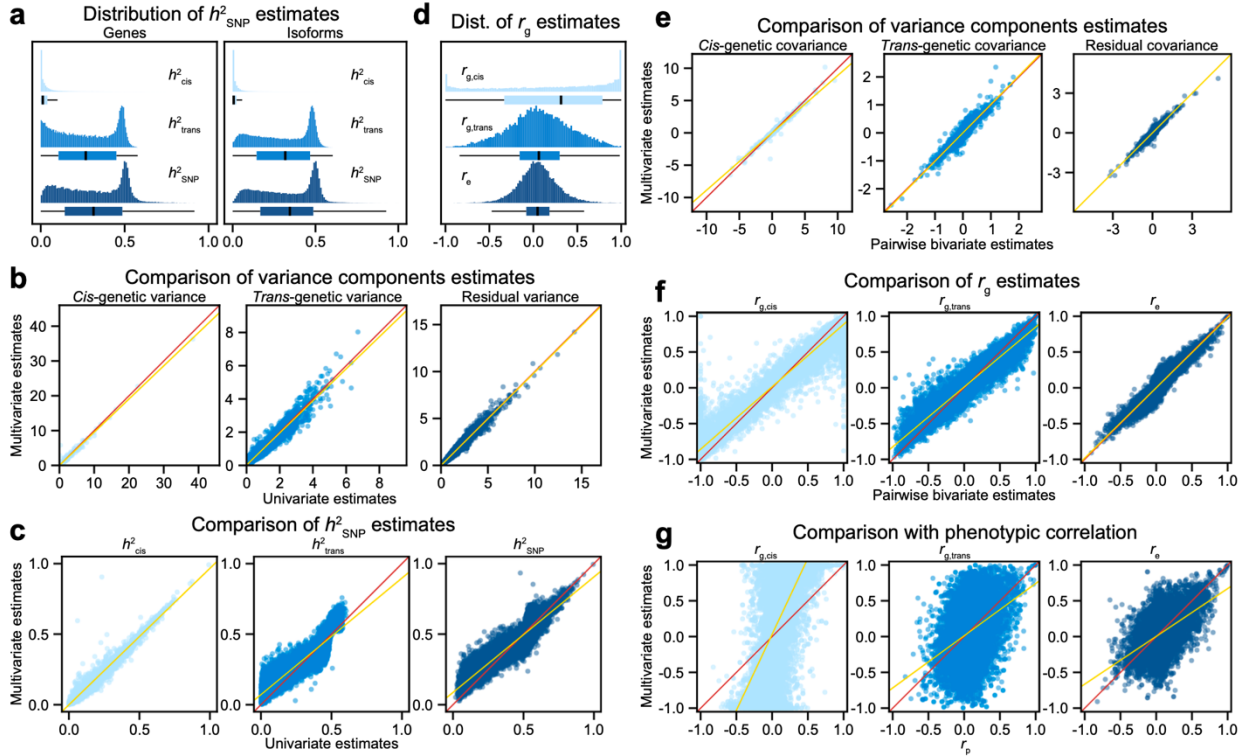
Figure 3.2: Comparison of h^2_{SNP} and r_g estimates from different variance components models. a, Shown are the distributions of h^2_{SNP} estimates for gene and isoform expression from fitting a univariate variance components model (22,965 genes + 89,926 isoforms). Note that $h^2_{\text{SNP}} = h^2_{\text{cis}} + h^2_{\text{trans}}$. **b**, Shown are variance components estimates from fitting univariate and multivariate variance components models. **c**, Shown are h^2_{SNP} estimates from univariate and multivariate variance components models. **d**, Shown is the distribution of r_g estimates from fitting a multivariate variance components model. Note that $r_{g,\text{cis}}$ captures aggregate degree of pleiotropy among *cis*-SNPs, while $r_{g,\text{trans}}$ captures aggregate degree of pleiotropy among *trans*-SNPs. r_e captures residual correlation due to (unknown) biological and technological factors. The bimodal distribution of $r_{g,\text{cis}}$ with two modes in the extremes suggests that there are substantial shared genetic influences among *cis*-SNPs. **e**, Shown are (co)variance components estimates from pairwise bivariate and multivariate variance components models. **f**, Shown are r_g estimates from pairwise bivariate and multivariate variance components models. **g**, Shown are r_g estimates from a multivariate variance components model and expression (phenotypic) correlation (r_p). The same r_p values are plotted on the x-axis. All estimates are from REML estimation. The red lines are the diagonal lines, while the yellow lines denote the lines from linear regression.

3.3.4 *ATP9B* gene as a case study for h^2_{SNP} and r_g analyses

We now focus on *ATP9B* gene to make our h^2_{SNP} and r_g results more concrete (**Figure 3.3**). In Gencode v19, *ATP9B* had 27 annotated isoforms, 17 of which were determined to be brain-expressed in PsychENCODE based on the criteria TPM > 0.1 in at least 25% of samples (Gandal et al. 2018b). Nine of these isoforms were found heritable at $P < 0.05$ after fitting the univariate variance components model (**Figure 3.3a**). Pairwise bivariate and multivariate models were fit for

the nine heritable isoforms, which resulted in concordant h^2_{SNP} estimates with similar magnitudes (**Figure 3.3b**). Standard errors for h^2_{SNP} estimates were comparable across univariate, pairwise bivariate, and multivariate models, while standard errors for h^2_{trans} were consistently larger than those of h^2_{cis}. Because we modeled nine isoforms, we note that there were eight pairwise bivariate estimates for each h^2_{SNP} parameter (**Figure 3.3b**). For r_{g} parameter, pairwise bivariate and multivariate models yielded estimates that were concordant in direction of effect, but less so in magnitude. That is, pairwise bivariate estimates were relatively inflated compared to multivariate estimates for both $r_{\text{g,cis}}$ and $r_{\text{g,trans}}$ (**Figure 3.3c**). Meanwhile, for both pairwise bivariate and multivariate models, $r_{\text{g,cis}}$ estimates were substantially larger than $r_{\text{g,trans}}$ estimates, suggesting that there were shared genetic influences, particularly among *cis*-SNPs. Not to mention, $r_{\text{g,cis}}$ and $r_{\text{g,trans}}$ estimates were generally larger than r_{e} estimates. Further comparison of r_{g} and r_{e} estimates with r_{p} estimates revealed that phenotypic correlation resembled residual correlation most closely (**Figure 3.3d**), indicating that for this particular gene *ATP9B*, residual (e.g. biological and technological) factors overall exert larger effects than genetic factors (i.e. *cis*- and *trans*-SNPs). We note that one might be confused by this observation of high $r_{\text{g,cis}}$ estimates but little resemblance between $r_{\text{g,cis}}$ and r_{p} estimates. This is possible because r_{g} and r_{e} estimates are on a standardized scale, and hence the magnitude of residual variance-covariance can be larger than genetic variance-covariance while r_{e} is smaller than r_{g} (van Rheenen et al. 2019). Finally, upon visual inspection of eQTL results, we found that there were distinct isoform-level eQTL signals (**Figure 3.3e**). For *ATP9B*, gene-level expression was also heritable and harbored strong eQTL signals (**Figure 3.3e**) unlike in subsequent examples (**Figure 3.4**).
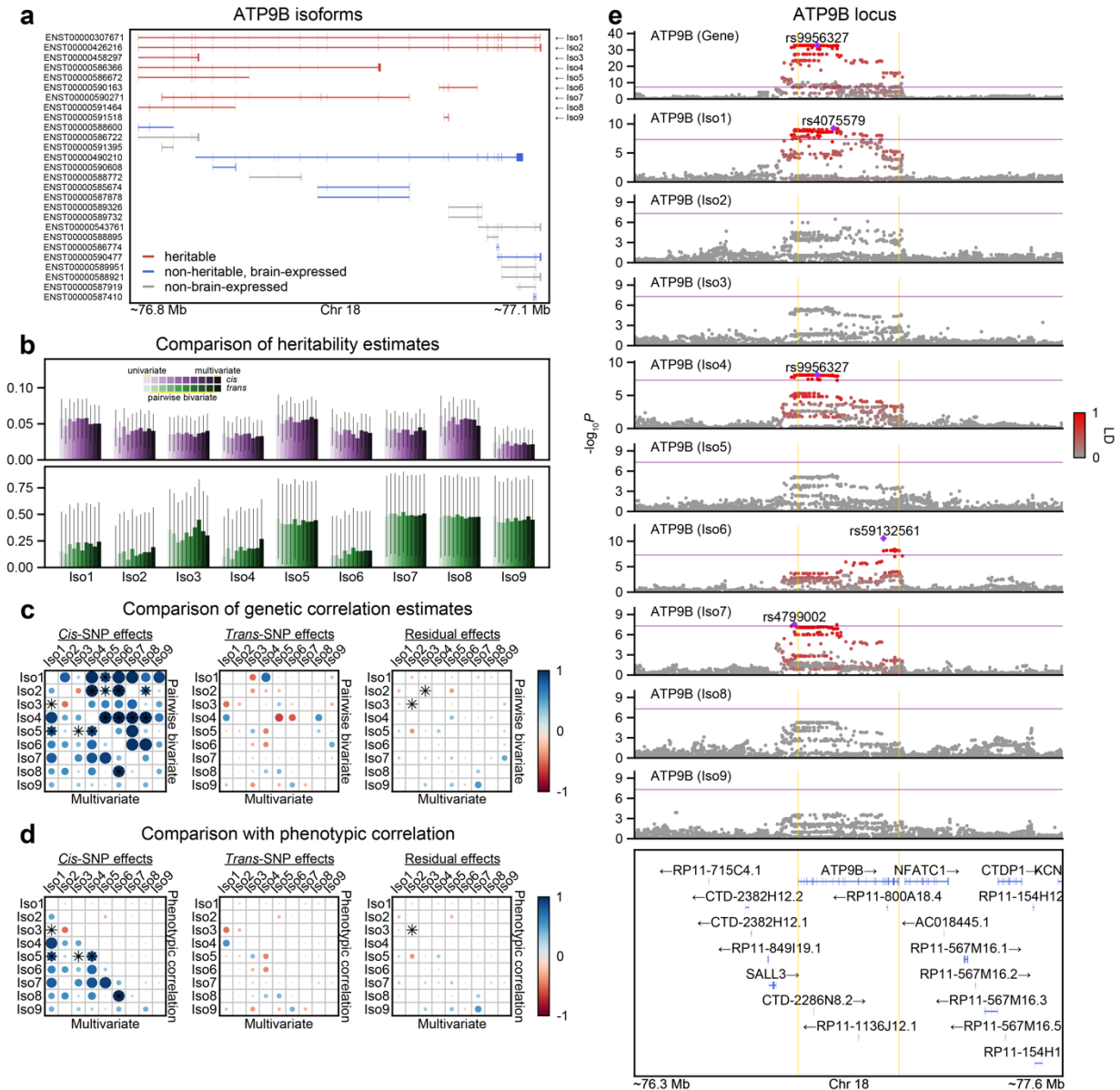
Figure 3.3: *ATP9B* as an example gene in h^2_{SNP}, r_g, and *cis*-eQTL analyses. a, Shown are *ATP9B* isoforms that are present in Gencode v19. Nine isoforms that are found heritable in a univariate variance components model with a variation of the likelihood ratio test (LRT) (Molenberghs and Verbeke 2007) are highlighted in red, and eight isoforms that are determined to be brain-expressed (Gandal et al. 2018b) but not heritable are highlighted in blue. **b,** h^2_{cis} and h^2_{trans} estimates are shown in top and bottom rows, respectively. Colors encode estimates from univariate, pairwise bivariate, and multivariate variance components models. Because there are nine isoforms that are modeled jointly, there are eight pairwise bivariate estimates. All error bars denote ± one standard errors calculated from the Fisher information matrix. **c,** Shown are $r_{g,cis}$, $r_{g,trans}$, r_e estimates from pairwise bivariate and multivariate variance components models for *ATP9B*. The lower triangular elements represent multivariate estimates, while the upper triangular elements represent pairwise bivariate estimates. Asterisks denote significance from the Wald test at $P < 0.05$. **d,** Alike panel **c**, except that the upper triangular elements now represent phenotypic correlations (r_p). The same r_p values are plotted. **e,** LocusZoom plot for *ATP9B* gene and isoform expression. Top row shows gene-level results. Index SNPs for features passing $P = 5 \times 10^{-8}$ threshold are shown and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes significance of $P = 5 \times 10^{-8}$, and yellow lines denote gene start and end sites for *ATP9B* gene. LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel.

35

3.3.5 Isoform-level eQTL signals prioritize candidate causal genes in GWAS loci

We next sought to understand whether isoform-resolution analyses could help prioritize disease genes in GWAS loci. Here, as a proof of concept, we nominate and share four such cases. The first is *XRN2* gene, which resides in one of two GWAS loci for autism spectrum disorder (ASD) (Grove et al. 2019). Of note, the other ASD GWAS locus is a pleiotropic genomic region characterized by long-range, complex LD (Anderson et al. 2010; **Supplementary Figure 3.7**), which complicates fine-mapping and gene prioritization efforts. For *XRN2*, there were three brain-expressed isoforms, one of which was significantly heritable (ENST00000430571; $h^2_{cis} = 0.07$, $h^2_{trans} = 0.46$, $P = 5.3 \times 10^{-13}$). eQTL signal of this isoform colocalized strongly with the ASD GWAS signal (**Figure 3.4a**; **Supplementary Figures 3.8-3.9**), suggesting that this isoform might be the causal isoform underlying the ASD GWAS signal. The second example is *SYNE1* gene, a SFARI ASD risk gene, of which five isoforms were heritable (**Supplementary Figure 3.10**). Three of these isoforms harbored strong and distinct eQTL signals, one of which colocalized with a well-known bipolar disorder (BD) GWAS signal (Mullins et al. 2021; **Figure 3.4b**; **Supplementary Figures 3.11-3.12**). The third example is *TBL1XR1* gene, another SFARI gene, for which loss-of-function is implicated in increased risk for ASD and developmental delay disorder (DDD) (Kaplanis et al. 2020; Satterstrom et al. 2020). Two *TBL1XR1* isoforms were found heritable, and they harbored distinct isoform-level eQTL signals, one of which colocalized with a SCZ GWAS signal (**Figure 3.4c**; **Supplementary Figures 3.13-3.14**). The last example is *SYT1* gene, which is another SFARI gene and a high-confidence DDD risk gene (Kaplanis et al. 2020). Three *SYT1* isoforms were found heritable (**Supplementary Figure 3.15**), two of which colocalized with an educational attainment (EA) GWAS signal (Okbay et al. 2022; **Figure 3.4d**; **Supplementary Figures 3.16-3.17**). Given the strong implication of these example genes in neurodevelopmental disorders and
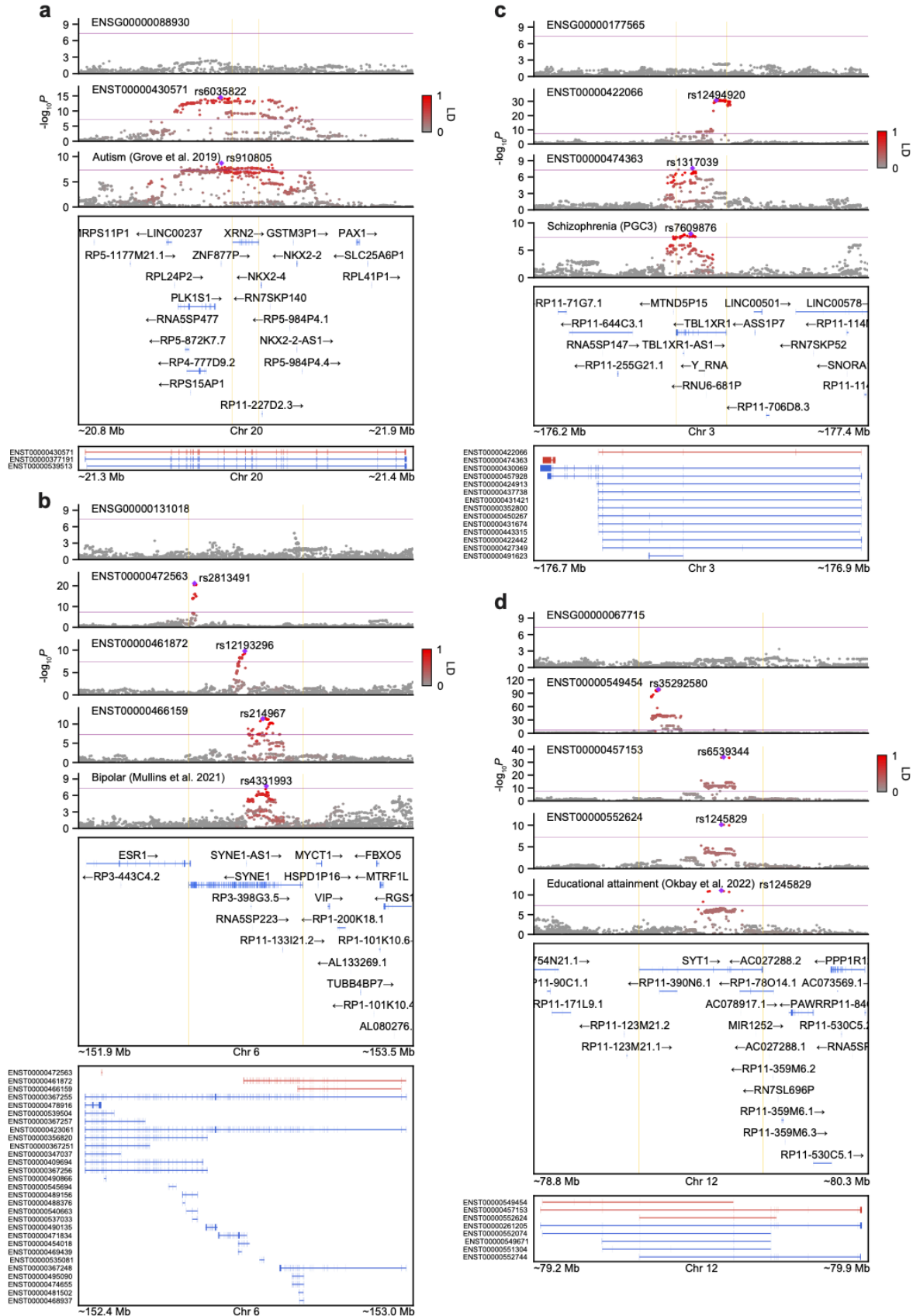
Figure 3.4: Isoform-level eQTL signals prioritize candidate causal genes in established GWAS loci. Shown are LocusZoom plots for **a**, *ATP9B*, **b**, *SYNE1*, **c**, *TBL1XR1*, and **d**, *SYT1* genes along with several complex phenotypes. Select isoforms that are included in LocusZoom plots are highlighted in red, while all other isoforms that are brain-expressed are highlighted in blue. Due to space constraints, isoforms that are present in Gencode v19 but not found brain-expressed are omitted. Index SNPs for features passing $P = 5 \times 10^{-8}$ threshold are shown and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes significance of $P = 5 \times 10^{-8}$, and yellow lines denote gene start and end sites. LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel.

their proximity to corresponding GWAS signals, these genes are likely the true causal genes driving the GWAS signals. Relatively straightforward patterns of linkage disequilibrium (LD) and reasonable gene densities also help in reaching this conclusion (**Figure 3.4**). It is important to note that all these example genes were missing gene-level eQTL signals, while harboring isoform-level eQTL signals, which highlights the potential of isoform-level analyses. Due to incomplete isoform annotations, the isoforms we identify may not be the true causal isoforms driving the GWAS signals, but with much more refined and complete transcriptome annotations along with more accurate quantifications of isoform-level expression (e.g. from long-read RNA sequencing), we hypothesize that we would gain more fine-grained resolution in both gene- and isoform-level expression and thereby uncover more disease genes in GWAS loci.

3.3.6 Replication of *XRN2* isoform-level eQTL signals in the developing human brain with improved isoform annotations

As another proof of concept, we next sought to definitely fine-map the *XRN2* locus for ASD with improved isoform annotations. This was partly motivated by its relatively simple gene structure, for example, compared to that of *SYNE1* (**Figures 3.4b**). In Gencode v19, there were three isoforms for this gene, all of which were determined to be brain-expressed. Upon close inspection of *XRN2* exons, we observed that the only difference between the ASD-associated isoform ENST00000430571 and the canonical isoform ENST00000377191 was skipped exon 2. On the

contrary, the difference between the remaining isoform ENST00000539513 and the canonical isoform was a different exon 1 and the rest of the sequence was the same. Interestingly, we then noticed that all isoforms except the canonical isoform had been removed in subsequent Gencode versions past v19 due to low transcript support level. To ensure that our isoform is expressed and to improve the transcriptome annotation of the human brain, we combined the latest Gencode v40 annotation with existing long-read data from six studies (Leung et al. 2021; Palmer et al. 2021; Glinos et al. 2021; Methods). The data for these studies were mostly generated from fetal and adult human brain samples. To compile a list of high-confidence isoforms, we filtered for isoforms that were found in at least two or three different sources. Remarkably, we recovered all three *XRN2* isoforms in Gencode v19 with ENST00000377191, ENST00000430571, and ENST00000539513 found in six, four, and two studies, respectively (**Figure 3.5a**). Given that ASD genetic risk factors are known to converge in cell-types and biological pathways during brain development (Parikshak et al. 2013; Walker et al. 2019; Willsey et al. 2013), we then sought to replicate our adult brain isoform-level eQTL findings in fetal brain samples (O'Brien et al. 2018). The previous work (O'Brien et al. 2018) performed isoform-level eQTL analyses with Gencode v23 and did not detect any significant association for *XRN2*, presumably because ENST00000377191 is the only isoform present in that version of Gencode. In contrast, when we re-quantified fetal brain RNA-seq data, using Salmon (Patro et al. 2017) with the updated transcriptome annotation, and repeated isoform-level *cis*-eQTL analyses, we observed a strong and specific eQTL signal for ENST00000430571 (or TCONS_00530677 equivalently) (**Figure 3.5b**). Further, the index SNP for ASD GWAS result (rs910805) and its risk increasing (minor) allele G was associated with reduced expression of ENST00000430571 (**Figure 3.5c**), which direction of effect is consistent with the adult brain findings (**Figure 3.4a**). As another sanity check, we tested all genes within ±0.5 Mb window of

39
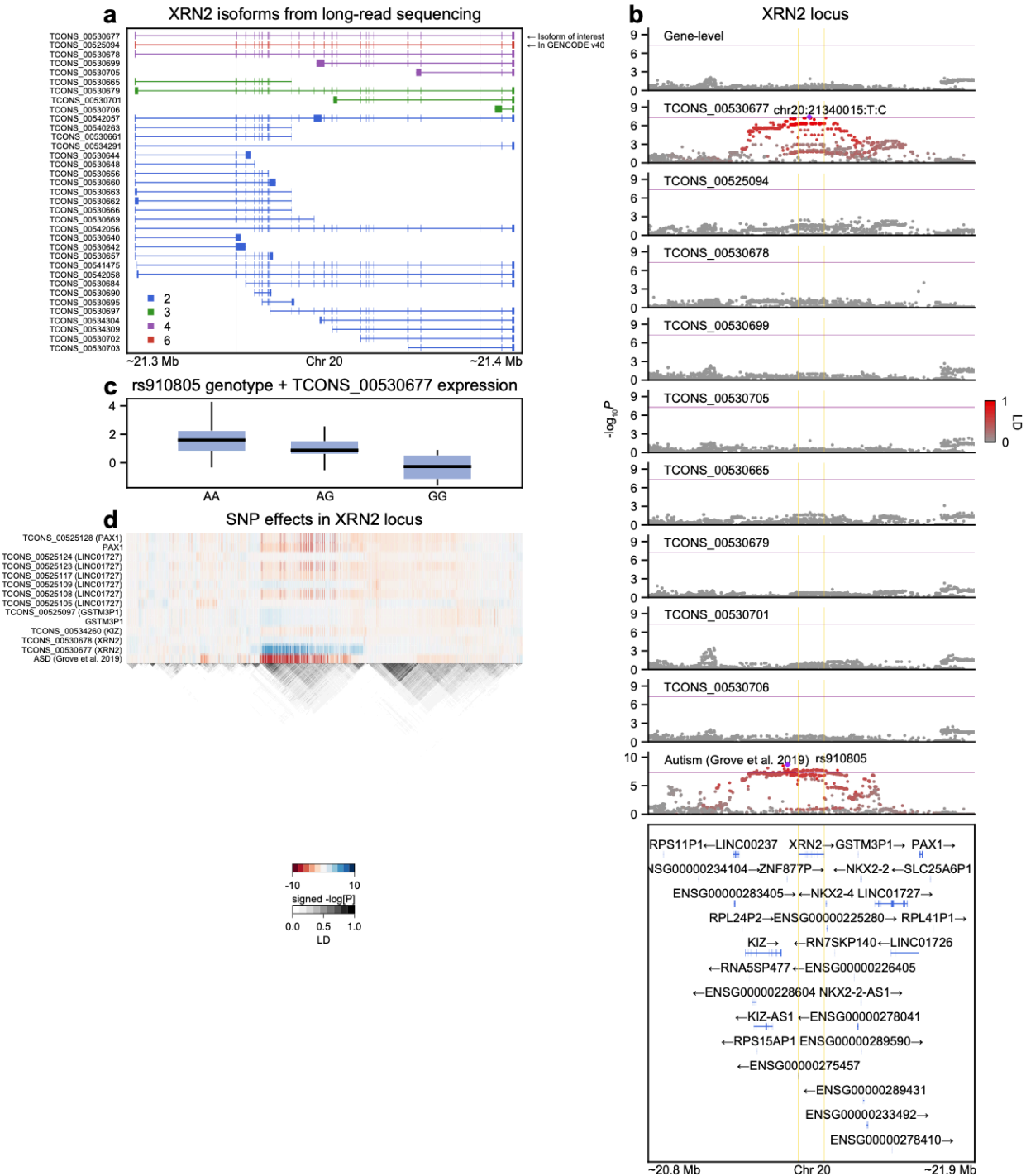
Figure 3.5: Fine-mapping of top ASD GWAS locus with isoform-level eQTL signals in an independent fetal human brain dataset. a, Shown are high-confidence, credible *XRN2* isoforms compiled from six existing long-read studies (Leung et al. 2021; Palmer et al. 2021; Glinos et al. 2021). Isoforms are colored with respect to the number of studies each isoform is found in. TCONS_00530677, TCONS_00525094, and TCONS_00541475 correspond to ENST00000430571, ENST00000377191, and ENST00000539513 in Gencode v19, respectively. The skipped exon 2 in TCONS_00530677 is shaded in grey. We subsequently filtered for isoforms found in at least three different studies and updated the transcriptome accordingly. **b**, Shown is LocusZoom plot for *XRN2* gene and its isoform expression based on the updated transcriptome annotation. Isoform expression was quantified in fetal human brain samples (O'Brien et al. 2018). **c**, Index SNP for ASD GWAS (rs910805) and its risk increasing allele G was associated with

40

reduced expression of TCONS_00530677. Note that G is also the minor allele. **d**, Shown is a heatmap of eQTL signals for both gene- and isoform-level expression for features within ±0.5 Mb window of (collapsed) gene start and end sites for *XRN2* gene. Although eQTL analyses were conducted for a total of 30 genes within this region and all their isoforms, we only plot the results of features with minimum *P* values less than 10^{-4} for visual clarity. For isoforms, their cognate genes are shown in parentheses. LD is calculated with 86 individuals of European ancestry in O'Brien et al. 2018, which are the same individuals used for eQTL analyses. These analyses were conducted using the GRCh38 human genome build, which leads to slightly different genomic coordinates relative to previous figures that were based on the hg19 reference genome.

XRN2 and their constituent isoforms from the updated transcriptome annotation and observed that no other gene- or isoform-level expression harbored eQTL signals that colocalized with the ASD GWAS signal (**Figure 3.5d**). Altogether, we replicated our adult brain *XRN2* findings with more complete isoform annotations in an independent fetal brain dataset and confidently narrowed down expression changes in ENST00000430571 as the causal signal for the ASD GWAS signal.

The *XRN2* isoform ENST00000430571 has exon 2 skipped, which in theory should be uncovered by annotation-free methods that detect local splicing patterns (Li et al. 2018b). Indeed, the intron that corresponds to the end of exon 1 and the start of exon 3 were shown to harbor a significant splicing QTL (sQTL) in previous studies (Aygün et al. 2021; Li et al. 2019; Walker et al. 2019). However, several other studies failed to detect a significant sQTL (Raj et al. 2018; GTEx Consortium 2020). We suspect that this discrepancy is due to a combination of low level of expression for ENST00000430571 and differences in RNA-seq library preparation across studies with polyA selection methods being more susceptible to 3' bias than rRNA depletion methods. Indeed, changes in splicing patterns between rs910805 genotypes were subtle (**Supplementary Figure 3.18**), which could be overshadowed by polyA selection methods that induce 3' bias.

3.4 Discussion

In this study, we leveraged large-scale genetic and transcriptomic datasets from PsychENCODE (Gandal et al. 2018b; Wang et al. 2018) to dissect the genetic influences (i.e. h^2_{SNP} and r_g) on human brain gene and isoform expression. We find a substantial proportion of the human brain transcriptome to be under genetic control. We find *cis*-SNP effects to be sparse and large individually, while *trans*-SNP effects to be polygenic and large in aggregate. We jointly model isoform expression— for the first time, to the best of our knowledge—with multivariate variance components linear mixed models by implementing the MM algorithm (Zhou et al. 2019). We conclude that the isoforms of a given gene are under shared genetic influences, particularly among *cis*-SNPs. By comparing pairwise bivariate to multivariate models, we empirically show that multivariate models yield more accurate estimates. Finally, several genes are found heritable only at the isoform-level and their genetic signals colocalize strongly with GWAS signals, suggesting that isoform expression changes might be the underlying drivers of these GWAS signals. One notable example is *XRN2*, a specific isoform of which we prioritize for increasing ASD risk with reduced expression. Overall, the present study thoroughly partakes in an intellectual endeavor of estimating genetic parameters for brain gene and isoform expression and highlights the utility of conducting isoform-resolution analyses.

We first note a major and important caveat of isoform-level analyses: the underlying generative model for the state-of-the-art short-read RNA-seq quantification tools (Bray et al. 2016; Li and Dewey 2011; Sterne-Weiler et al. 2018) start with an assumption that all expressed isoforms of a given transcriptome are known. But the human transcriptome annotation is far from complete, especially for the brain, which can result in inaccurate isoform quantifications and hence inaccurate downstream analyses. However, this also means that with more complete transcriptome

annotations, we can obtain more accurate estimates of isoform-level expression even with short-read RNA-seq data (Sterne-Weiler et al. 2018). Obviously, it is impossible to account for all possible (known and unknown) sample- and isoform-specific biases when resolving multi-mapped short-read RNA-seq reads and estimating isoform-level expression, but the current methods and their statistical models (Bray et al. 2016; Li and Dewey 2011) offer a reasonable starting point. With extensive efforts to comprehensively catalog all brain-expressed isoforms (Clark et al. 2020; Flaherty et al. 2019; Leung et al. 2021; Palmer et al. 2021; Glinos et al. 2021), we foresee that the accuracy of isoform-level quantification will improve for short-read RNA-seq data.

The current study is limited by its sample size. This is evident by relatively large standard errors in our estimates, particularly for *trans*-SNP effects (Visscher et al. 2014). Hence with larger sample sizes, we expect to uncover more genes and isoforms that are genetically regulated. We also note that the PsychENCODE dataset is an output of mega-analysis of six different studies, which can lead to decreased signal-to-noise ratio from heterogeneity and various technical artifacts. We envision increased genetic discovery with more homogenous genotype and RNA-seq data that have sufficient sample sizes. Moreover, the choice of RNA-seq normalization methods can have an impact on h^2_{SNP} and r_g estimates. The PsychENCODE expression data was processed in $\log_2$-CPM-TMM scale, which better accounts for differences in library composition between samples in a large mega-analysis dataset such as PsychENCODE, but complicates the relationship between gene and isoform-level expression, including h^2_{SNP} estimates (Methods). In the main text, we only highlight a handful of genes which isoform expression changes seem to drive GWAS signals such as *XRN2*, *SYNE1*, *TBL1XR1*, and *SYT1*. Whether isoform-level genetic signals are more broadly enriched for disease risk and how many additional putative disease genes we can prioritize from

isoform-resolution analyses remains to be further investigated (Giambartolomei et al. 2014; Gusev et al. 2016; Zhu et al. 2016).

Several important questions remain regarding the genetic architecture of brain gene and isoform expression. First, other sources of genetic variance-covariance (e.g. epistasis and dominance) have not been explored, as we focus on additive effects from *cis-* and *trans-*SNPs. Although we look at the degree of pleiotropy among constituent isoforms of a given gene, nearby genes can be influenced by the same genetic factors as well (Ribeiro et al. 2021; Wainberg et al. 2019). The relationship between effect size and minor allele frequency is not similarly addressed in the present study. Finally, sources of variance-covariance in gene and isoform expression are multifactorial with many non-genetic factors playing a role. It was not feasible to quantify the effects of such factors in this study due to the limited availability of relevant metadata.

The *XRN2* gene encodes an RNA-binding protein that possesses 5'-3' exoribonuclease activity. It is known to promote termination of transcription by degrading RNA to resolve R-loops. *XRN2* is mildly constrained (pLI = 0.35, LOEUF = 0.35), shows peak expression in the fetal brain (Li et al. 2018a), and is enriched in neuronal cell-types (Polioudakis et al. 2019; Wang et al. 2018). Interestingly, visual inspection of phenome-scale LocusZoom plots reveals that there are two distinct GWAS signals within the *XRN2* locus, one for ASD and the other for blood cell-related phenotypes (**Supplementary Figure 3.19**). Note that the same GWAS signal for ASD reached genome-wide significance in the latest GWAS for ADHD. We find that at the gene-level, *XRN2* is not heritable and does not harbor a significant eQTL, but at the isoform-level, one specific *XRN2* isoform is heritable and harbors a significant eQTL that colocalizes with the ASD GWAS signal.

As a result of the skipping of the second exon, this isoform is missing 76 amino acids in the 5' end compared to the canonical *XRN2* isoform. Structural and biological consequences of such changes in protein sequences need to be studied in the future. We note that the prioritized and canonical isoforms belong in different brain WGCNA modules (isoM23 and isoM35; Gandal et al. 2018b), which capture known neurobiological pathways and cell-types, potentially suggesting that the two isoforms might be involved in different biological pathways in the human brain.

In sum, we advocate for running isoform-level analyses in conjunction with traditional gene-level analyses, since regulation of gene expression is far more complex than previously thought and isoform-level analyses can provide additional granularity. In comparison to usual sQTL analyses (Aygün et al. 2021; Garrido-Martín et al. 2021; GTEx Consortium 2020; Gusev et al. 2018; Li et al. 2018b; Li et al. 2019; Raj et al. 2018; Takata et al. 2017; Walker et al. 2019; Zhang et al. 2020a), isoform-level analyses are more interpretable in that they can pinpoint specific isoforms that are influenced by genetic variation. Furthermore, preliminary analyses in the fetal brain suggest that isoform-level eQTL signals recapitulate most, if not all, sQTL signals (data not shown), highlighting the informativeness of these analyses despite their limitations as discussed above.

3.5 Methods

The PsychENCODE genotype dataset

Genotype array and frontal cortex RNA-seq data from Freeze 1 and 2 of PsychENCODE were obtained from www.doi.org/10.7303/syn12080241. This consisted of uniformly processed data from six studies: BipSeq, LIBD_szControl, CMC-HBCC, CommonMind, BrainGVEX, and UCLA-ASD (see Table S1 and Fig. S33 in Wang et al. 2018). Genotype data for these individual studies were previously harmonized through phasing and imputation with the Haplotype Reference Consortium (HRC) reference panel. We focused on 860 unique European individuals with matching genotype and frontal cortex RNA-seq data. We started with 5,312,508 HRC imputed SNPs and filtered for SNPs with minor allele frequency (MAF) > 0.01, genotype and individual missingness rate < 0.05, and Hardy-Weinberg equilibrium P values > 1e-6. Five pairs of individuals had classic genetic relationship matrix (GRM) values > 0.05 when using all filtered SNPs, while 647 pairs of individuals had GRM values > 0.025. We kept one individual from each of five pairs and only SNPs belonging to autosomal chromosomes, resulting in a total of 855 unrelated European individuals and 4,685,674 SNPs for downstream analyses.

The PsychENCODE RNA-seq dataset

We used post-QC RNA-seq data that were fully processed, filtered, and normalized (see Materials/Methods and Fig. S3 in Gandal et al. 2018b). Of note, RNA-seq reads were previously aligned to the hg19 reference genome with STAR 2.4.2a and gene and isoform-level quantifications calculated using RSEM v1.2.29. Genes and isoforms were filtered to include those with TPM > 0.1 in at least 25% of samples. Gene and isoform expression were separately normalized using TMM normalization in edgeR and $\log_2$-transformed. The same set of known

biological and technical covariates were used for mean or fixed effects, which include age, age^2, study, sex, diagnosis, RNA integrity number (RIN), RIN^2, post-mortem interval (PMI), 24 sequencing principal components (PCs), and 5 genetic PCs. RNA-seq data was also restricted to frontal cortex samples from European individuals as well as genes and isoforms belonging to autosomal chromosomes, resulting in a total of 24,905 genes and 93,293 isoforms based on Gencode v19 annotations. The same expression data was used for all downstream analyses unless otherwise stated. We note that we did not correct for latent factors for the main analyses such as hidden covariates with prior (HCP) (Mostafavi et al. 2013), given that we are modeling *cis*- and *trans*-SNP effects simultaneously and we risk removing trans effects by adjusting for such latent factors (Dahl et al. 2019).

Multiple variance components linear mixed model for univariate response

Given an $n \times 1$ response vector $\mathbf{y}$ and $n \times p$ predictor matrix $\mathbf{X}$, the variance components linear mixed model assumes $\mathbf{y} \sim N(\mathbf{X\beta}, \mathbf{\Omega})$, where

$$\mathbf{\Omega} = \sum_{i=1}^{m} \sigma_i^2 \, \mathbf{V}_i,$$

and $\mathbf{V}_1, \dots, \mathbf{V}_m$ are m known positive semidefinite matrices. The parameters of the model include mean effects ($\mathbf{\beta}$) and variance components ($\sigma_1^2, \dots, \sigma_m^2$). $\mathbf{\Omega}$ is assumed to be positive definite. The simplest heritability and GWAS model assumes two variance components, where $\mathbf{V}_1$ is a kinship matrix and $\mathbf{V}_2 = \mathbf{I}$. This model is misspecified for gene expression, since SNPs in the vicinity of a gene (i.e. *cis*-SNPs) are known to exert stronger effects on its expression. Therefore, for gene and isoform expression, we specified three variance components, where

$$\mathbf{\Omega} = \sigma_1^2 \mathbf{V}_1 + \sigma_2^2 \mathbf{V}_2 + \sigma_e^2 \mathbf{I},$$

and $\mathbf{V}_1$ and $\mathbf{V}_2$ are empirical kinship matrices constructed from *cis-* and all other SNPs (i.e. *trans-*SNPs), respectively, such that σ_1^2 and σ_2^2 capture aggregate genetic effects of corresponding sets of SNPs. The classic genetic relationship matrix (GRM) was used for both empirical kinship matrices, under which $\frac{1}{n-1}\mathrm{E}[\mathbf{y}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{y}] \approx \frac{n}{n-1}(\sigma_1^2 + \sigma_2^2) + \sigma_e^2 + \frac{1}{n-1}\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}\boldsymbol{\beta}$. This equality is approximate, since the diagonal elements of GRM are not exactly one. When there are no mean effects or there is only the intercept term, the overall variance can be decomposed into the sum of variance components as above, so it makes natural sense to define SNP-based heritability as

$$h_{\text{SNP}}^2 := \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + \sigma_e^2}.$$

Multiple variance components linear mixed model for multivariate response

The multivariate response variance components model assumes an $n \times d$ response matrix $\mathbf{Y}$ with vec $\mathbf{Y} \sim N(\text{vec}(\mathbf{XB}), \boldsymbol{\Omega})$, where

$$\boldsymbol{\Omega} = \sum_{i=1}^{m} \boldsymbol{\Gamma}_i \otimes \mathbf{V}_i.$$

vec $\mathbf{Y}$ creates an $nd \times 1$ vector from $\mathbf{Y}$ by stacking its columns and $\otimes$ denotes the Kronecker product. The parameters of the model include $p \times d$ mean effects $\mathbf{B}$ and $d \times d$ variance components $(\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_m)$ that are positive semidefinite. $\boldsymbol{\Omega}$ is assumed to be positive definite. The univariate response model is subsumed under the multivariate response model when there is only a single phenotype. The simplest genetic correlation model looks at phenotypes pairwise and assumes two variance components. However, this model is likely inadequate for isoform-level

expression, so for constituent isoforms of a given gene, we modeled them jointly and specified three variance components such that

$$\mathbf{\Omega} = \mathbf{\Gamma}_1 \otimes \mathbf{V}_1 + \mathbf{\Gamma}_2 \otimes \mathbf{V}_2 + \mathbf{\Gamma}_e \otimes \mathbf{I}.$$

Corresponding to the univariate response model, $\mathbf{\Gamma}_1$ captures genetic variances and covariances among *cis*-SNPs, while $\mathbf{\Gamma}_2$ captures genetic variances and covariances among *trans*-SNPs. Let $(\mathbf{Y})_{\cdot j}$ denote the jth column of $\mathbf{Y}$ and $(\mathbf{Y})_{jk}$ the (j, k)th element of $\mathbf{Y}$. Then $\frac{1}{n-1}\mathrm{E}[(\mathbf{Y})_{\cdot j}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(\mathbf{Y})_{\cdot k}] \approx \frac{n}{n-1}[(\mathbf{\Gamma}_1)_{jk} + (\mathbf{\Gamma}_2)_{jk}] + (\mathbf{\Gamma}_e)_{jk} + \frac{1}{n-1}(\mathbf{B})_{\cdot j}^T \mathbf{X}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}(\mathbf{B})_{\cdot k}$. In other words, just as the overall variance could be decomposed into the sum of genetic variances and residual variance, the overall covariance can be decomposed into the sum of genetic covariances and residual covariance following this model. Genetic correlation is defined as genetic covariance divided by the product of square root of corresponding genetic variances. For example, genetic correlation for j, kth isoforms among *cis*-SNPs is $\frac{(\mathbf{\Gamma}_1)_{jk}}{\sqrt{(\mathbf{\Gamma}_1)_{jj}(\mathbf{\Gamma}_1)_{kk}}}$. It is worthwhile to note that the magnitude of genetic covariance or genetic correlation can differ between *cis*- and *trans*-SNPs.

Residual variance could be further decomposed into $\mathbf{\Gamma}_e = \mathbf{\Gamma}_\tau + \mathbf{\Gamma}_\epsilon$, where $\mathbf{\Gamma}_\tau$ captures covariance from sampling error and measurement error from shared transcript structures among isoforms, while $\mathbf{\Gamma}_\epsilon$ captures covariance arising from other biological factors. In practice, these two are indistinguishable from one another, but one could make use of bootstrap samples (i.e. also known as technical replicates) from Kallisto (Bray et al. 2016) or Salmon (Patro et al. 2017) to estimate $\mathbf{\Gamma}_\tau$. Then we can let $\hat{\mathbf{\Gamma}}_\epsilon = \hat{\mathbf{\Gamma}}_e - \hat{\mathbf{\Gamma}}_\tau$.

Relationship between gene and isoform expression h_{SNP}^2

Gene-level expression in unit of transcripts per million (TPM) is conventionally assumed to be the sum of expression of its constituent isoforms in TPM. Suppose there are d isoforms for a given gene, then

$$\mathbf{Y1}_d \sim N\left(\mathbf{XB1}_d, \sum_{i=1}^{2} \mathbf{1}_d^T \mathbf{\Gamma}_i \mathbf{1}_d \mathbf{V}_i + \mathbf{1}_d^T \mathbf{\Gamma}_e \mathbf{1}_d \mathbf{I}\right).$$

As a result, the overall variance in gene expression is approximately equal to $\mathbf{1}_d^T \mathbf{\Gamma}_1 \mathbf{1}_d + \mathbf{1}_d^T \mathbf{\Gamma}_2 \mathbf{1}_d + \mathbf{1}_d^T \mathbf{\Gamma}_e \mathbf{1}_d$ and gene-level SNP-based heritability becomes

$$h_{\text{SNP}}^2 := \frac{\mathbf{1}_d^T \mathbf{\Gamma}_1 \mathbf{1}_d + \mathbf{1}_d^T \mathbf{\Gamma}_2 \mathbf{1}_d}{\mathbf{1}_d^T \mathbf{\Gamma}_1 \mathbf{1}_d + \mathbf{1}_d^T \mathbf{\Gamma}_2 \mathbf{1}_d + \mathbf{1}_d^T \mathbf{\Gamma}_e \mathbf{1}_d}.$$

Four scenarios are possible based on the above equation: 1. both gene and isoforms are not heritable, 2. both gene and (a subset of) isoforms are heritable, 3. only gene is heritable, 4. only (a subset of) isoforms are heritable.

In practice, gene and isoform-level expression are $\log_2$-transformed, which leads to the distortion of the above-mentioned relationship, where $\log_2\left(\sum_{j=1}^{d} \text{TPM}_j\right) \geq \frac{1}{d}\sum_{j=1}^{d} \log_2\left(d \cdot \text{TPM}_j\right) = \log_2 d + \frac{1}{d}\sum_{j=1}^{d} \log_2\left(\text{TPM}_j\right)$. The PsychENCODE expression data was further processed not in TPM but instead in $\log_2$-CPM-TMM, which normalization better accounts for differences in library composition across samples in a large mega-analysis dataset such as PsychENCODE, but complicates the relationship between gene and isoform-level expression.

Estimation and inference in variance components model

We consider both maximum likelihood (ML) and restricted (or residual) maximum likelihood (REML) approaches. REML estimates tend to give less biased estimates of variance components, while ML estimates can still have smaller mean squared error (MSE) and are useful for the likelihood ratio test (LRT). For the univariate response model, the log-likelihood function is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\text{lndet}(\boldsymbol{\Omega}) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_m^2)$. The corresponding score (gradient) vector is

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}}\mathcal{L} &= \mathbf{X}^T\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
\frac{\partial}{\partial\sigma_i^2}\mathcal{L} &= -\frac{1}{2}\text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{V}_i) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}\mathbf{V}_i\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).
\end{aligned}$$

The observed information matrix has elements

$$\begin{aligned}
-\nabla_{\boldsymbol{\beta}}^2\mathcal{L} &= \mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X} \\
-\frac{\partial}{\partial\sigma_i^2\partial\boldsymbol{\beta}}\mathcal{L} &= \mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{V}_i\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
-\frac{\partial}{\partial\sigma_i^2\partial\sigma_j^2}\mathcal{L} &= -\frac{1}{2}\text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{V}_j\boldsymbol{\Omega}^{-1}\mathbf{V}_i) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^T)\boldsymbol{\Omega}^{-1}\mathbf{V}_j\boldsymbol{\Omega}^{-1}\mathbf{V}_i\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).
\end{aligned}$$

The expected (Fisher) information matrix has elements

$$\begin{aligned}
\mathrm{E}\left[-\nabla_{\boldsymbol{\beta}}^2\mathcal{L}\right] &= \mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X} \\
\mathrm{E}\left[-\frac{\partial}{\partial\sigma_i^2\partial\boldsymbol{\beta}}\mathcal{L}\right] &= \mathbf{0} \\
\mathrm{E}\left[-\frac{\partial}{\partial\sigma_i^2\partial\sigma_j^2}\mathcal{L}\right] &= \frac{1}{2}\text{tr}(\boldsymbol{\Omega}^{-1}\mathbf{V}_j\boldsymbol{\Omega}^{-1}\mathbf{V}_i).
\end{aligned}$$

Suppose $\text{rank}(\mathbf{X}) = r$ and columns of $\mathbf{A} \in \mathbb{R}^{n\times(n-r)}$ span the null space of $\mathbf{X}^T$, then the log-likelihood function for REML estimation is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\sigma}^2) &= -\frac{n-r}{2}\ln(2\pi) - \frac{1}{2}\text{lndet}(\mathbf{A}^T\boldsymbol{\Omega}\mathbf{A}) - \frac{1}{2}\mathbf{y}^T\mathbf{A}(\mathbf{A}^T\boldsymbol{\Omega}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \\
&= -\frac{n-r}{2}\ln(2\pi) - \frac{1}{2}\text{lndet}(\mathbf{A}^T\boldsymbol{\Omega}\mathbf{A}) - \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{y},
\end{aligned}$$

51

where $\mathbf{P} = \mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{\Omega}^{-1}$. Due to the uniqueness of $\mathbf{P}$, we can choose the columns of $\mathbf{A}$ to form an orthogonal basis such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. The score vector, observed and Fisher information matrices for this residual log-likelihood are similar to the above with $\mathbf{\Omega}$ replaced by $\mathbf{A}^T\mathbf{\Omega}\mathbf{A}$, $\mathbf{V}_i$ by $\mathbf{A}^T\mathbf{V}_i\mathbf{A}$ for $i = 1, \dots, m$, and $\mathbf{y}$ by $\mathbf{A}^T\mathbf{y}$, and hence

$$\frac{\partial}{\partial\sigma_i^2}\mathcal{L} = -\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{V}_i) + \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{V}_i\mathbf{P}\mathbf{y}$$

$$-\frac{\partial}{\partial\sigma_i^2\partial\sigma_j^2}\mathcal{L} = -\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{V}_j\mathbf{P}\mathbf{V}_i) + \mathbf{y}^T\mathbf{P}\mathbf{V}_j\mathbf{P}\mathbf{V}_i\mathbf{P}\mathbf{y}$$

$$\mathrm{E}\left[-\frac{\partial}{\partial\sigma_i^2\partial\sigma_j^2}\mathcal{L}\right] = \frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{V}_j\mathbf{P}\mathbf{V}_i).$$

For the multivariate response model, the log-likelihood function is

$$\mathcal{L}(\mathbf{B}, \mathbf{\Gamma}) = -\frac{nd}{2}\ln(2\pi) - \frac{1}{2}\mathrm{lndet}(\mathbf{\Omega}) - \frac{1}{2}\mathrm{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T\mathbf{\Omega}^{-1}\mathrm{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}),$$

where $\mathbf{\Gamma} = (\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_m)$. To ensure that $\mathbf{\Gamma}_i$ is positive semidefinite, we reparametrize with its Cholesky factor $\mathbf{L}_i$ such that $\mathbf{\Gamma}_i = \mathbf{L}_i\mathbf{L}_i^T$ for $i = 1, \dots, m$. Then the score vector is

$$\mathrm{D}_{\mathrm{vec}\ \mathbf{B}}\mathcal{L} = \mathrm{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T\mathbf{\Omega}^{-1}(\mathbf{I}_d \otimes \mathbf{X})$$

$$\mathrm{D}_{\mathrm{vech}\ \mathbf{L}_i}\mathcal{L}$$
$$= \frac{1}{2}[\mathrm{vec}(\mathbf{\Omega}^{-1}\mathrm{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B})\mathrm{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T\mathbf{\Omega}^{-1}) - \mathrm{vec}\ \mathbf{\Omega}^{-1}]^T$$
$$\times (\mathbf{I}_d \otimes \mathbf{K}_{nd} \otimes \mathbf{I}_n)([(\mathbf{I}_{d^2} + \mathbf{K}_{dd})(\mathbf{L}_i \otimes \mathbf{I}_d)] \otimes \mathrm{vec}\ \mathbf{V}_i)\mathbf{D}_d,$$

where $\mathbf{K}_{nd}$ is the $nd \times nd$ commutation matrix and $\mathbf{D}_d$ the $d^2 \times \frac{d(d+1)}{2}$ duplication matrix. Let $\mathbf{W}_i = (\mathbf{I}_d \otimes \mathbf{K}_{nd} \otimes \mathbf{I}_n)([(\mathbf{I}_{d^2} + \mathbf{K}_{dd})(\mathbf{L}_i \otimes \mathbf{I}_d)] \otimes \mathrm{vec}\ \mathbf{V}_i)\mathbf{D}_d$ and $\mathbf{R}$ be the $n \times d$ matrix satisfying $\mathrm{vec}\ \mathbf{R} = \mathbf{\Omega}^{-1}\mathrm{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B})$. Then the observed information matrix has elements

$$-\frac{\partial^2}{\partial(\text{vec } \mathbf{B})^T \partial(\text{vec } \mathbf{B})}\mathcal{L}$$
$$= (\mathbf{I}_d \otimes \mathbf{X}^T)\mathbf{\Omega}^{-1}(\mathbf{I}_d \otimes \mathbf{X})$$
$$-\frac{\partial^2}{\partial(\text{vech } \mathbf{L}_i)^T \partial(\text{vec } \mathbf{B})}\mathcal{L}$$
$$= (\text{vec } \mathbf{R}^T \otimes [(\mathbf{I}_d \otimes \mathbf{X}^T)\mathbf{\Omega}^{-1}])\mathbf{W}_i$$
$$-\frac{\partial^2}{\partial(\text{vech } \mathbf{L}_j)^T \partial(\text{vech } \mathbf{L}_i)}\mathcal{L}$$
$$= -\frac{1}{2}\mathbf{W}_i^T[\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1} - (\text{vec } \mathbf{R}\text{vec } \mathbf{R}^T) \otimes \mathbf{\Omega}^{-1}$$
$$-\mathbf{\Omega}^{-1} \otimes (\text{vec } \mathbf{R}\text{vec } \mathbf{R}^T)]\mathbf{W}_j.$$

The expected (Fisher) information matrix has elements

$$\text{E}\left[-\frac{\partial^2}{\partial(\text{vec } \mathbf{B})^T \partial(\text{vec } \mathbf{B})}\mathcal{L}\right] = (\mathbf{I}_d \otimes \mathbf{X}^T)\mathbf{\Omega}^{-1}(\mathbf{I}_d \otimes \mathbf{X})$$

$$\text{E}\left[-\frac{\partial^2}{\partial(\text{vech } \mathbf{L}_i)^T \partial(\text{vec } \mathbf{B})}\mathcal{L}\right] = \mathbf{0}$$

$$\text{E}\left[-\frac{\partial^2}{\partial(\text{vech } \mathbf{L}_j)^T \partial(\text{vech } \mathbf{L}_i)}\mathcal{L}\right] = \frac{1}{2}\mathbf{W}_i^T(\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1})\mathbf{W}_j.$$

As a sanity check, one can observe that the above Fisher information boils down to the Fisher information for the univariate response model when $d = 1$. For REML estimation in the multivariate response model, the log-likelihood function is

$$\mathcal{L}(\mathbf{\Gamma}) = -\frac{(n-r)d}{2}\ln(2\pi) - \frac{1}{2}\text{lndet}(\widetilde{\mathbf{\Omega}}) - \frac{1}{2}\text{vec}(\mathbf{A}^T\mathbf{Y})^T\widetilde{\mathbf{\Omega}}^{-1}\text{vec}(\mathbf{A}^T\mathbf{Y}),$$

where $\widetilde{\mathbf{\Omega}} = \sum_{i=1}^{m}\mathbf{\Gamma}_i \otimes (\mathbf{A}^T\mathbf{V}_i\mathbf{A})$. Let $\widetilde{\mathbf{V}}_i = \mathbf{A}^T\mathbf{V}_i\mathbf{A}$ and $\widetilde{\mathbf{W}}_i = (\mathbf{I}_d \otimes \mathbf{K}_{(n-r)d} \otimes \mathbf{I}_{n-r})([(\mathbf{I}_{d^2} + \mathbf{K}_{dd})(\mathbf{L}_i \otimes \mathbf{I}_d)] \otimes \text{vec } \widetilde{\mathbf{V}}_i)\mathbf{D}_d$, then the corresponding score vector, observed and Fisher information matrices are

$$
\begin{aligned}
\mathrm{D}_{\mathrm{vech}\mathbf{L}_i}\mathcal{L} &= \frac{1}{2}\mathrm{vec}(\widetilde{\boldsymbol{\Omega}}^{-1}[\mathrm{vec}(\mathbf{A}^T\mathbf{Y})\mathrm{vec}(\mathbf{A}^T\mathbf{Y})^T - \widetilde{\boldsymbol{\Omega}}]\widetilde{\boldsymbol{\Omega}}^{-1})^T\widetilde{\mathbf{W}}_i
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial(\mathrm{vech}\mathbf{L}_j)^T\partial(\mathrm{vech}\mathbf{L}_i)}&\mathcal{L} \\
&= -\frac{1}{2}\widetilde{\mathbf{W}}_i^T[\widetilde{\boldsymbol{\Omega}}^{-1}\otimes\widetilde{\boldsymbol{\Omega}}^{-1} - (\widetilde{\boldsymbol{\Omega}}^{-1}\mathrm{vec}(\mathbf{A}^T\mathbf{Y})\mathrm{vec}(\mathbf{A}^T\mathbf{Y})^T\widetilde{\boldsymbol{\Omega}}^{-1})\otimes\widetilde{\boldsymbol{\Omega}}^{-1} \\
&\quad -\widetilde{\boldsymbol{\Omega}}^{-1}\otimes(\widetilde{\boldsymbol{\Omega}}^{-1}\mathrm{vec}(\mathbf{A}^T\mathbf{Y})\mathrm{vec}(\mathbf{A}^T\mathbf{Y})^T\widetilde{\boldsymbol{\Omega}}^{-1})]\widetilde{\mathbf{W}}_j
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[-\frac{\partial^2}{\partial(\mathrm{vech}\mathbf{L}_j)^T\partial(\mathrm{vech}\mathbf{L}_i)}\mathcal{L}\right] &= \frac{1}{2}\widetilde{\mathbf{W}}_i^T(\widetilde{\boldsymbol{\Omega}}^{-1}\otimes\widetilde{\boldsymbol{\Omega}}^{-1})\widetilde{\mathbf{W}}_j.
\end{aligned}
$$

Variance components parameters are usually estimated using an iterative algorithm such as Fisher scoring and expectation-maximization (EM) algorithms. Fisher scoring uses the expected information matrix derived above instead of the observed information matrix in Newton's method. To derive the EM algorithm, we let $\boldsymbol{\Omega}_i = \boldsymbol{\Gamma}_i \otimes \mathbf{V}_i$, $r_i = \mathrm{rank}(\mathbf{V}_i)$, $s_i = \mathrm{rank}(\boldsymbol{\Gamma}_i)$, and $\det^+(\boldsymbol{\Gamma}_i)$ denote the pseudo-determinant of $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Gamma}_i^+$ the pseudo-inverse of $\boldsymbol{\Gamma}_i$. Additionally, if we let $\mathrm{vec}\,\mathbf{R}^{(t)} = \boldsymbol{\Omega}^{-(t)}\mathrm{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)})$, then Q-function for the EM algorithm in the multivariate response model is

$$
\begin{aligned}
Q(\boldsymbol{\Gamma} \mid \mathbf{B}^{(t)}, \boldsymbol{\Gamma}^{(t)}) \\
= \sum_{i=1}^{m}\left[-\frac{nd}{2}\ln(2\pi) - \frac{r_i}{2}\mathrm{lndet}^+(\boldsymbol{\Gamma}_i) - \frac{s_i}{2}\mathrm{lndet}^+(\mathbf{V}_i)\right] \\
- \sum_{i=1}^{m}\frac{1}{2}\mathrm{vec}\,\mathbf{R}^{(t)T}(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^+\boldsymbol{\Gamma}_i^{(t)}\otimes\mathbf{V}_i)\mathrm{vec}\,\mathbf{R}^{(t)} \\
- \sum_{i=1}^{m}\frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^+\otimes\mathbf{V}_i\mathbf{V}_i^+) \\
+ \sum_{i=1}^{m}\frac{1}{2}\mathrm{tr}[(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^+\boldsymbol{\Gamma}_i^{(t)}\otimes\mathbf{V}_i)\boldsymbol{\Omega}^{-(t)}]
\end{aligned}
$$

$$= \sum_{i=1}^{m} \left[-\frac{nd}{2} \ln(2\pi) - \frac{r_i}{2} \operatorname{lndet}^+(\mathbf{\Gamma}_i) - \frac{s_i}{2} \operatorname{lndet}^+(\mathbf{V}_i) \right]$$

$$- \sum_{i=1}^{m} \frac{1}{2} \operatorname{tr}(\mathbf{R}^{(t)^T} \mathbf{V}_i \mathbf{R}^{(t)} \mathbf{\Gamma}_i^{(t)} \mathbf{\Gamma}_i^{+} \mathbf{\Gamma}_i^{(t)}) - \sum_{i=1}^{m} \frac{r_i}{2} \operatorname{tr}(\mathbf{\Gamma}_i^{(t)} \mathbf{\Gamma}_i^{+})$$

$$+ \sum_{i=1}^{m} \frac{1}{2} \operatorname{tr}\left[(\mathbf{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i) \odot \mathbf{\Omega}^{-(t)}] (\mathbf{I}_d \otimes \mathbf{1}_n) \mathbf{\Gamma}_i^{(t)} \mathbf{\Gamma}_i^{+} \mathbf{\Gamma}_i^{(t)} \right],$$

where $\odot$ denotes the Hadamard (elementwise) product. Hence, the ECM updates are

$$\operatorname{vec} \mathbf{B}^{(t)}$$
$$= [(\mathbf{I}_d \otimes \mathbf{X}^T) \mathbf{\Omega}^{-(t)} (\mathbf{I}_d \otimes \mathbf{X})]^{-1} (\mathbf{I}_d \otimes \mathbf{X}^T) \mathbf{\Omega}^{-(t)} \operatorname{vec} \mathbf{Y}$$
$$\mathbf{\Gamma}_i^{(t+1)}$$
$$= \frac{1}{r_i} \mathbf{\Gamma}_i^{(t)} [\mathbf{R}^{(t)^T} \mathbf{V}_i \mathbf{R}^{(t)} - (\mathbf{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i) \odot \mathbf{\Omega}^{-(t)}] (\mathbf{I}_d \otimes \mathbf{1}_n)] \mathbf{\Gamma}_i^{(t)} + \mathbf{\Gamma}_i^{(t)}.$$

The EM algorithm is a special case of the minorization-maximization (MM) algorithm. A different formulation of the MM algorithm implements a block ascent strategy by alternatively updating $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ (Zhou et al. 2019) such that

$$\boldsymbol{\beta}^{(t)} = (\mathbf{X}^T \mathbf{\Omega}^{-(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Omega}^{-(t)} \mathbf{y}$$

$$\sigma_i^{2(t+1)} = \sigma_i^{2(t)} \sqrt{\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \mathbf{\Omega}^{-(t)} \mathbf{V}_i \mathbf{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})}{\operatorname{tr}(\mathbf{\Omega}^{-(t)} \mathbf{V}_i)}}.$$

For the multivariate response model, the updates are

$$\operatorname{vec} \mathbf{B}^{(t)}$$
$$= [(\mathbf{I}_d \otimes \mathbf{X}^T) \mathbf{\Omega}^{-(t)} (\mathbf{I}_d \otimes \mathbf{X})]^{-1} (\mathbf{I}_d \otimes \mathbf{X}^T) \mathbf{\Omega}^{-(t)} \operatorname{vec} \mathbf{Y}$$
$$\mathbf{\Gamma}_i^{(t+1)}$$
$$= \mathbf{L}_i^{-(t)T} [\mathbf{L}_i^{(t)T} (\mathbf{\Gamma}_i^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \mathbf{\Gamma}_i^{(t)}) \mathbf{L}_i^{(t)}]^{1/2} \mathbf{L}_i^{-(t)},$$

where $\mathbf{L}_i$ is the Cholesky factor of $(\mathbf{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i) \odot \mathbf{\Omega}^{-(t)}] (\mathbf{I}_d \otimes \mathbf{1}_n)$ and $\operatorname{vec} \mathbf{R}^{(t)} = \mathbf{\Omega}^{-(t)} \operatorname{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)})$. As another sanity check, one can see that the above updates become the updates for the univariate response model when $d = 1$. For complete derivation, please see Zhou et al. 2019 and for REML estimation, the updates are basically of the same form. Compared to

Newton's method or Fisher scoring, the MM algorithm is numerically stable and computationally efficient. Although the number of iterations it takes is usually longer than Fisher scoring, the computational burden is smaller in each iteration. The MM algorithm also gracefully respects the nonnegativity (or positive semidefinite) constraint of variance components. Compared to the EM algorithm, the MM updates converge more quickly, since the surrogate function hugs the log-likelihood function more tightly (Zhou et al. 2019). In the simplest heritability and GWAS model where there are two variance components, repeated matrix inversion can be avoided by the generalized eigenvalue decomposition of the two kernel matrices (Zhou et al. 2019). When there are more than two variance components, a matrix inversion is inevitable in each update, so the MM algorithm is not scalable to biobank-level data, and we recommend applying this method for datasets of size up to $n \times d = 50000$. In this study, we take advantage of the aforementioned MM algorithm in estimating parameters in variance components models.

Univariate variance components linear mixed models were fit for gene and isoform-level expression with the above-mentioned mean effects covariates. We specified three variance components, two of which capture *cis*- and *trans*-SNP genetic effects. We defined *cis*-SNPs as those within 1 Mb window of gene start and gene end sites, and *trans*-SNPs as all the other SNPs. Based on this definition, 24,754 genes and 93,030 isoform had non-zero *cis*-SNPs. The same set of *cis*-SNPs was used for a given gene and its constituent isoforms for direct comparison. The mean number of *cis*-SNPs were 3,264 and 3,274 for these genes and isoforms, respectively. Variance components parameters were estimated using both ML and REML. The maximum number of iterations was set to 3,000. In total, 22,965 genes and 89,926 isoforms had converged

estimates. In most (if not all) cases where the estimates failed to converge, ML estimation was the issue.

As part of sensitivity analyses, we fit additional univariate models, including three variance components models with different definitions of *cis*-SNP windows (i.e. ±250 kb window and the entire chromosome), two variance components model with a single parameter capturing the entire genetic effects, and two variance components model with only *cis*-SNP effects (1 Mb window). In theory, model selection that best fits the data using a statistical criterion such as AIC metric is possible, but we did not conduct such analyses.

Multivariate variance components linear mixed models were fit for isoform-level expression. We specified three variance components, one of which captures *cis*-SNP genetic effects and the other *trans* effects. We used the same set of *cis*-SNPs that were used for univariate models. To reduce computational burden and the number of variance components parameters that need to be estimated, given limited sample size of the PsychENCODE dataset, we ran the multivariate model for isoforms with significant heritability estimates in a univariate model at $P < 0.05$. For isoforms that are perfectly correlated in expression, we included only one isoform of the two. Variance components parameters were estimated using REML. Note that the MM algorithm is numerically stable to fit even the non-heritable isoforms, but we chose not to. Lastly, pairwise bivariate variance components models were fit with the same scheme as multivariate models. For each gene with at least two heritable isoforms, the model was fit to all pairwise combinations of isoforms.

Inference on variance components parameters was done using a variation of the likelihood ratio test (LRT). Here, we tested the null hypothesis that both variance components corresponding to *cis*- and *trans*-SNP genetic effects are zero by fitting a null model with only a single residual variance components parameter. Then we compared the log-likelihood from ML estimation, which difference times two is assumed to follow a mixture of χ^2 distributions (Molenberghs and Verbeke 2007) as

$$2(\mathcal{L} - \mathcal{L}_0) \sim \frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2.$$

Standard errors for both genetic variances and genetic covariances were estimated using the Fisher information matrix, where

$$\mathrm{E}\left[-\frac{\partial^2}{\partial(\text{vech } \mathbf{\Gamma}_j)^T \partial(\text{vech } \mathbf{\Gamma}_i)}\mathcal{L}\right] = \frac{1}{2}\mathbf{U}_i^T(\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1})\mathbf{U}_j$$

and $\mathbf{U}_i = (\mathbf{I}_d \otimes \mathbf{K}_{nd} \otimes \mathbf{I}_n)(\mathbf{I}_{d^2} \otimes \text{vec } \mathbf{V}_i)\mathbf{D}_d$. Note that this form is slightly different from the one shown earlier with $\mathbf{U}_i$ replaced by $\mathbf{W}_i$. This is because we previously imposed the positive semidefinite constraint by reparametrizing $\mathbf{\Gamma}_i = \mathbf{L}_i\mathbf{L}_i^T$. Then by using the vec-transpose operator,

$$\begin{aligned}
\mathbf{U}_i &= \mathbf{I}_d \otimes (\mathbf{K}_{nd} \otimes \mathbf{I}_n)(\mathbf{I}_d \otimes \text{vec } \mathbf{V}_i)\mathbf{D}_d \\
&= \mathbf{I}_d \otimes (\mathbf{K}_{nd} \otimes \mathbf{I}_n)(\mathbf{I}_d \otimes \mathbf{V}_i^{(n)})\mathbf{D}_d \\
&= \mathbf{I}_d \otimes (\mathbf{K}_{nd} \otimes \mathbf{I}_n)(\mathbf{I}_d \otimes \mathbf{V}_i)^{(n)}\mathbf{D}_d \\
&= \mathbf{I}_d \otimes [(\mathbf{I}_d \otimes \mathbf{V}_i)\mathbf{K}_{dn}]^{(n)}\mathbf{D}_d.
\end{aligned}$$

Naïve computation of the Fisher information matrix is not computationally feasible for even reasonable nd, so we must take advantage of the structure inherent in $\mathbf{U}_i$. Standard errors for heritability and genetic correlation estimates were subsequently calculated using the delta method. Inference on genetic covariance or genetic correlation parameters was done using the Wald test for both bivariate and multivariate models.

Penalization of variance components

To investigate the polygenicity of gene and isoform expression, we also fit a univariate, 23 variance components model with separate SNP effects from each autosome such that

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sum_{i=1}^{22} \sigma_i^2 \mathbf{V}_i + \sigma_e^2 \mathbf{I}).$$

However, due to the limited sample size of PsychENCODE, these estimates were noisy and hence we instead minimized the lasso-penalized negative log-likelihood function (Kim et al. 2021). One of the advantages of the MM algorithm is that it separates the parameters of a problem such that penalized estimation is conducive (Kim et al. 2021; Zhai et al. 2018). For each chromosome, we focused on heritable genes and isoforms ($P < 0.05$) that reside in that chromosome, and for these features, we calculated the number of times each chromosome appeared in the solution path.

Different starting point for the MM algorithm

Due to the non-concavity of the log-likelihood, our estimates can get stuck in a local maximum, which we can mitigate by trying different starting points. For example, we could use a method of moments estimator for variance components parameters $(\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \cdots, \boldsymbol{\Gamma}_m)$ by minimizing the least squares criterion

$$\min_{\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \cdots, \boldsymbol{\Gamma}_m} \| \text{vec}(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})\text{vec}(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^T - \boldsymbol{\Omega} \|_{\text{F}}^2,$$

where $\widehat{\mathbf{B}}$ is the ordinary least squares estimate. Suppose $\mathbf{R} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}$ and $f = \| \text{vec } \mathbf{R} \text{ vec } \mathbf{R}^T - \boldsymbol{\Omega} \|_{\text{F}}^2$, then

$$
\begin{aligned}
f \;&= \operatorname{tr}([\operatorname{vec}\mathbf{R}\,\operatorname{vec}\mathbf{R}^T - \boldsymbol{\Omega}]^T[\operatorname{vec}\mathbf{R}\,\operatorname{vec}\mathbf{R}^T - \boldsymbol{\Omega}]) \\
&= \operatorname{tr}(\operatorname{vec}\mathbf{R}\,\operatorname{vec}\mathbf{R}^T\operatorname{vec}\mathbf{R}\,\operatorname{vec}\mathbf{R}^T \\
&\quad -2\operatorname{vec}\mathbf{R}^T\boldsymbol{\Omega}\operatorname{vec}\mathbf{R} \;+\; \sum_{j=1}^{m}\sum_{i=1}^{m}\boldsymbol{\Gamma}_i\,\boldsymbol{\Gamma}_j \otimes \mathbf{V}_i\mathbf{V}_j) \\
&= \operatorname{tr}(\operatorname{vec}\mathbf{R}\,\operatorname{vec}\mathbf{R}^T\operatorname{vec}\mathbf{R}\,\operatorname{vec}\mathbf{R}^T) \\
&\quad -2\operatorname{vec}\mathbf{R}^T\boldsymbol{\Omega}\operatorname{vec}\mathbf{R} \;+\; \sum_{j=1}^{m}\sum_{i=1}^{m}\operatorname{tr}(\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_j)\operatorname{tr}(\mathbf{V}_i\mathbf{V}_j)
\end{aligned}
$$

Then if we differentiate with respect to $\boldsymbol{\Gamma}_i$,

$$
\begin{aligned}
\mathrm{d}f \;&= -2\operatorname{tr}(\mathbf{R}^T\mathbf{V}_i\mathbf{R}\,\mathrm{d}\boldsymbol{\Gamma}_i) + 2\sum_{j=1}^{m}\operatorname{tr}(\mathbf{V}_i\mathbf{V}_j)\operatorname{tr}(\boldsymbol{\Gamma}_j\,\mathrm{d}\boldsymbol{\Gamma}_i) \\
\therefore \frac{\partial}{\partial\boldsymbol{\Gamma}_i}f \;&= -2\mathbf{R}^T\mathbf{V}_i\mathbf{R} + 2\sum_{j=1}^{m}\operatorname{tr}(\mathbf{V}_i\mathbf{V}_j)\boldsymbol{\Gamma}_j \\
&\rightarrow \sum_{j=1}^{m}\operatorname{tr}(\mathbf{V}_i\mathbf{V}_j)\boldsymbol{\Gamma}_j = \mathbf{R}^T\mathbf{V}_i\mathbf{R}.
\end{aligned}
$$

This leads to m system of linear equations, which can be solved easily. The estimator derived here boils down to the same estimators previously derived for simpler variance components linear mixed models. For example, the following estimator for multivariate, two variance components models was used in Ge et al. 2016, which is the same as the above.

$$
\begin{aligned}
\boldsymbol{\Gamma}_g \;&= \frac{1}{\operatorname{tr}(\mathbf{A}^2)n - \operatorname{tr}(\mathbf{A})^2}\mathbf{Y}^T[n\mathbf{A} - \operatorname{tr}(\mathbf{A})\mathbf{I}_n]\mathbf{Y}, \\
\boldsymbol{\Gamma}_e \;&= \frac{1}{\operatorname{tr}(\mathbf{A}^2)n - \operatorname{tr}(\mathbf{A})^2}\mathbf{Y}^T[\operatorname{tr}(\mathbf{A})^2\mathbf{I}_n - \operatorname{tr}(\mathbf{A})\mathbf{A}]\mathbf{Y}.
\end{aligned}
$$

The same estimator was previously derived in Ge et al. 2017 and Wu and Sankararaman 2018 for univariate, two variance components models, Pazokitoroudi et al. 2020 for univariate, multiple variance components models, and Wu et al. 2022 for bivariate, two variance components models.

cis-eQTL analyses

We ran eQTL analyses with the same _cis_-SNPs used in variance components models for each gene and isoform using a linear model as implemented in QTLtools (Delaneau et al. 2017). To account for linkage disequilibrium (LD), sample labels were permuted 1,000 times, and each iteration the most significantly associated P value was saved. Such a null distribution of P values was then fit to a beta distribution and the observed P value was subsequently adjusted to give an empirical P value. To account for multiple testing, empirical P values transcriptome-wide were FDR corrected. Conditional analyses were further performed through QTLtools to discover independently associated eQTL signals for gene and isoform expression. The top associated eQTL SNP for each gene and isoform was then used to calculate R^2 or variance explained by index SNPs.

Targeted long-read sequencing of _XRN2_ and isoform quantification in short-read bulk RNA-seq data

To better complete the human brain transcriptome annotation, we compiled long-read sequencing data of the human brain from six different studies (Leung et al. 2021; Palmer et al. 2021; Glinos et al. 2021). This included in-house fetal brain single-cell Iso-Seq data and two adult human brain data shared by PacBio (https://www.pacb.com/connect/datasets/). Note that Glinos et al. 2021 includes data from non-brain tissues, while other studies are strictly from the adult or developing human brain. We started with finalized (e.g. filtered and merged) GTF files from each study, harmonized genomic coordinates via liftover to the GRCh38 human genome build, ran gffcompare v0.12.6 with Gencode v40 as a reference, and filtered for isoforms found in at least two or three studies. gffcompare searches for identical matches for internal exons, while it implements fuzzy matches for terminal exons, which is appropriate here since these exons can be sequenced to

different lengths. We subsequently used the above two filtered/munged GTF files to re-quantify O'Brien et al. 2018 fetal brain bulk RNA-seq data using Salmon v1.8.0. Gene-level expression was calculated to be the sum of its constituent isoform TPM. To repeat gene-level and isoform-level eQTL analyses, genotype data for O'Brien et al. 2018 was filtered for typed variants with individual call rate (> 0.95), minor allele frequency (> 0.01), Hardy-Weinberg equilibrium (HWE) P value ($> 10^{-6}$) and individuals based on genotype call rate (> 0.9) using PLINK 1.9. Genotype data was then phased and imputed with the TOPMed freeze 5 reference panel on the Michigan Imputation Server. After removing SNPs with low imputation quality (R2 < 0.3) and subsetting to 86 individuals of European ancestry, we again filtered SNPs based on individual call rate (> 0.95), minor allele frequency (> 0.05), HWE P value ($> 10^{-6}$) and individuals based on genotype call rate (> 0.95). For eQTL analyses, we modeled either TPM or $\log_2$TPM with age, sex, 4 genetic PCs, RIN, and 21 sequencing PCs from sequencing metrics generated from PicardTools as covariates.

Data availability

PsychENCODE raw data are available at www.doi.org/10.7303/syn12080241 and processed summary-level data are available at Resource.PsychENCODE.org.

Code availability

The code used to perform bioinformatic analyses are available at

https://github.com/mmkim1210/isoforms-genetics. All plots were generated programmatically

using Makie.jl (Danisch and Krumbiegel 2021) and GeneticsMakie.jl (Kim et al. 2022).

The Julia implementation of the MM algorithm is available at

https://github.com/Hua-Zhou/MultiResponseVarianceComponentModels.jl.

3.6 Supplementary Figures
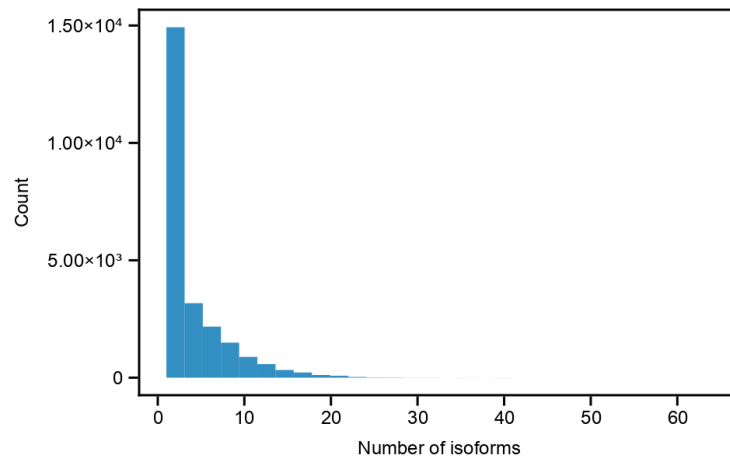
Supplementary Figure 3.1: Distribution of the number of isoforms for 24,905 genes in PsychENCODE. The maximum number of isoforms is 64.
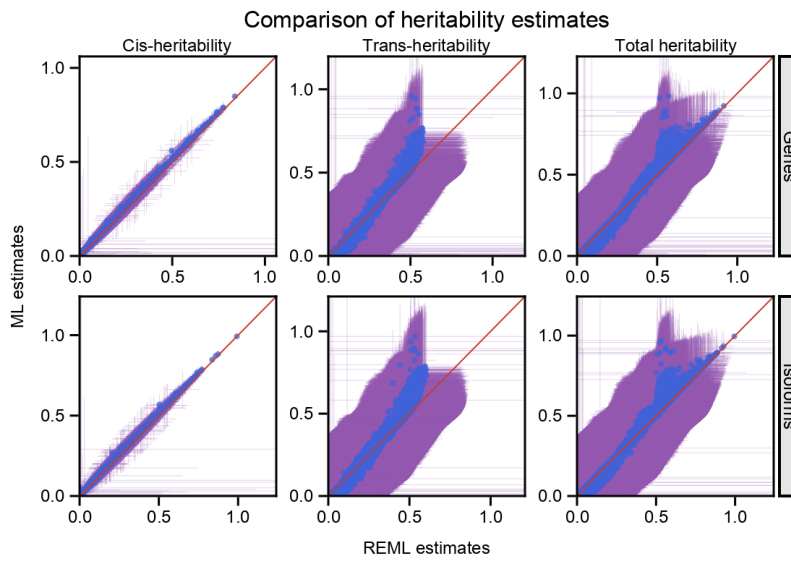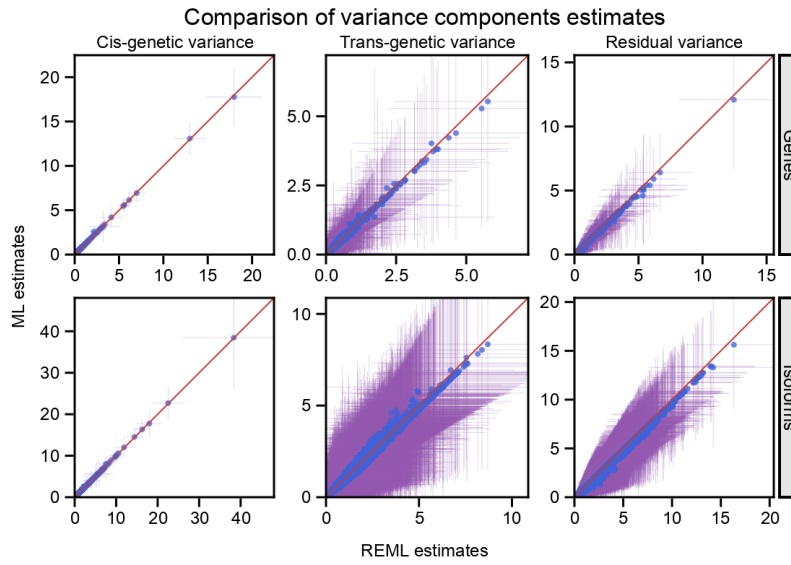
Supplementary Figure 3.2: Comparison of REML and ML estimates in the univariate variance components models. REML and ML estimates were highly concordant for h^2_{cis} and to a lesser degree for h^2_{trans} and h^2_{SNP}. Error bars denote $\pm$ one standard errors. Note higher standard errors for h^2_{trans} and h^2_{SNP} relative to h^2_{cis}. Standard errors are calculated using the Fisher information matrix and the Delta method.
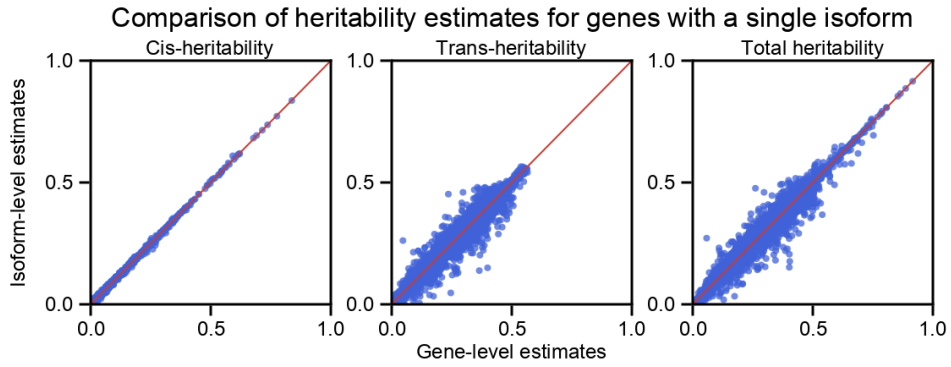
Supplementary Figure 3.3: Comparison of heritability estimates for genes with a single isoform. h^2_{cis} estimates were highly concordant, while more variability was observed for h^2_{trans} and h^2_{SNP} estimates.
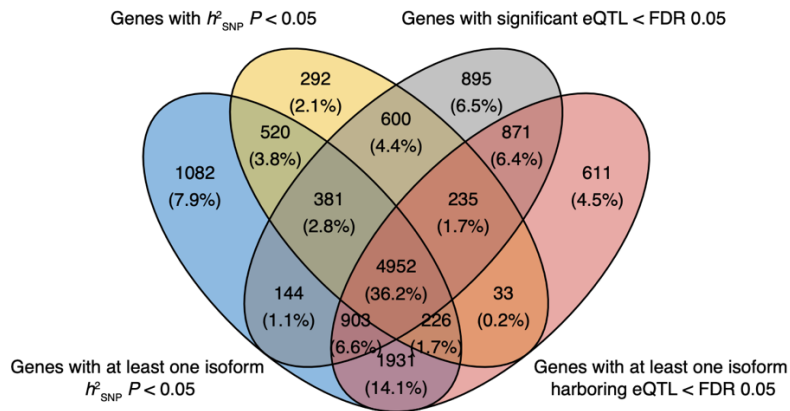
Supplementary Figure 3.4: Comparison between heritability and *cis*-eQTL results. Shown is Venn diagram of overlap between heritable genes and genes harboring eQTL at FDR < 0.05.
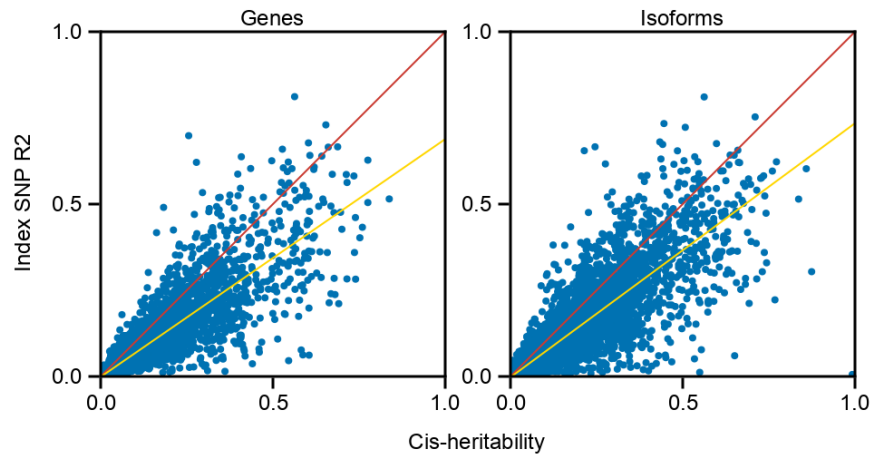
Supplementary Figure 3.5: Sparsity of *cis*-SNP effects. Left, comparison between h^2_{cis} estimates and variance explained by top associated eQTL SNPs for heritable genes. Right, the same comparison but for heritable isoforms.
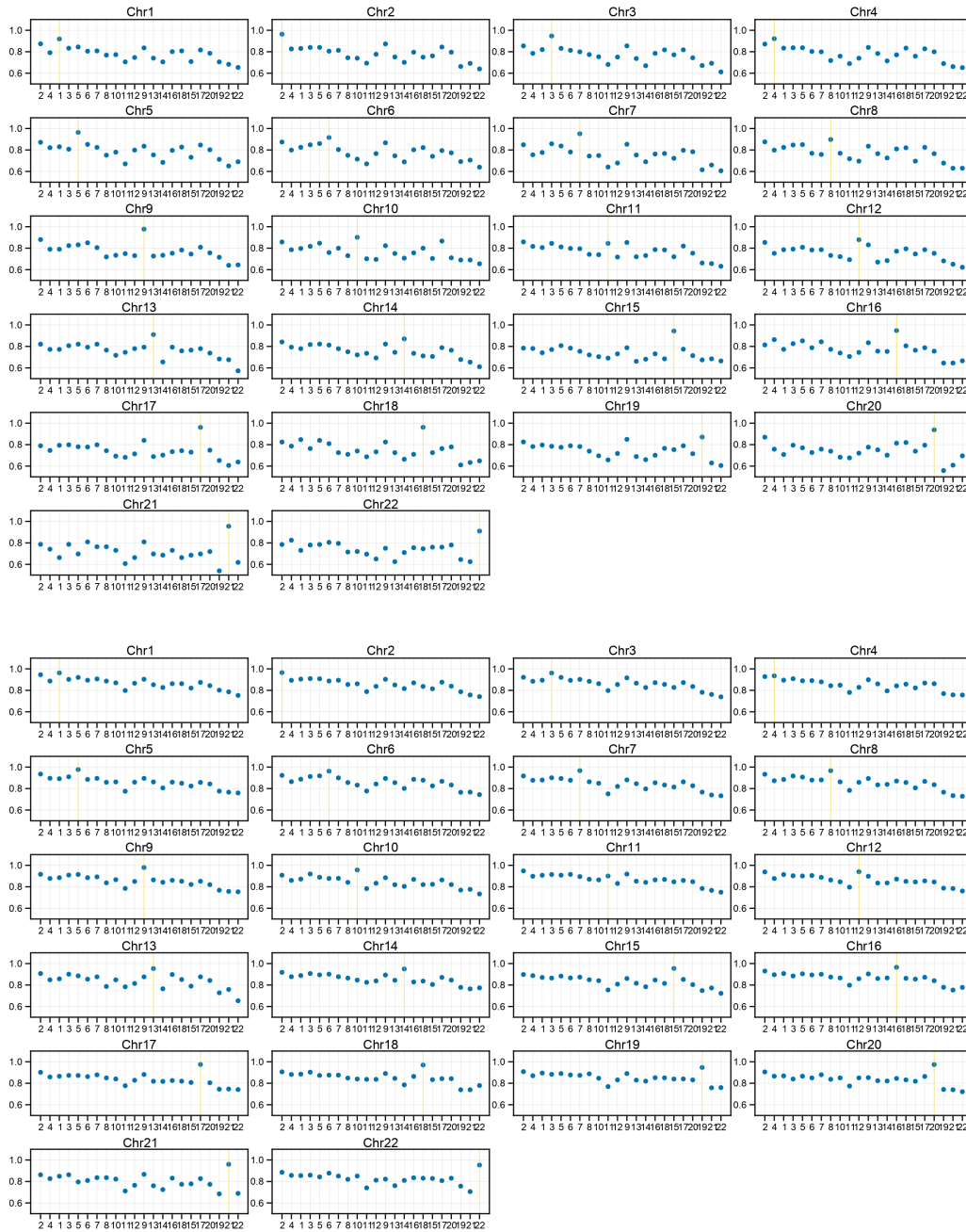
Supplementary Figure 3.6: Polygenicity of *trans*-SNP effects. We fit a univariate penalized variance components model with 23 variance components parameters that capture genetic effects from each autosome and lasso penalty. Top, shown are results for genes. Bottom, shown are results for isoforms. For each chromosome, we focus on heritable genes and isoforms ($P < 0.05$) that reside in that chromosome, and for these features, we calculate the number of times each chromosome appeared in the solution path (Kim et al. 2021). The chromosome the feature is from almost always appeared in the solution path (i.e. yellow lines have the highest proportions), while the rest of the chromosomes appeared in proportion to their number of SNPs. Chromosomes are listed in descending order of the number of SNPs on the x-axis.

67

SOX7 locus

Supplementary Figure 3.7: A close look at the ASD GWAS locus in chromosome 8. GWAS results for 56 complex phenotypes are shown, which span autoimmune, endocrine, psychiatric, cardiovascular disorders, and cancer. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *SOX7* gene. LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), RBC (red blood cell), SCZ (schizophrenia).
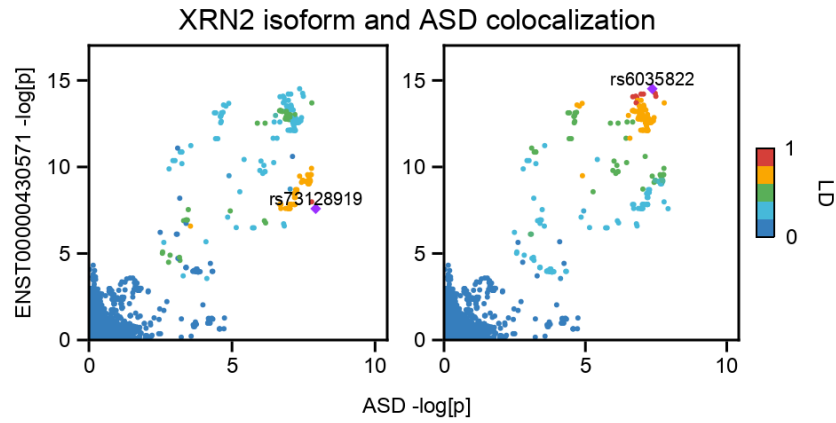
Supplementary Figure 3.8: Colocalization between *XRN2* isoform-level eQTL and ASD GWAS results. Left, LD is colored with respect to the index SNP for ASD GWAS. Right, LD is colored with respect to the index SNP for ENST00000430571 eQTL. LD is calculated with individuals of European ancestry in PsychENCODE.

XRN2 locus
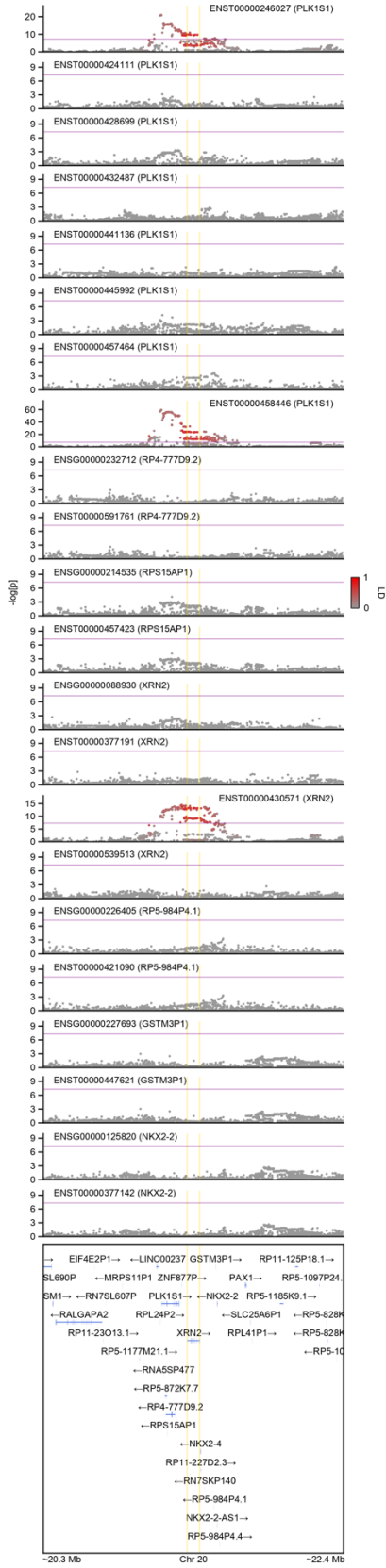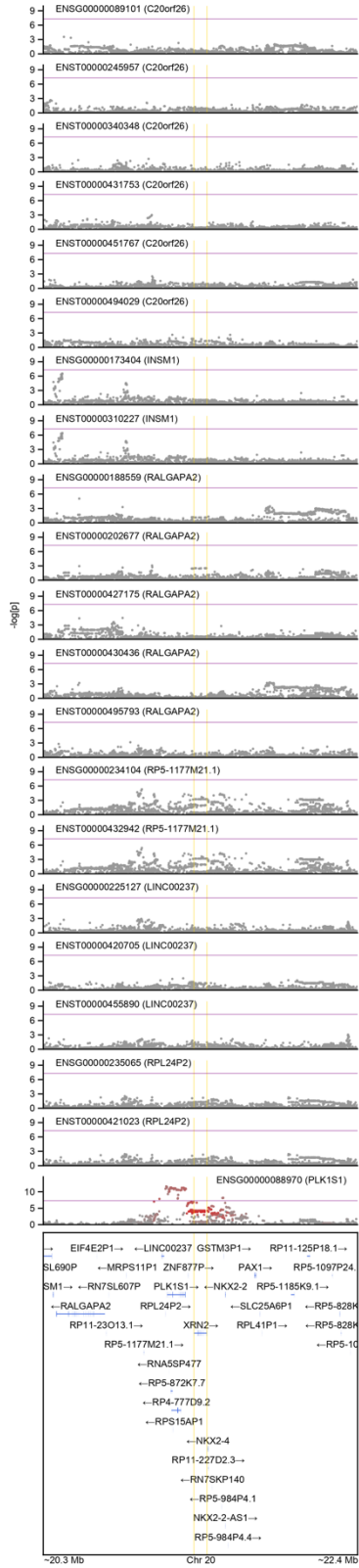
Supplementary Figure 3.9: Absence of colocalization for all other features within the *XRN2* locus for the ASD GWAS signal. Shown is LocusZoom of eQTL signals for both gene- and isoform-level expression for all features within ±1 Mb window of (collapsed) gene start and end sites for *XRN2* gene. Gene names are shown in parentheses. LD is colored with respect to the index SNP for ASD GWAS (rs910805). LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel.

Supplementary Figure 3.10: h^2_{SNP}, r_g, and *cis*-eQTL results for *SYNE1* gene. Follows the same outline as **Figure 3.3** in the main text.

Supplementary Figure 3.11: Colocalization between *SYNE1* isoform-level eQTL and BD GWAS results. Left, LD is colored with respect to the index SNP for BD GWAS. Right, LD is colored with respect to the index SNP for ENST00000466159 eQTL. LD is calculated with individuals of European ancestry in PsychENCODE.

SYNE1 locus
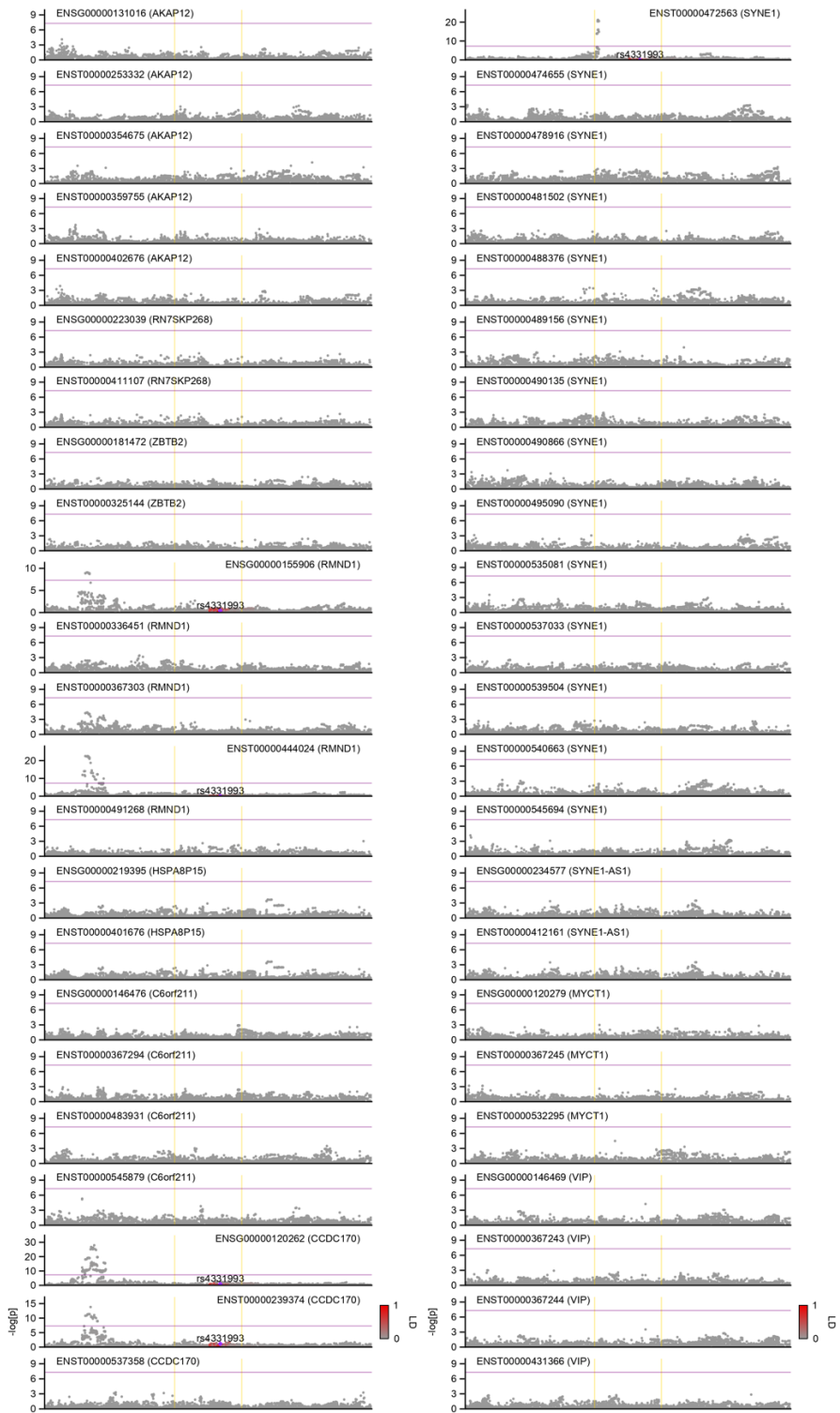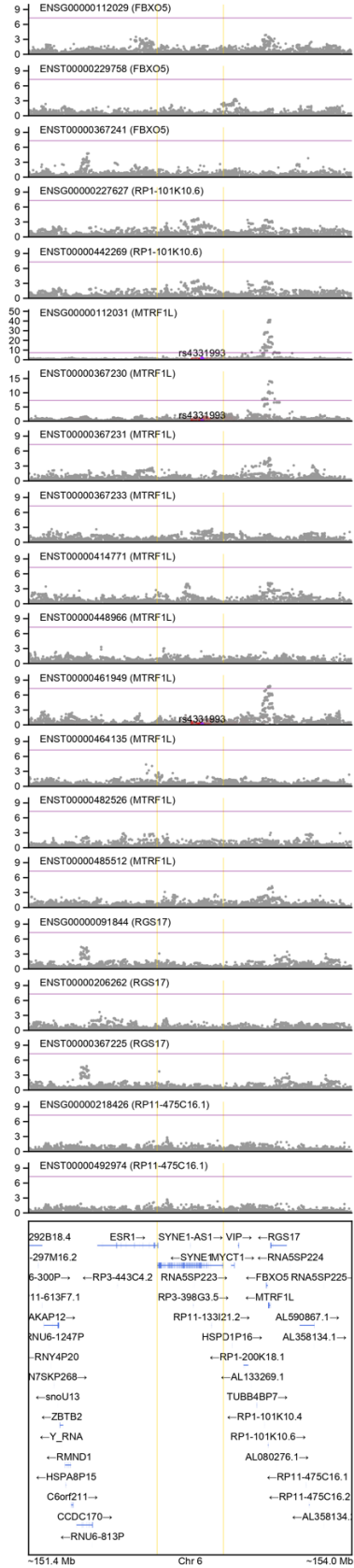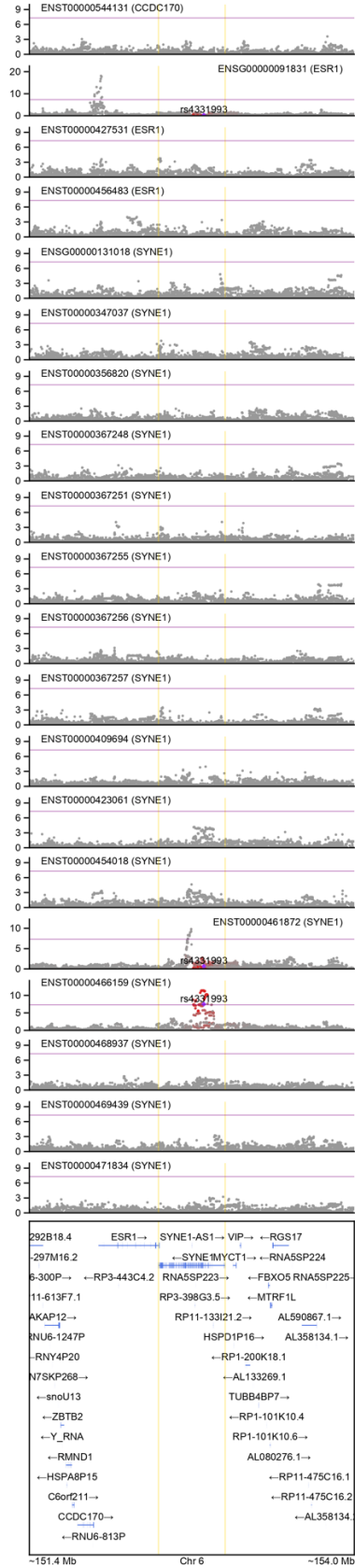
Supplementary Figure 3.12: Absence of colocalization for all other features within the *SYNE1* locus for the BD GWAS signal. Shown is LocusZoom of eQTL signals for both gene- and isoform-level expression for all features within ±1 Mb window of (collapsed) gene start and end sites for *SYNE1* gene. Gene names are shown in parentheses. LD is colored with respect to the index SNP for BD GWAS (rs4331993). LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel.
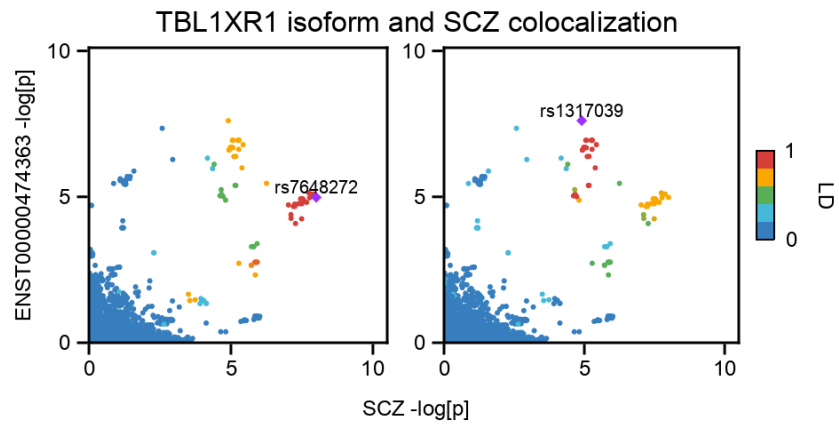
Supplementary Figure 3.13: Colocalization between *TBL1XR1* isoform-level eQTL and SCZ GWAS results. Left, LD is colored with respect to the index SNP for SCZ GWAS. Right, LD is colored with respect to the index SNP for ENST00000474363 eQTL. LD is calculated with individuals of European ancestry in PsychENCODE.
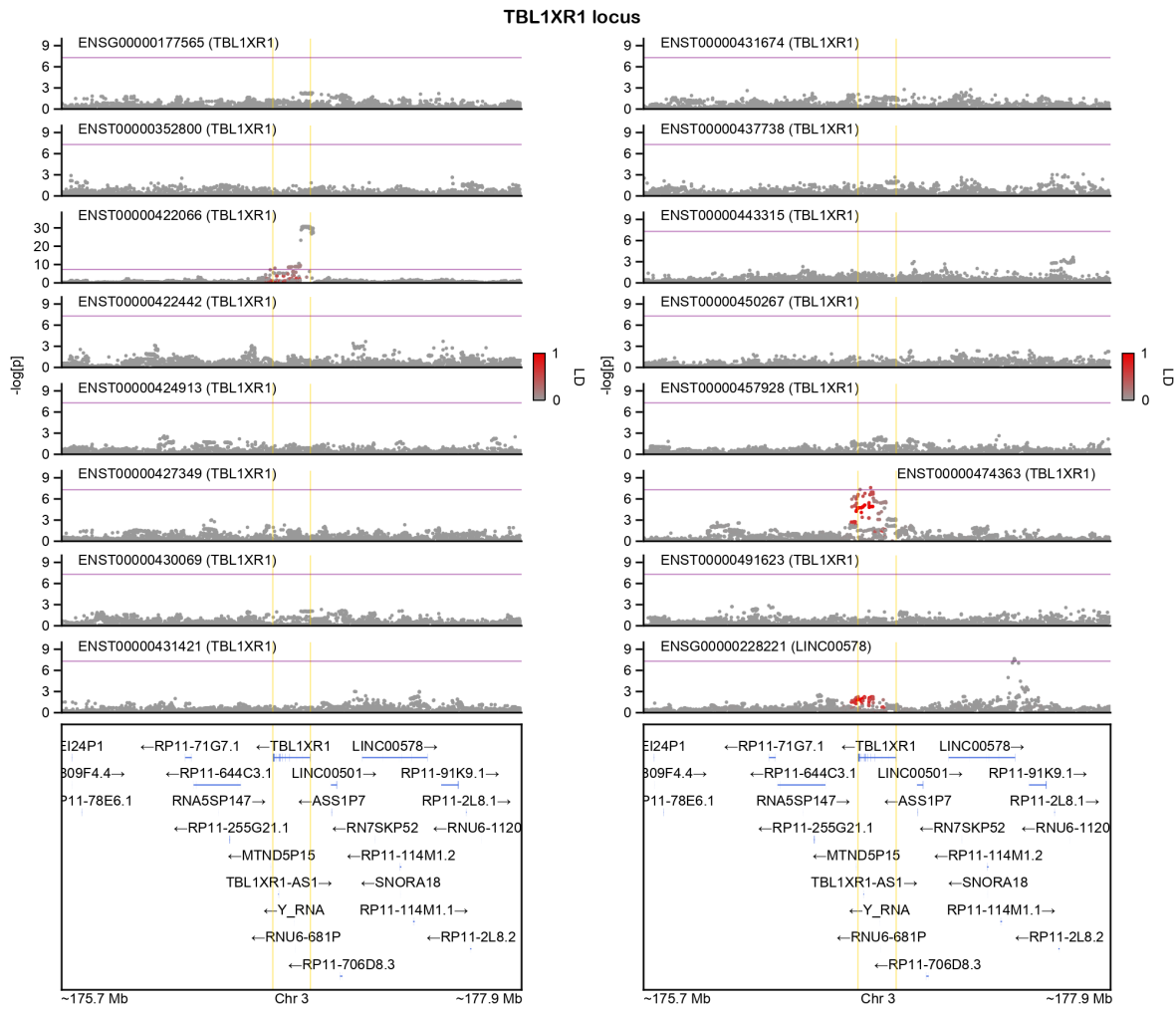
Supplementary Figure 3.14: Absence of colocalization for all other features within the *TBL1XR1* locus for the SCZ GWAS signal. Shown is LocusZoom of eQTL signals for both gene- and isoform-level expression for all features within ±1 Mb window of (collapsed) gene start and end sites for *TBL1XR1* gene. Gene names are shown in parentheses. LD is colored with respect to the index SNP for SCZ GWAS (rs7609876). LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel.
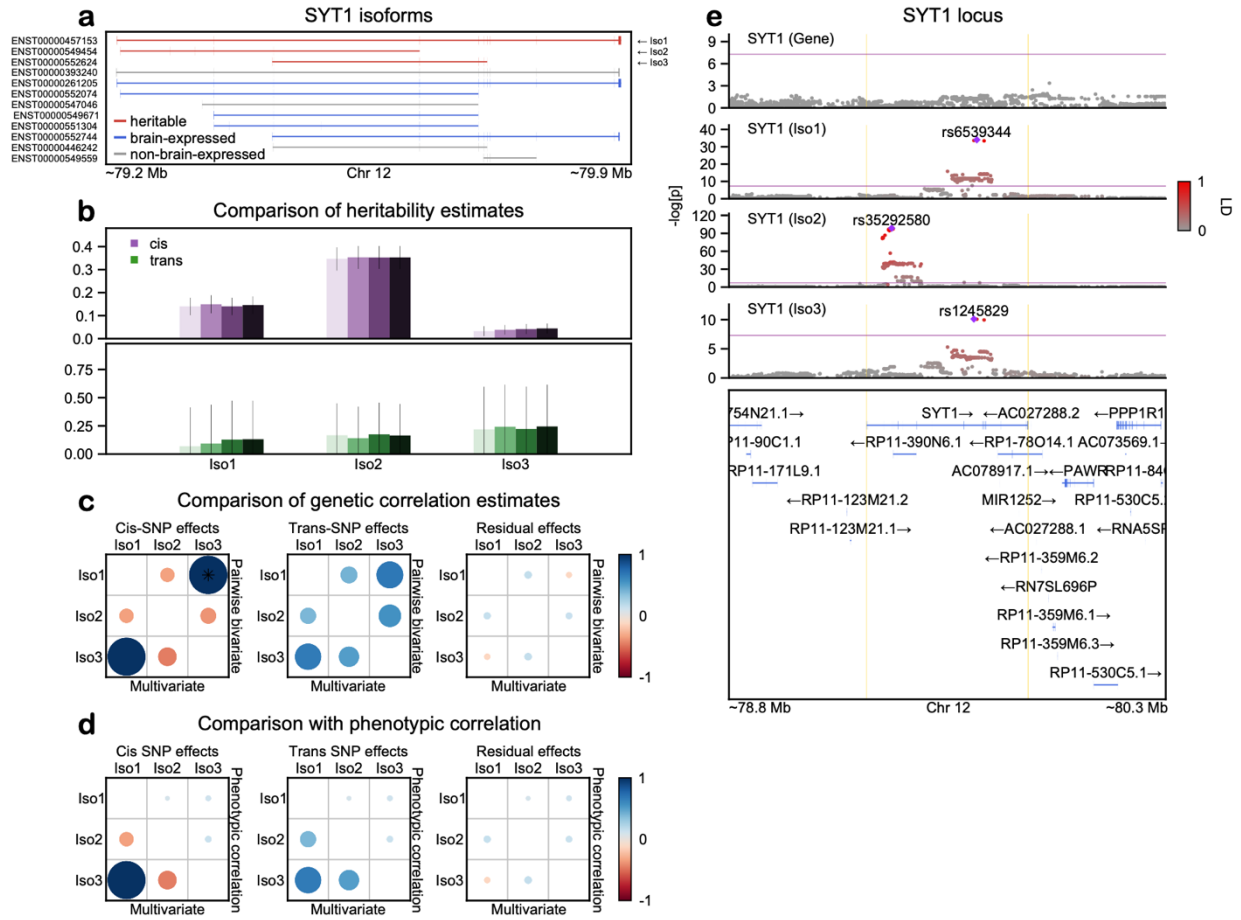
Supplementary Figure 3.15: h^2_{SNP}, r_g, **and** *cis*-**eQTL results for** *SYT1* **gene.** Follows the same outline as **Figure 3.3** in the main text.
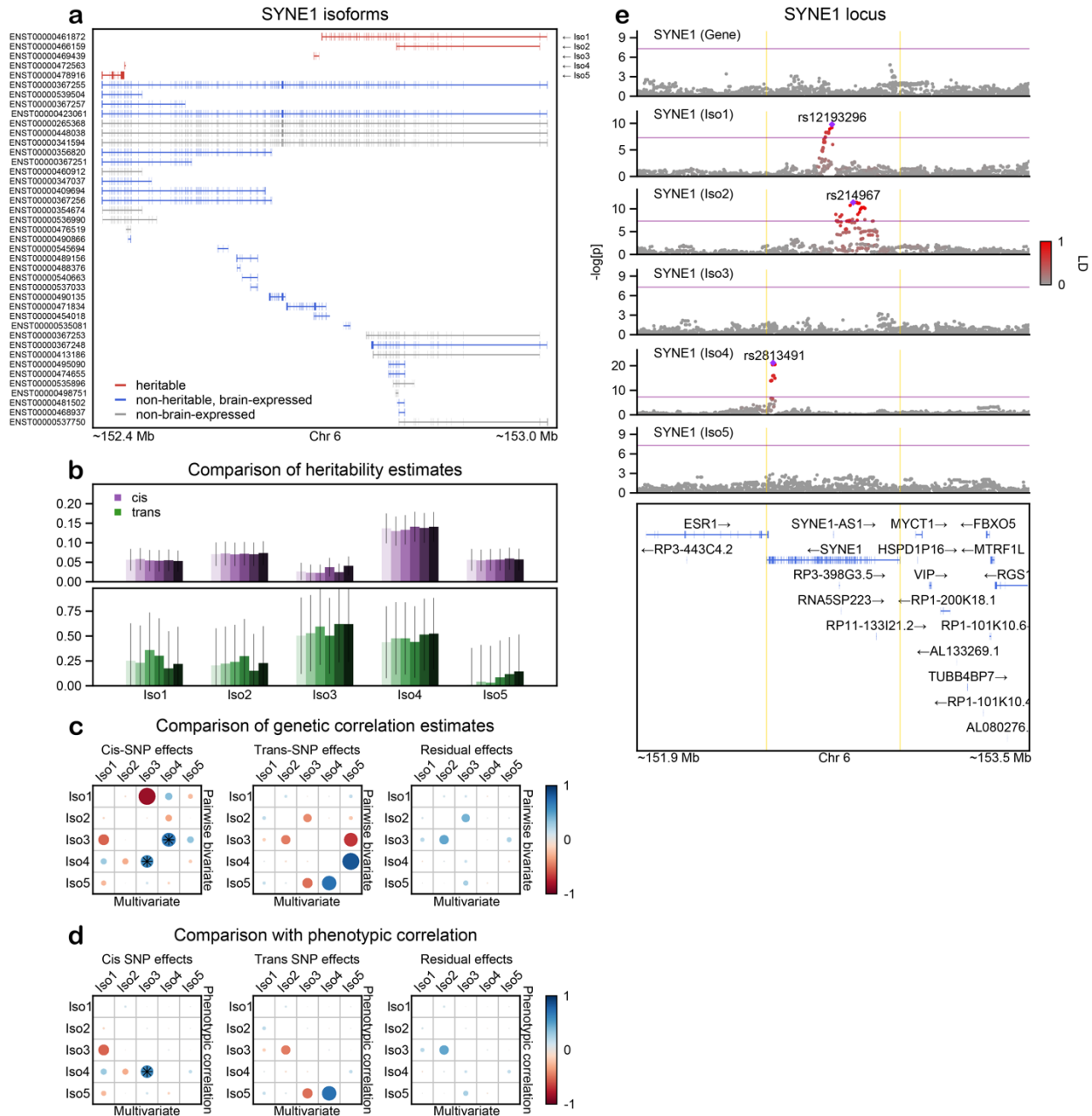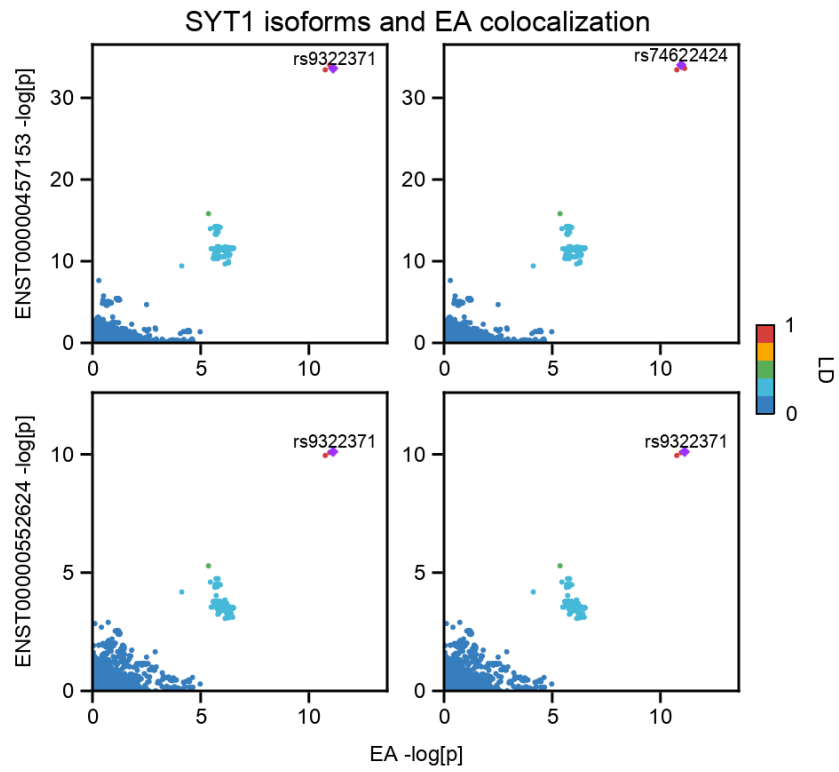
Supplementary Figure 3.16: Colocalization between *SYT1* isoform-level eQTL and EA GWAS results. Left, LD is colored with respect to the index SNP for EA GWAS. Right, LD is colored with respect to the index SNP for corresponding isoform-level eQTL. LD is calculated with individuals of European ancestry in PsychENCODE.
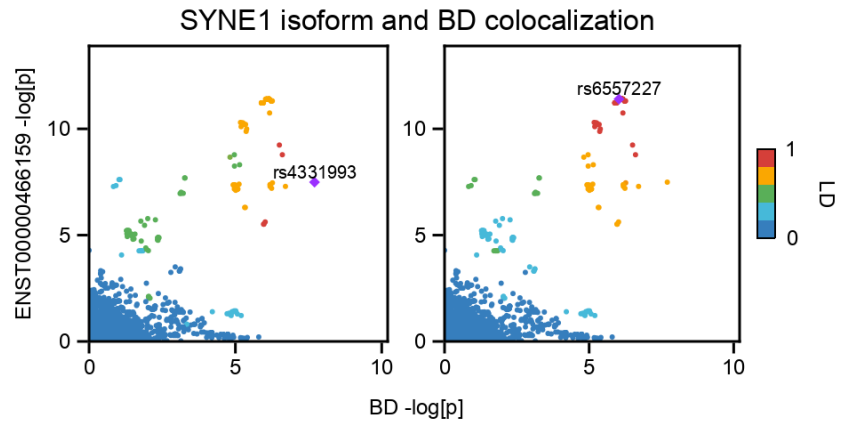
SYT1 locus

Supplementary Figure 3.17: Absence of colocalization for all other features within the *SYT1* locus for the EA GWAS signal. Shown is LocusZoom of eQTL signals for both gene- and isoform-level expression for all features within ±1 Mb window of (collapsed) gene start and end sites for *SYT1* gene. Gene names are shown in parentheses. LD is colored with respect to the index SNP for EA GWAS (rs1245829). LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel.
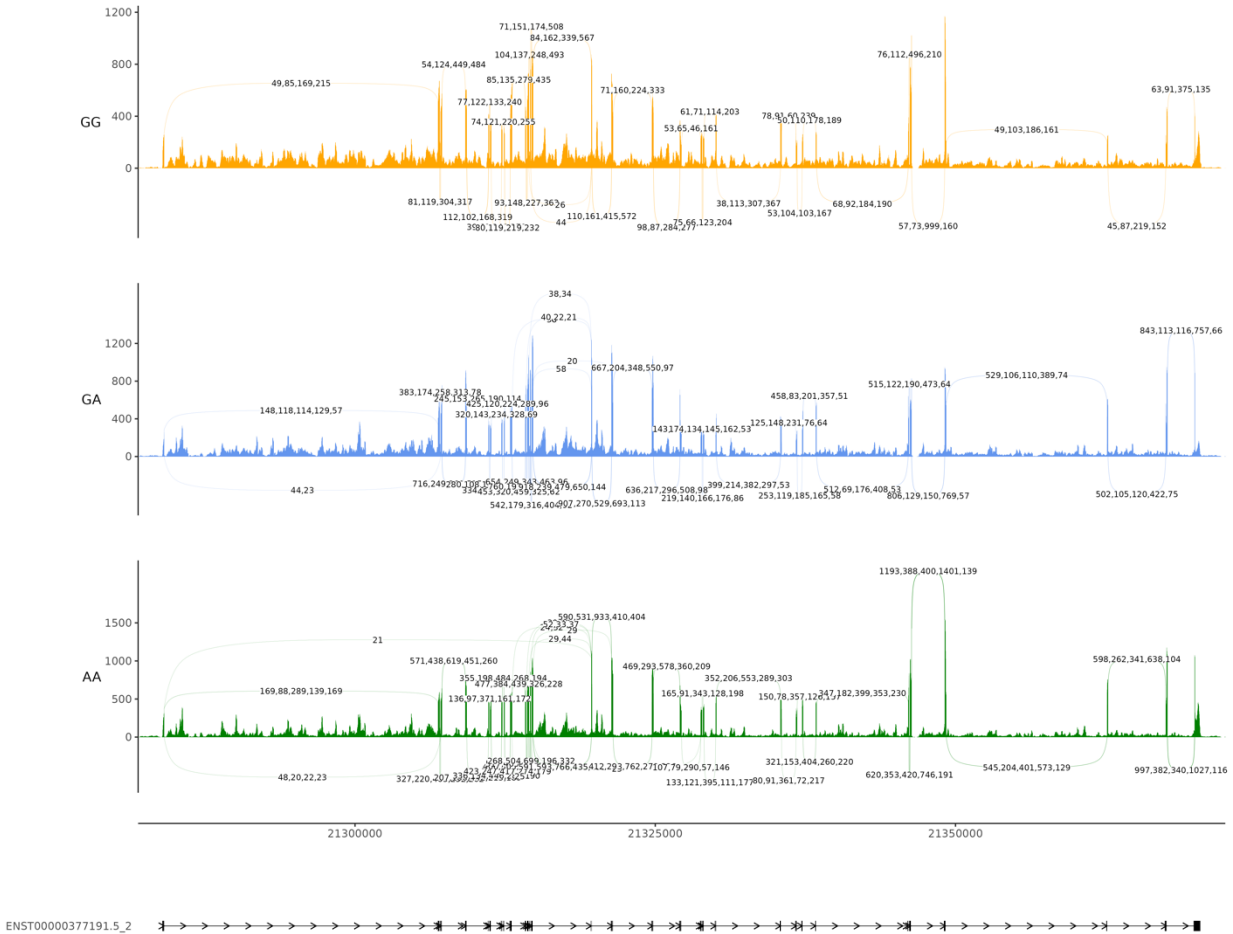

Supplementary Figure 3.18: Local splicing events in the *XRN2* gene body across rs910805 genotypes. Five RNA-seq samples (O'Brien et al. 2018) with approximately equal library (or read) depths are plotted for each rs910805 genotype. As above, the increase in the number of 1st to 3rd exon junction reads in the 5' end with respect to the A major allele was subtle. This plot was generated with ggsashimi GitHub repository (https://github.com/guigolab/ggsashimi).

XRN2 locus

Supplementary Figure 3.19: A close look at the pleiotropic *XRN2* locus. GWAS results for 56 complex phenotypes are shown, which span autoimmune, endocrine, psychiatric, cardiovascular disorders, and cancer. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *XRN2* gene. LD is calculated with individuals of European ancestry in the 1000 Genomes Project reference panel. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), RBC (red blood cell), SCZ (schizophrenia).

CHAPTER 4

Brain gene co-expression networks link complement signaling with

convergent synaptic pathology in schizophrenia

4.1 Abstract

The most significant common variant association for schizophrenia (SCZ) reflects increased expression of the complement component 4A (*C4A*). Yet, it remains unclear how *C4A* interacts with other SCZ risk genes and whether the complement system is more broadly implicated in SCZ pathogenesis. Here, we integrate several existing, large-scale genetic and transcriptomic datasets to interrogate the functional role of the complement system and *C4A* in the human brain. Surprisingly, we find no significant genetic enrichment among known complement system genes for SCZ. Conversely, brain co-expression network analyses using *C4A* as a seed gene revealed that genes down-regulated when *C4A* expression increased exhibit strong and specific genetic enrichment for SCZ risk. This convergent genomic signal reflected neuronal, synaptic processes and was sexually dimorphic and most prominent in frontal cortical brain regions. Overall, these results indicate that synaptic pathways—rather than the complement system—are the driving force conferring SCZ risk.

4.2 Introduction

SCZ is a highly heritable and disabling neurodevelopmental, psychiatric disorder that affects ~1% of the general population (Gandal et al. 2016; Sullivan et al. 2003). Despite its immense contribution to public health burden worldwide, there have been no fundamental advances in the treatment of SCZ since the 1980s, due in large part to the lack of novel, robust therapeutic targets.

The recent success of genome-wide association studies (GWAS) (Pardiñas et al. 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Visscher et al. 2017) brings hope that genetics can provide novel insights into underlying disease mechanisms and identify new biological pathways for intervention. However, the transition from GWAS to mechanistic insights is challenged by daunting levels of polygenicity and small effect sizes of associated variants (Gandal et al. 2016; Hyman 2018). One potential solution has been to incorporate GWAS results within the context of known molecular and cellular pathways, leveraging prior knowledge that genes do not act in isolation, to identify biological processes exhibiting robust genetic convergence (Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium 2015; Parikshak et al. 2013; Willsey et al. 2013).

The strongest and first-identified GWAS signal for SCZ lies in the major histocompatibility complex (MHC) region, traditionally known for its role in immunity. This association was subsequently shown to reflect in part complex genetic variation within the *C4* locus (Sekar et al. 2016), where human *C4* is encoded by two genes—*C4A* and *C4B*—which exist in different combinations of copy numbers, commonly ranging from zero to four copies of each gene per individual. Previous work demonstrated that such multiallelic copy number variation (mCNV) of *C4* influences gene expression and that elevated expression of *C4A*, but not *C4B*, confers SCZ risk (Sekar et al. 2016). *C4A* encodes an early component of the classical complement pathway, a part of the innate immune system that serves to clear cellular debris and provide the first line of antimicrobial defense. The strength and novelty of this association has prompted speculation that *C4A*—and the complement system more broadly—may represent a novel therapeutic target for SCZ. However, apart from *C4A*, surprisingly little is known about the broader relevance of the

complement system in SCZ pathogenesis. Furthermore, it remains unclear whether *C4A* interacts with other established risk factors.

Within the brain, the complement system plays a distinct, non-inflammatory role as a mediator of synaptic pruning (Coulthard et al. 2018; Sekar et al. 2016; Stephan et al. 2012), where it tags synapses for microglia-dependent elimination. Intriguingly, excessive pruning has long been hypothesized in SCZ (Feinberg 1982; Glantz and Lewis 2000; Keshavan et al. 1994) and thought to reflect reduced cortical thickness (van Erp et al. 2018) as well as dendritic spine abnormalities (MacDonald et al. 2017) observed in SCZ cases. However, these links have yet to be proven or tied to a concrete genetic mechanism. Complicating matters, the lack of evolutionary conservation of *C4A* has hindered direct investigation in model organisms. Whereas human stem cell-based assays have been used to study aspects of synapse elimination relevant to SCZ (Sellgren et al. 2019), these systems fail to recapitulate the complete range of neuronal-glial interactions present in the human brain, nor have they been shown to reach postnatal levels of maturity (Stein et al. 2014) when pruning largely occurs. As such, we reasoned that direct assessment in the human brain is an important first step to elucidate the specific molecular processes through which *C4A* increases risk for disease.

In this study, we integrated large-scale genetic and brain transcriptomic datasets from PsychENCODE (Gandal et al. 2018b; Wang et al. 2018) and GTEx (GTEx Consortium et al. 2017) to interrogate the functional role of *C4A* in the human brain and its relation to other SCZ risk factors. We used gene co-expression networks to capture coherent biological processes that covary across samples (Parikshak et al. 2015) and hence provide an unbiased functional annotation for

C4A. We took a seeded approach, identifying genes whose expression is either positively or negatively correlated with *C4A* expression. Genes positively correlated with *C4A* captured the known complement components as well as astrocyte, microglial, and NFkB signaling pathways, but they showed no genetic enrichment for SCZ. In contrast, genes negatively correlated with *C4A* reflected neuronal and synaptic pathways and exhibited strong convergent enrichment for SCZ genetic risk. Altogether, these results highlight the human brain-specific function of *C4A* and provide evidence for complex interplay between *C4A* and synaptic processes to confer SCZ risk.

4.3 Results

4.3.1 Limited evidence for SCZ genetic association within the known complement system

We first sought to determine whether genetic evidence supported SCZ association for any of the 57 genes annotated within the complement system (Methods). As GWAS loci are difficult to definitively map to causal genes, we assessed several lines of evidence supporting a putative association (**Figure 4.1a**). We first evaluated the proximity of these genes to SCZ GWAS loci (Pardiñas et al. 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). Outside the MHC region, nine genes were within 1 Mb of genome-wide significant loci (**Figure 4.1a**). Of these, three were not considered brain-expressed in PsychENCODE (Gandal et al. 2018b), and several were within the same genomic region. Three genes—*CD46*, *CSMD1*, and *CLU*—were the closest gene to their respective index single-nucleotide polymorphism (SNP). *CSMD1* and *CD46* had support from Hi-C interactions in fetal and adult brain (Mah and Won, 2019), and *CLU* and *CD46* had additional support from summary-data-based Mendelian randomization (SMR) (Zhu et al. 2016) at FDR < 0.05 and $P_{\text{HEIDI}} > 0.05$ (**Figure 4.1a**). Altogether,

these findings provided a moderate level of evidence supporting SCZ association for up to four genes within the complement system.

To determine whether this putative association of four complement system genes is greater than expected by chance, and to test whether the complement system as a whole is broadly enriched for SCZ GWAS signals, we used stratified LD score regression (sLDSC) (Finucane et al. 2018). We found no significant enrichment of SNP-based heritability in SCZ, despite testing a range of window sizes around each gene (**Figure 4.1b**). A similar lack of enrichment was found using a second method, MAGMA (de Leeuw et al. 2015) (**Figure 4.1c**). To account for the small number of genes in this pathway, we further expanded the annotation to include high-confidence protein-protein interactions (PPIs) (Li et al. 2017) for the complement system and still observed no significant enrichment. Finally, we tested whether any of these gene sets were enriched for genes implicated in SCZ through rare variant association studies, again finding no evidence of enrichment (**Figure 4.1d**). These included genes within the eight recurrent CNV regions associated with SCZ (Marshall et al. 2017) and genes harboring an excess of rare, likely gene-disrupting (LGD) variants in SCZ probands (Genovese et al. 2016; Singh et al. 2022). Together, these results do not support broad genetic association for SCZ within the complement system.
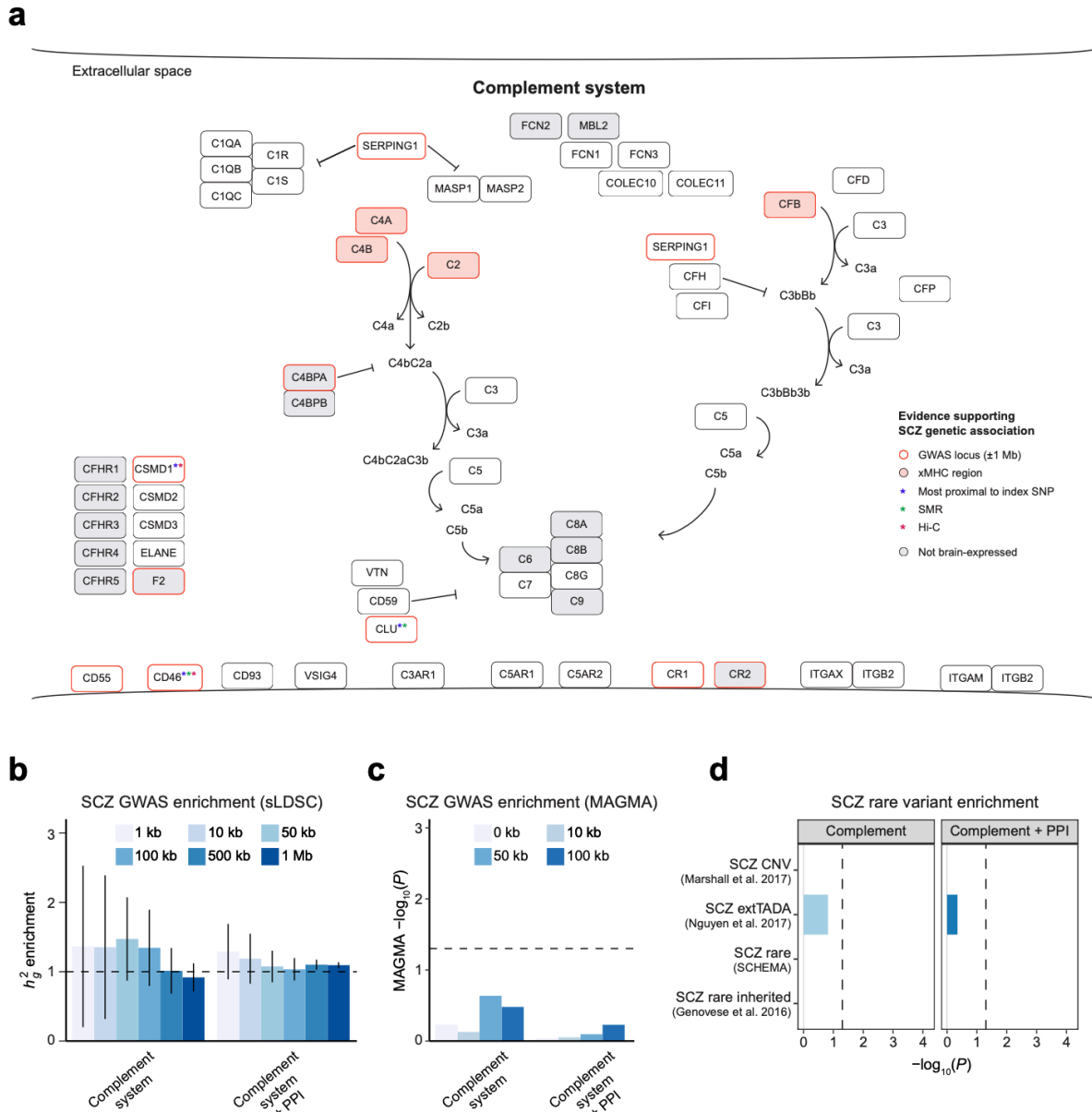
Figure 4.1: Limited evidence for broad genetic enrichment within the complement system. a, The complement system is composed of 57 genes which function together in a cascade to clear cellular debris, opsonize microbes, and mediate synaptic pruning. Here, we plot genes annotated within the complement system and corresponding evidence for SCZ genetic association, based on proximity to GWAS loci, support from SMR (summary-data-based Mendelian randomization), and Hi-C interactions in fetal and adult brain. No enrichment of SCZ GWAS signals was observed for the complement system or an expanded annotation including high-confidence PPIs (InWeb3; n = 545 genes), using **b**, sLDSC or **c**, MAGMA with varying window sizes around each gene. All error bars denote standard errors of estimates of heritability enrichment, where the enrichment is defined as the proportion of SNP-based heritability over proportion of SNPs. Dashed lines denote enrichment of one and significance at $P < 0.05$. **d**, The complement system did not show enrichment for SCZ risk genes from rare variant studies.
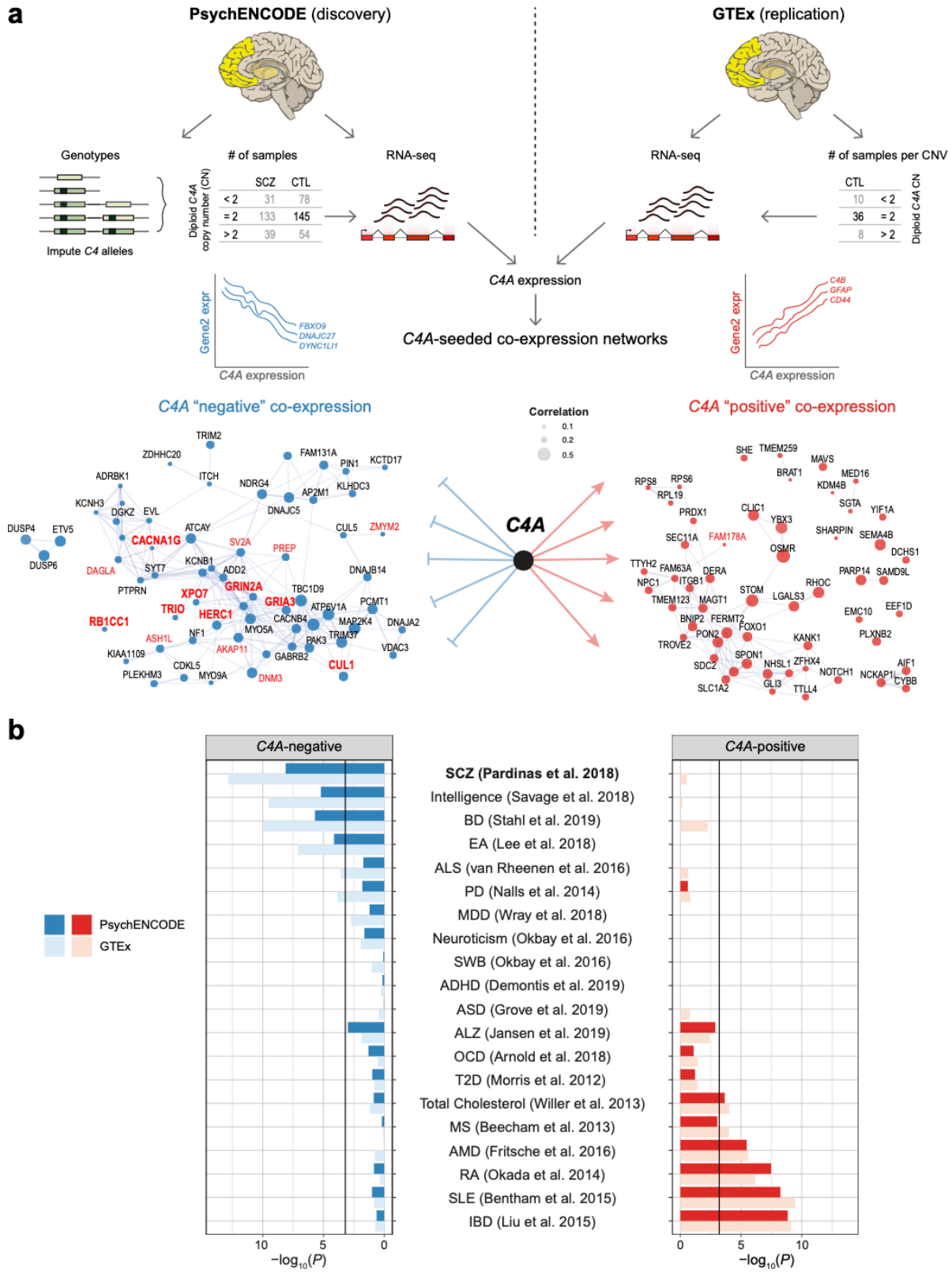
Figure 4.2: *C4A*-seeded co-expression networks capture convergent genetic risk for SCZ. a, Overview of the generation of *C4A*-seeded networks, using control samples from PsychENCODE and GTEx. Node size is proportional

to |correlation| with *C4A* expression and edges represent gene-gene co-expression. Shown in red labels are SCZ risk genes from SCHEMA reaching FDR or exome-wide (bold) significance. **b**, *C4A*-positive and *C4A*-negative genes showed enrichment for distinct GWAS signals, where *C4A*-negative, but not *C4A*-positive, genes showed enrichment for SNP-based heritability in SCZ. Results replicated in the independent GTEx dataset. The black line denotes Bonferroni-adjusted *P* value at 0.05/80. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), ALZ (Alzheimer disease), AMD (age-related macular degeneration), ASD (autism spectrum disorder), BD (bipolar disorder), EA (educational attainment), IBD (inflammatory bowel disease), MDD (major depressive disorder), MS (multiple sclerosis), OCD (obsessive-compulsive disorder), PD (Parkinson's disease), RA (rheumatoid arthritis), SLE (systemic lupus erythematosus), SWB (subjective well-being), T2D (type 2 diabetes).

4.3.2   Seeded co-expression networks provide brain-specific functional annotation for *C4A*

The previous analyses relied on known gene set annotations which are often incomplete, especially for biological processes occurring in the human brain (Koopmans et al. 2019). Additionally, the non-inflammatory role of *C4A*—and the complement system—as an effector of synaptic pruning may not be fully reflected in these annotations. To address this, we turned to gene co-expression network analyses, which can provide an orthogonal, unbiased functional annotation based on correlated gene expression patterns across samples (Parikshak et al. 2015). Here, we took a 'seeded' approach, identifying genes either positively or negatively correlated with *C4A* expression and using such 'guilt-by-association' to draw biological inference.

We first constructed a *C4A*-seeded co-expression network from frontal cortex samples of non-psychiatric controls in PsychENCODE (Gandal et al. 2018b; Wang et al. 2018) (**Figure 4.2a**). To mitigate the potential influence of germline mCNV, we imputed *C4* structural alleles from nearby SNP genotypes (Sekar et al. 2016) in individuals of European ancestry (N = 812; **Supplementary Figure 4.1**). We then selected control samples with high-quality imputation results carrying the most common diploid *C4A* copy number (CN = 2, N = 145; **Supplementary Figure 4.2**; Methods). Using these samples, we identified 3,021 genes co-expressed with *C4A* at FDR < 0.05. These included 1,869 positively co-expressed genes as well as 1,152 negatively co-expressed genes (herein referred to as "*C4A*-positive" and "*C4A*-negative" genes). As a positive control, the known

complement signaling pathway was overrepresented among *C4A*-positive genes (odds ratio (OR) = 17.2, $P < 10^{-16}$), but not *C4A*-negative genes (OR = 0, $P = 1$). In addition, *C4A*-positive genes were most strongly enriched for "immune effector process" and "response to cytokine" Gene Ontology (GO) terms (FDR's $< 10^{-41}$), whereas *C4A*-negative genes were most strongly enriched for "anterograde trans-synaptic signaling" and "chemical synaptic transmission" GO terms (FDR's $< 10^{-12}$).

For replication, we generated an analogous seeded network in the independent GTEx dataset (GTEx Consortium et al. 2017). We observed highly significant overlap among *C4A*-positive and *C4A*-negative genes across these datasets (OR's = 19 and 16, P's $< 10^{-16}$, respectively; **Supplementary Figure 4.3**). As an additional control, we generated 10,000 seeded networks using randomly sampled seed genes (Methods). The original *C4A*-positive network showed greater enrichment for the known complement components than 98% of all other networks generated in this manner (**Supplementary Figure 4.4a**).

4.3.3 *C4A*-negative, but not *C4A*-positive, genes show strong SCZ genetic enrichment

We next sought to determine whether this network-based, brain-specific functional annotation for *C4A* better captured convergent genetic risk for SCZ. Consistent with our results above, we did not find enrichment of SNP-based heritability for SCZ among *C4A*-positive genes (**Figure 4.2b**). These were instead associated with autoimmune and chronic inflammatory conditions, such as inflammatory bowel disease (IBD), rheumatoid arthritis (RA), and lupus (SLE). In contrast, *C4A*-negative genes were strongly enriched for SNP-based heritability in SCZ and in several other neuropsychiatric disorders, to a lesser degree (**Figure 4.2b**). These findings were replicated in

GTEx, so we subsequently combined both networks from PsychENCODE and GTEx to yield a high-confidence seeded network (Methods). Notably, in this network, among the ten genes harboring rare loss-of-function variants in SCZ probands at exome-wide significance (Singh et al. 2022), eight were negatively co-expressed with *C4A* at FDR < 0.1 (*TRIO*, *GRIN2A*, *XPO7*, *CUL1*, *GRIA3*, *HERC1*, *RB1CC1*, and *CACNA1G*), suggesting convergence of polygenic effects across the allelic spectrum (logistic regression, FDR = 9.0×10^{-4}; **Figure 4.2a** and **Supplementary Figure 4.4b**; Methods). The remaining two genes (*SETD1A*, *SP4*) show peak expression in the fetal brain, suggesting alternative developmental mechanisms (Li et al. 2018a).

4.3.4 Network expansion with increased *C4A* copy number

C4A expression is likely influenced by both genetic and environmental factors. In PsychENCODE, we observed that ~22% of the variation in *C4A* expression can be explained by germline mCNV (**Supplementary Figure 4.5**). However, it remains unknown what effect these genetic factors have on *C4A* co-expression. To address this, we stratified all PsychENCODE samples with high-quality imputation results (N = 552) into three CNV groups based on diploid *C4A* copy number of < 2, 2, and > 2, representing a gradient of increasing genetic risk for SCZ (**Supplementary Figure 4.2**; Methods). We then generated *C4A*-seeded networks for each group, using bootstrap to match the sample size (100 samples + 10,000 iterations). Remarkably, we observed a large increase in network size as *C4A* copy number increased (**Figure 4.3a**). With increased genomic copy number, the number of both *C4A*-positive and *C4A*-negative genes was substantially larger, indicating that *C4A* is more strongly connected and likely plays more of a driver role (**Supplementary Figure 4.6**). This network expansion was preserved across a range of correlation and FDR thresholds (**Figure 4.3b**) and was not associated with technical factors such as postmortem interval (PMI) or

RIN. Furthermore, this network expansion was not observed for *C4B*-seeded networks, demonstrating the specificity of this association (**Figure 4.3c** and **Supplementary Figure 4.7**; Methods). Together, these results indicate that genotypes conferring increased risk for SCZ are associated with strong and specific remodeling of brain gene co-expression networks, providing novel means to characterize the molecular processes underlying SCZ risk.

Figure 4.3: Strong network expansion with increased *C4A* copy number. a, *C4A*-seeded co-expression networks were generated following stratification of the PsychENCODE dataset by imputed *C4A* copy number. A substantial network expansion was observed with increased *C4A* copy number. Each network was generated via bootstrap (100 samples, 10,000 iterations) for robustness. Edges represent Pearson's correlation coefficient (PCC) > 0.5 and edge weights represent the strength of the correlation. Probable SCZ risk genes implicated by common or rare variant studies are highlighted in bold. **b**, *C4A*-seeded networks expanded in size regardless of the applied PCC threshold. **c**, The nonlinear network expansion was specific to *C4A* as a seed gene, and not observed for *C4B*. Two genes, *GRIA3* and *MVP*, are shown to illustrate this specificity. Shown are fitted linear models with 95% confidence bands.

4.3.5 Seeded networks capture *C4A*-associated pathways and cell-types

We then sought to understand the biological pathways and cell-types captured by these *C4A*-seeded networks. As above, *C4A*-positive and *C4A*-negative genes were enriched for distinct GO terms: *C4A*-positive genes for inflammatory pathways and *C4A*-negative genes for synapse-related pathways (**Supplementary Figure 4.8**). Overlap of these genes with a set of previously characterized brain co-expression modules (Gandal et al. 2018b) confirmed their broad relationship to inflammatory and synaptic function, respectively (**Figure 4.4a** and **Supplementary Figure 4.9**).

Notably, *C4A*-positive genes were strongly enriched for co-expression modules previously shown to represent astrocyte, microglial, and NFkB signaling pathway genes. These included several canonical markers of astrocytes (e.g. *GFAP*, *AQP4*) and microglia (e.g. *FCGR3A*, *TYROBP*); critical components of the NFkB signaling pathway (e.g. *NFKB2*, *IL4R*, *RELA*); as well as known members of the classical complement pathway (e.g. *C1R*, *C1S*). Conversely, *C4A*-negative genes showed enrichment for multiple neuronal and synaptic processes, stronger at higher copy number (**Figure 4.4a** and **Supplementary Figure 4.9**). These included several glutamate receptors (e.g. *GRIN2A*, *GRM1*, *GRIA3*), calcium regulators (e.g. *CAMK4*, *CAMTA1*, *CAMKK2*), and potassium channels (e.g. *KCNK1*, *KCNQ5*, *KCNIP3*). Other notable *C4A*-negative genes included the serotonin receptor *HTR2A*, the dopamine receptor *DRD1*, the major neuronal splicing regulator *NOVA1*, and the zinc transporter *SLC39A10*. These *C4A*-positive and *C4A*-negative genes were also strongly enriched for genes up- and down-regulated in SCZ brain (Collado-Torres et al. 2019; Gandal et al. 2018b), respectively (**Figure 4.4b**), further connecting *C4A* expression to dysregulated molecular pathways in SCZ brain.
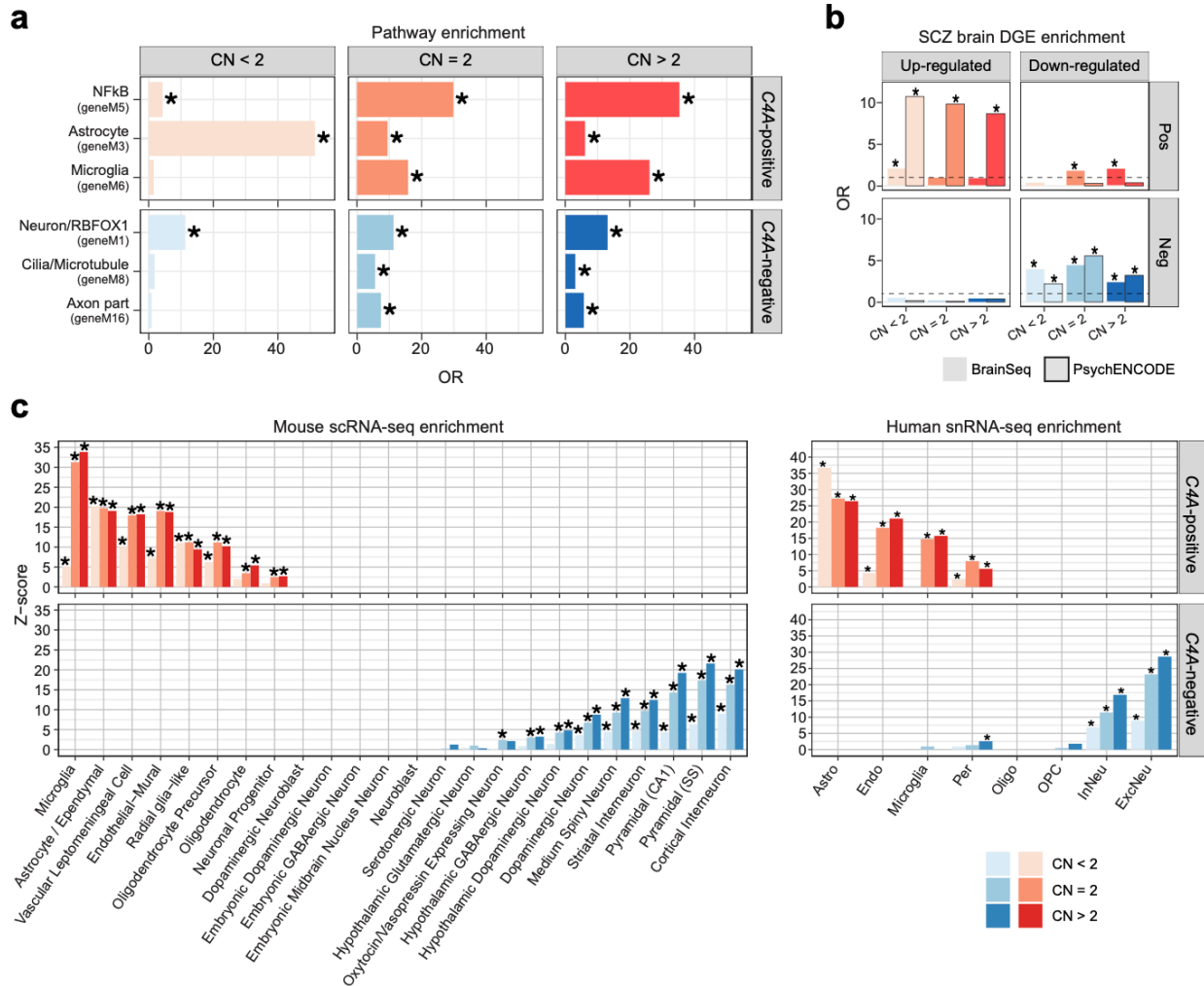
Figure 4.4: *C4A*-seeded co-expression networks identify transcriptional correlates of synaptic pruning. a, The top *C4A*-positive and *C4A*-negative genes showed distinct enrichments for neurobiological pathways and cell-types. With increasing *C4A* copy number, *C4A*-positive genes showed greater enrichment for microglia and NFkB pathways, while *C4A*-negative genes showed greater enrichment for neuron- and synapse-related modules. OR = odds ratio from two-sided Fisher's exact test. Asterisks denote significance at Bonferroni-corrected $P < 0.05$. **b**, *C4A*-positive and *C4A*-negative genes were enriched for differentially expressed genes in SCZ brain from PsychENCODE and LIBD BrainSeq (Collado-Torres et al. 2019). Asterisks denote significance from Fisher's exact test at nominal $P < 0.05$. **c**, *C4A*-positive and *C4A*-negative genes were expressed in distinct cell-types. Expression-weighted cell-type enrichment (EWCE) was performed using mouse cortical/subcortical single-cell RNA-seq data (Skene et al. 2018) and human cortical single-nucleus RNA-seq data (Wang et al. 2018). Asterisks denote significance at FDR < 0.05. *C4A*-positive and *C4A*-negative genes are shown in red and blue, respectively.

To further refine the cell-types associated with these networks, we evaluated whether *C4A*-positive

and *C4A*-negative genes were expressed in specific cell-types defined by single-cell/nucleus RNA-

seq (Skene and Grant 2016). At low copy number (i.e. CN < 2), *C4A*-positive genes showed the

strongest association in astrocytes, but with subsequently higher copy number, they became more broadly associated with microglia and endothelial cells (**Figure 4.4c**). In contrast, *C4A*-negative genes were most highly expressed in five neuronal cell-types—cortical interneurons, pyramidal (hippocampus CA1), pyramidal (somatosensory cortex), medium spiny neurons, and striatal interneurons. Remarkably, these cell-types have all been previously shown to be enriched for SCZ GWAS signals (Skene et al. 2018) (**Figure 4.4c** and **Supplementary Figure 4.10**). These findings were replicated across multiple other single-cell/nucleus RNA-seq datasets from either mouse or human brain (**Figure 4.4c** and **Supplementary Figure 4.11**). Taken together, these results indicate that higher *C4A* copy number is associated with brain co-expression changes leading to down-regulation of neuronal, synaptic genes—a putative transcriptomic signature of synaptic pruning.

4.3.6 Sexual dimorphism of *C4A* effects in the human brain

SCZ is more prevalent in males compared with females, and recent work has identified larger effect sizes of *C4* alleles in males compared with females (Kamitaki et al. 2020). Although no sex differences in *C4A* expression were reported in GTEx, protein levels of *C3* and *C4* were elevated in cerebrospinal fluid (CSF) from males (Kamitaki et al. 2020). Here, in the independent PsychENCODE dataset, we replicated these findings, finding no sex differences in *C4A* expression in the brain (**Figure 4.5a**). Notably, however, we observed a significant increase in *C4A* network size in males, consistent with larger effects in males (**Figure 4.5b**; Methods). Females showed a reduction in the number of both *C4A*-positive and *C4A*-negative genes, indicating broad sex-specific effects (**Supplementary Figure 4.12a**).
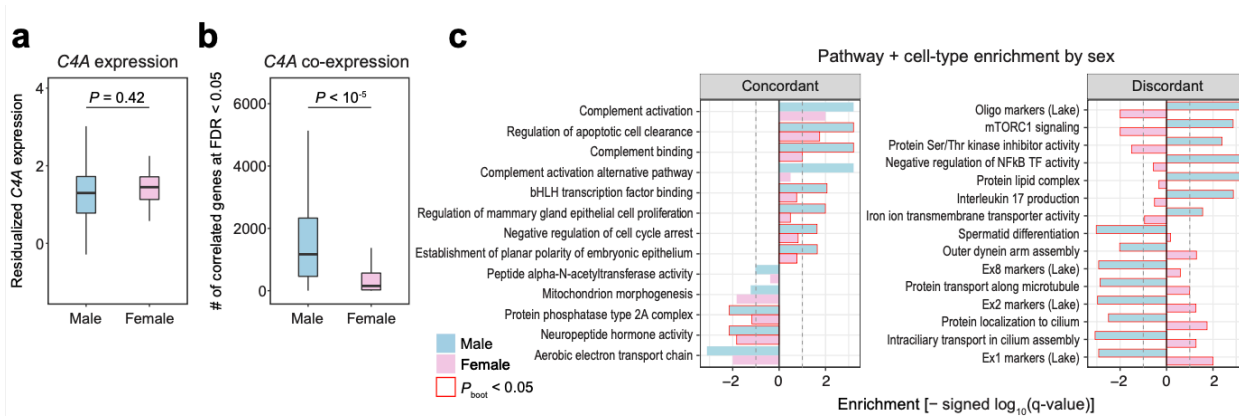
Figure 4.5: Sex differences in *C4A* co-expression highlight male-accentuated effects on mTOR signaling and neuronal cilia. a, Overall expression levels of *C4A* did not differ between sexes in PsychENCODE (N = 98 and 37 for male and female samples, respectively; two-sided Welch's t-test, $P = 0.42$). **b**, Conversely, C*4A* co-expression network size was much larger in males (N = 98, 37 for males and females; permutation test, $P < 10^{-5}$). Bootstrapped distributions were generated to match for sample size between sexes. **c**, To identify biological pathways and cell-types reflected by these sex-specific *C4A* co-expression patterns, we performed gene set enrichment analysis (GSEA). Genes were ranked by their *C4A* co-expression magnitude in male and female networks separately, and resulting enrichments were compared. Left, sex-concordant terms included positively associated complement activation. Right, sex-discordant terms included lipid and mTOR signaling genes as well as excitatory neuron markers and cilia-related pathways. Enrichment differences that were significant when compared to a null distribution of 10,000 random seed genes are highlighted in red. All boxplots show median and interquartile range (IQR) with whiskers denoting 1.5 × IQR.

To more systematically interrogate the neurobiological mechanisms contributing to these sexually dimorphic effects, we next sought to identify the specific pathways and cell-types that were differentially co-expressed with *C4A* across sexes. To do so, we ranked genes by the magnitude of *C4A* co-expression in males and females separately, performed gene set enrichment analysis (GSEA) on this ranked list, and compared the resulting enrichments (Methods). To ensure the robustness of these results, we further generated an empirical null distribution of enrichment differences between males and females with 10,000 randomly sampled seed genes (**Supplementary Figure 4.13**; Methods). As a positive control, complement-related pathways showed concordant enrichment among *C4A*-positive genes across both sexes (**Figure 4.5c**). In contrast, several pathways and cell-types showed significantly discordant effects across sexes. In males, *C4A*-positive genes were strongly associated with lipid and mTOR signaling genes, while

these enrichments were absent in females or even showed the opposite direction of effect. Likewise, strong sex-discordant effects were observed for upper layer excitatory neuron markers (Lake et al. 2018) (e.g. Ex1 and Ex2) and several cilia-related pathways among *C4A*-negative genes. Together, these results suggest that heightened effects of *C4A* in males may reflect distinct activation of mTOR signaling and disruption of primary cilia-related processes in excitatory neurons.

4.3.7 Spatiotemporal profiles highlight frontal cortex-predominant *C4A* effects

Many biological processes occurring in the human brain are region-specific and developmentally regulated (Neniskyte and Gross 2017). To determine whether certain regions are more susceptible to the effects of *C4A*, we next compared *C4A* network size across eight distinct brain regions from GTEx. Remarkably, we observed large regional differences with frontal and anterior cingulate cortex exhibiting the greatest degree of *C4A* co-expression (**Figure 4.6a**; Methods). This result was robust to different threshold metrics (**Supplementary Figure 4.12b**) and was not driven by differences in expression level across brain regions (**Figure 4.6b**). These results indicate that frontal cortical regions may be particularly vulnerable to *C4A*-mediated neurobiological processes.

We next leveraged the fact that PsychENCODE contains the largest collection of uniformly processed brain samples from individuals with SCZ (N = 531) as well as neurotypical controls (N = 895) across the adult lifespan. To confer temporal resolution, we stratified samples into overlapping time windows, while controlling for *C4A* copy number, sex, and diagnosis (Methods). *C4A* co-expression reached its peak in the 50- to 80-year-old period for neurotypical controls. In comparison, a leftward age shift in co-expression peak was observed in SCZ cases (**Figure 4.6c**

and **Supplementary Figure 4.12c**). These findings are distinct from the temporal trajectory of

C4A expression, which increased monotonically with age (**Figure 4.6d**).
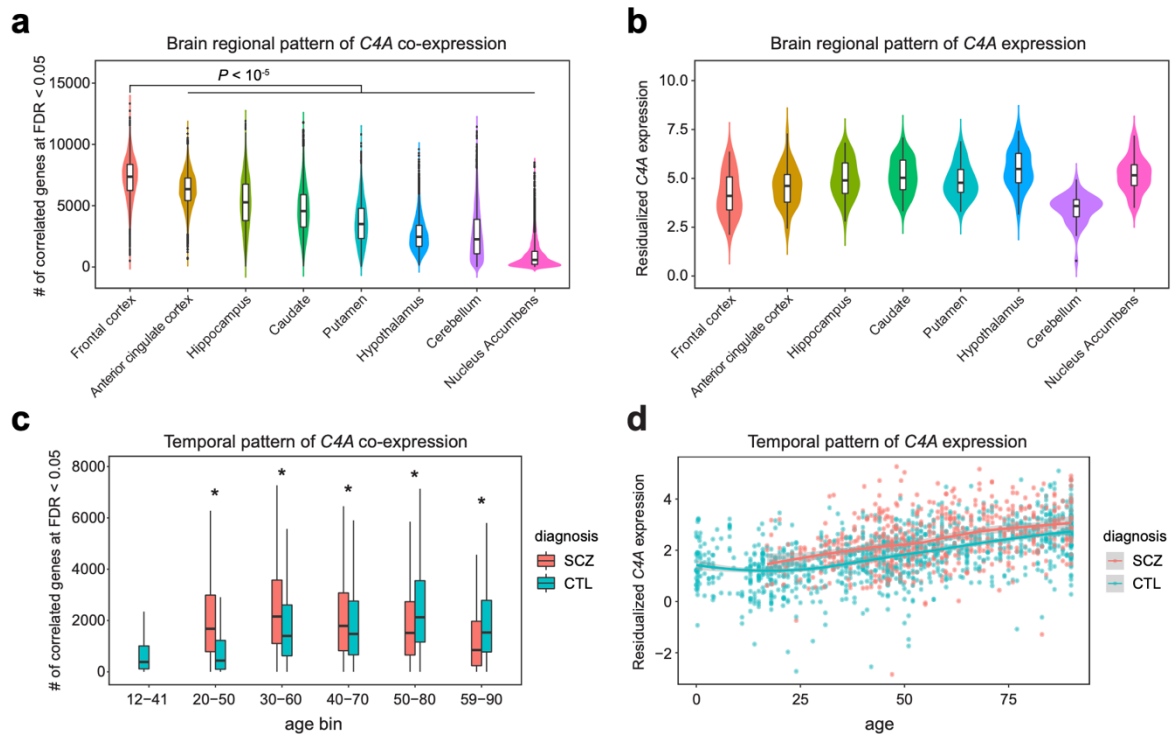
Figure 4.6: Spatiotemporal patterns of *C4A* co-expression implicate frontal cortical regions and early adult timepoints in SCZ. a, *C4A* exhibited the greatest degree of co-expression in frontal cortical brain areas. The plot shows the bootstrapped distribution of the number of co-expressed genes with *C4A* at FDR < 0.05 across eight different brain regions in GTEx (N = 36, 38, 45, 47, 39, 45, 39, and 45 for frontal cortex, anterior cingulate cortex, hippocampus, caudate, putamen, cerebellum, hypothalamus, and nucleus accumbens, respectively). All pairwise comparisons were statistically significant (permutation test, $P < 10^{-5}$). **b**, In contrast with co-expression patterns, frontal cortical regions did not show greater *C4A* expression. The plot shows *C4A* expression in GTEx samples used for the bootstrap (N = 36, 38, 45, 47, 39, 45, 39, and 45 for frontal cortex, anterior cingulate cortex, hippocampus, caudate, putamen, cerebellum, hypothalamus, and nucleus accumbens, respectively). **c**, The temporal peak of *C4A* co-expression was earlier in SCZ cases (30- to 60-year-old window) compared to controls (50- to 80-year-old window). Bootstrapped distributions were generated across overlapping time windows using samples from PsychENCODE (N = 30, 42, 57, 68, 47, and 32 for control samples in each age bin; N = 36, 46, 55, 45, and 47 for SCZ samples). Asterisks denote significant differences in the network size between SCZ cases and controls (permutation test, $P < 10^{-5}$). **d**, In contrast with co-expression patterns, *C4A* showed monotonically increasing expression across age in frontal cortex samples from PsychENCODE (N = 1730). Shown is a LOESS smooth curve with 95% confidence bands. All boxplots show median and interquartile range (IQR) with whiskers denoting 1.5 × IQR.

4.3.8 Genetic and environmental drivers of *C4A* expression alteration in SCZ brain

Finally, we sought to determine the extent to which *C4* mCNV could explain *C4A* expression

alteration in SCZ brain, using frontal cortex RNA-seq data from individuals with SCZ (N = 531)

and non-psychiatric controls (N = 895). As previously reported (Gandal et al. 2018b), we identified

strong up-regulation of *C4A* consistent with previous independent literature (Gandal et al. 2018a; Sekar et al. 2016). When we adjusted for *C4A* copy number, we continued to observe differential expression for *C4A* (**Figure 4.7a** and **Supplementary Figure 4.14**), suggesting that additional factors contribute to overexpression of *C4A* in SCZ (Sekar et al. 2016; Sellgren et al. 2019). Similar results were observed for several other complement system genes previously found (Gandal et al. 2018b) to be differentially expressed in SCZ (**Figure 4.7a** and **Supplementary Figure 4.14**).

To assess the specificity of these findings for SCZ, we performed an analogous analysis using frontal cortex data from individuals with bipolar disorder (BD; N = 217) and the same controls (**Supplementary Figure 4.14**). Despite strong genetic and transcriptomic correlations between SCZ and BD (Gandal et al. 2018a), *C4A* expression was not altered in BD, and the broader complement system exhibited minimal differential expression. This notable contrast between SCZ and BD remained when downsampling to the same number of subjects, indicating that the SCZ-BD differences were not driven by statistical power (**Supplementary Figure 4.14**). Additionally, brain samples from individuals with SCZ and BD were of similar quality with respect to PMI or RIN (Welch's t-test, $P > 0.5$) and many of the same neuroleptic medications are used to treat both conditions, indicating that these factors are unlikely to be key drivers of observed differences.

This additional component of *C4A* up-regulation in SCZ brain could be driven by other genetic factors (e.g. *trans*-eQTL) and environmental influences, or may simply represent a consequence of disease. To begin to identify potential non-genetic contributors, we turned to GTEx which has systematically compiled donor medical history. In addition to *C4A* copy number, we identified several covariates that were significantly associated with increased brain *C4A* expression—namely,

age, smoking status, and a history of liver disease (**Figure 4.7b**). This is notable given the substantially elevated rate of smoking in individuals with SCZ and some epidemiological evidence that smoking may increase risk for SCZ (Kendler et al. 2015). Altogether, these data support potential convergent effects of genetic (i.e. *C4* variation) and environmental (i.e. smoking) risk factors in disease risk.
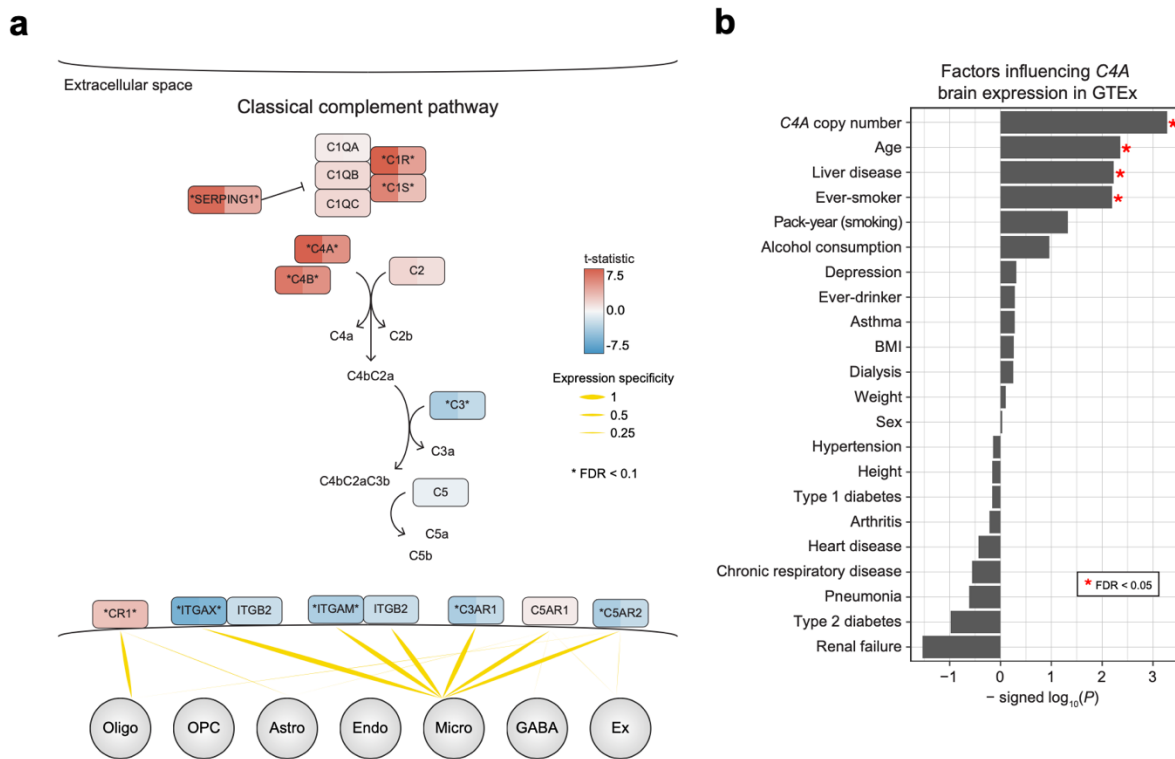
Figure 4.7: Broad, bimodal differential expression of genes within the classical complement pathway in postmortem brains from individuals with SCZ. a, Differential gene expression (DGE) in SCZ is shown for genes within the classical complement pathway. Early components were mostly up-regulated, whereas late components were down-regulated in SCZ. Genes are colored by DGE t-statistic on the left and t-statistic obtained while adjusting for *C4A* copy number on the right. Asterisks denote significance at FDR < 0.1. Bottom, cell-type specificity of complement receptors was calculated using snRNA-seq data from Hodge et al. 2019. Oligo (oligodendrocyte), OPC (oligodendrocyte progenitor cell), Astro (astrocyte), Endo (endothelial), Micro (microglia), GABA (interneuron), Ex (excitatory neuron). **b**, In GTEx, we characterized the effect of documented medical comorbidities and other relevant biological covariates on brain *C4A* expression. In addition to *C4A* copy number, age, smoking, and a history of liver disease showed significant positive associations (likelihood ratio test).

4.4 Discussion

In this study, we integrated multiple existing, large-scale genetic and transcriptomic datasets to interrogate the functional role of *C4A*—and the complement system more broadly—in the human brain and their relation to underlying core pathophysiology of SCZ. We find no evidence that the known complement system and its protein interactors are enriched for SCZ genetic signals. Using *C4A*-seeded co-expression networks, we again find that genes positively co-expressed with *C4A* show no appreciable enrichment for SCZ risk, whereas genes negatively co-expressed with *C4A* exhibit strong and specific enrichment for SCZ risk, identifying for the first time, a convergent genomic signal. These *C4A*-positive genes were associated with glial and inflammatory pathways, while *C4A*-negative genes were associated with neuronal and synaptic pathways, which is consistent with their interpretation as putative molecular correlates of synaptic pruning (Coulthard et al. 2018; MacDonald et al. 2017; Sekar et al. 2016; Sellgren et al. 2019; Stephan et al. 2012). Additionally, the seeded networks expanded in size with increased genomic copy number and exhibited sexual dimorphism and spatiotemporal specificity, suggesting potential vulnerability of the adult male frontal cortex to the effects of *C4A*. Overall, these results highlight convergence of SCZ polygenic effects and indicate that synaptic processes—rather than the complement system—are the driving force conferring SCZ risk (**Figure 4.8**).

We first observed that SCZ genetic risk is not enriched among complement system genes—despite testing multiple classes of genetic variation (i.e. GWAS, rare variants, large recurrent CNVs), using multiple statistical methods with varying genomic window sizes, and expanding the annotation to include high-confidence PPIs or *C4A*-positive genes. This was surprising given the integral role of *C4A* in the complement system (Stephan et al. 2012), the strength of the *C4A*

102
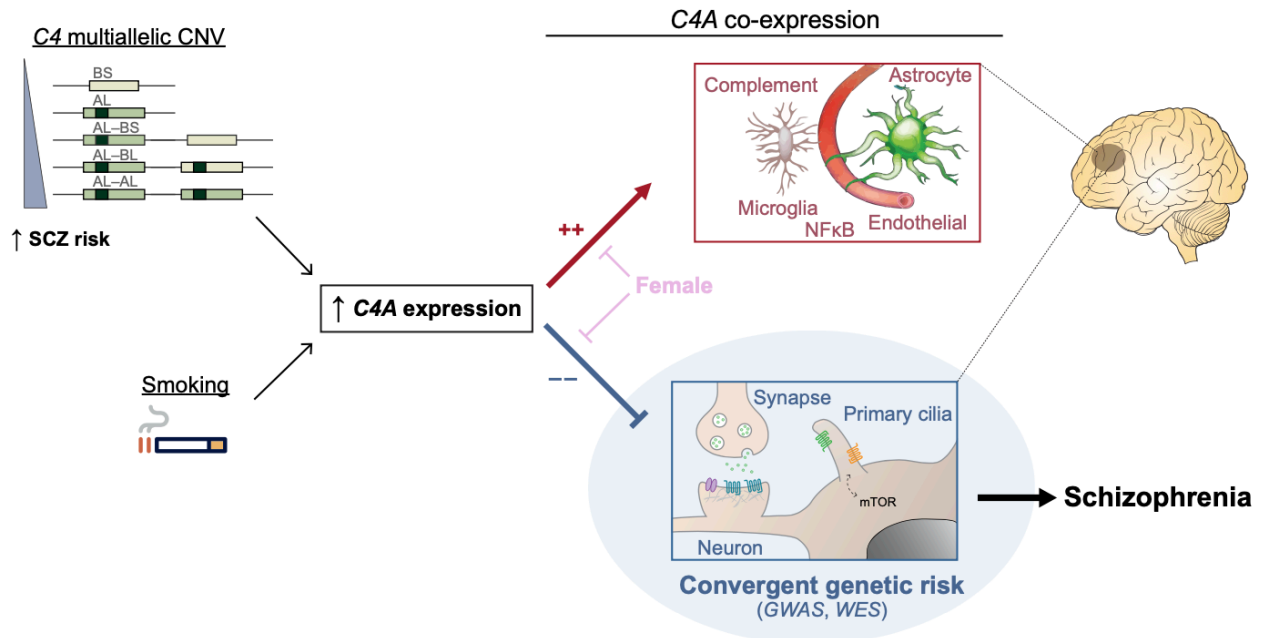
Figure 4.8: A model of the functional role of *C4A* in SCZ pathogenesis. mCNV of *C4* genes as well non-genetic factors such as smoking influence *C4A* expression. *C4A* expression is positively associated with glial and inflammatory processes and negatively associated with neuronal and synaptic processes, which in turn are enriched for SCZ genetic signals. Shown are the five most common *C4* structure alleles segregating in the general population (Sekar et al. 2016; Kamitaki et al. 2020): BS, haplotype carrying the short form of *C4B*; AL, haplotype carrying the long form of *C4A*; AL-BS, haplotype carrying one copy of the long form of *C4A* and one copy of the short form of *C4B*; AL-BL, haplotype carrying one copy of the long form of *C4A* and one copy of the long form of *C4B*; AL-AL, haplotype carrying two copies of the long form of *C4A*.

association (Sekar et al. 2016), and the high level of polygenicity observed in SCZ (Pardiñas et al. 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). This does, however, comport with recent East Asian SCZ GWAS results (Lam et al. 2019), which did not observe an MHC association (**Supplementary Figure 4.15**), despite a genetic correlation of 0.98 with European GWAS results. These findings imply that dysregulation of the complement system is neither necessary nor sufficient for the development of SCZ and fit with an alternative explanation that *C4A* may be more associated with the progression or severity of illness. Moreover, the logical extension of these observations predicts that drugs targeting this pathway are unlikely to be a panacea.

How then does *C4A* impart risk for SCZ? We reasoned that the functional role of *C4A* in the human brain may not be well captured by manually curated gene sets and pathway annotations, which are often incomplete. To address this, we leveraged co-expression networks and subsequent guilt-by-association to generate an unbiased, human brain-relevant functional annotation for *C4A*. As expected, *C4A*-positive genes capture the known complement system and reflect inflammatory processes, including astrocyte, microglial, and NFkB signaling pathways—all of which are dysregulated in SCZ brain (Gandal et al. 2018b), but none of which show an appreciable enrichment for genetic risk. Similar changes have been observed in other neuropsychiatric disorders (Gandal et al. 2018a; Gandal et al. 2018b) and may reflect environmental influences (e.g. smoking) or represent the consequence of a more proximal (e.g. synaptic) pathology. In contrast, *C4A*-negative genes reflect dysregulated neuronal and synaptic pathways, exhibiting strong genetic enrichment for SCZ. Notably, the network size and connectivity expand substantially with increased *C4A* copy number, indicating that *C4A* plays more of a driver role with increasing genetic risk for SCZ.

We find that *C4A* CNV is strongly associated with—but does not fully explain—the observed *C4A* up-regulation in SCZ brain. Similarly, although our results suggest that *C4A*-mediated SCZ risk occurs through synaptic mechanisms rather than complement signaling, several additional complement system genes exhibit differential expression in SCZ, even when controlling for *C4A* copy number. This included up-regulation of early components (e.g. *C1R*, *C1S*), but also significant down-regulation of downstream components including known complement receptors (e.g. *ITGAM*, *ITGAX*, *C3AR1*, *C5AR2*). We hypothesize that some of these observed transcriptomic alterations reflect a compensatory response to synaptic dysfunction, as *C4A* up-

regulation has also been observed in ASD brain (Gandal et al. 2018b), despite not being considered a genetic risk factor. Additionally, we find that brain *C4A* expression is elevated with smoking. Intriguingly, smoking is associated with diffuse, dose-dependent cortical thinning (Karama et al. 2015) and there is epidemiological evidence supporting a directional effect of smoking on SCZ risk (Kendler et al. 2015), although confounding factors (e.g. cannabis use) likely also contribute (Jones et al. 2018). Overall, these results highlight a neurobiological mechanism through which genetic and environmental risk factors converge and contribute to SCZ risk.

Finally, comparison of the network size provided additional insights into the spatiotemporal and sex-specific effects of *C4A*. Males showed greater degree of *C4A* co-expression, despite comparable *C4A* expression level across sexes, which is consistent with larger effects of *C4A* alleles in males relative to females (Kamitaki et al. 2020). Compared to its female counterpart, male *C4A*-seeded network showed greater activation of lipid and mTOR signaling pathways as well as greater disruption of cilia-related processes and excitatory neuron markers (**Figure 4.8**). Both mTOR signaling and primary cilia are known to be critical regulators of neurogenesis and synaptic pruning (Foerster et al. 2017; Han et al. 2008; Tang et al. 2014; Wang et al. 2015). Primary cilia, the solitary microtubule-based structure present in most neurons, glia, and their progenitors, also serve as a major hub for signaling pathways, including mTOR, Sonic hedgehog (Shh), Wnt, autophagy, and ubiquitin-proteasome system (Park et al. 2019; Wiegering et al. 2019), several of which have intriguing links to SCZ and other neurodevelopmental disorders (Marley and von Zastrow 2012; Singh et al. 2022) that warrant further experimental investigation. Together, these findings highlight several potential mechanisms underlying greater disease vulnerability in males.

Importantly, these observations are only evident through analysis of co-expression, rather than expression patterns alone, demonstrating the importance of this approach.

We note that several important questions remain for *C4A* in relation to SCZ. Although we identify *C4A*-specific interaction with *C4A* copy number variation, *C4A* and *C4B* co-expression partners are highly similar in general, making it difficult to disambiguate the effects of *C4A* from *C4B*. Further work in characterizing the biochemical properties of C4 proteins in the human brain is necessary to fully elucidate the mechanism through which *C4A* exerts larger effects in SCZ. In addition, human cell-types that express *C4A* in either physiology or pathophysiology remain unclear, due to dropout events in single-cell/nucleus RNA-seq. Although *C4A*-positive genes at low copy number (i.e. CN < 2) show strong and selective enrichment for astrocytes, and expression specificity of *C4A* is similarly the highest in astrocytes according to various mouse single-cell RNA-seq datasets (Skene et al. 2018), this remains to be validated for humans in future studies. Spatiotemporal resolution is also relatively restricted in this study, since the scope of our analyses is inherently limited to the range of available functional genomic resources, and our use of post-mortem samples is limited to retrospective analyses which cannot directly infer causal relationships. As larger and more diverse samples spanning all SCZ-relevant regions (e.g. striatum) and developmental time points become available, spatiotemporal specificity will undoubtedly improve. Likewise, as human brain genomic panels increase in size, we anticipate additional insights to be gained from distal genetic regulators (e.g. *trans*-eQTL) of *C4A*. Lastly, model systems capable of fully recapitulating postnatal neuronal-glial interactions in the human frontal cortex will be necessary for experimental validation (Forsingdal et al. 2018).

4.5 Methods

Annotation of the complement system and its protein-protein interactions (PPIs)

We compiled a list of 57 genes annotated as part of the complement system in the HUGO Gene Nomenclature Committee (HGNC) database (genenames.org). Of these, 42 genes were found to be expressed in the PsychENCODE RNA-seq data, after filtering for genes with TPM > 0.1 in at least 25% of samples (Gandal et al. 2018b). Those missing (n = 15 genes) due to low expression included: *C6*, *C8A*, *C8B*, *C9*, *FCN2*, *MBL2*, *C4BPA*, *C4BPB*, *CFHR1*, *CFHR2*, *CFHR3*, *CFHR4*, *CFHR5*, *F2*, and *CR2*. The annotation was also expanded by including high-confidence human PPIs for the complement system with score > 0.7 from the InWeb3 database (Li et al. 2017) (n = 57 + 488 = 545 genes).

Evaluation of the complement system for common variant association

We evaluated the proximity of the complement components to genome-wide significant loci from two recent SCZ GWAS studies (Pardiñas et al. 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). Four genes (*C4A*, *C4B*, *CFB*, *C2*) were within the MHC region. Excluding the MHC, nine genes (*SERPING1*, *CLU*, *CSMD1*, *CD46*, *CD55*, *CR1*, *CR2*, *C4BPA*, and *F2*) were within 1 Mb of GWAS loci. These genes were subsequently assessed for Hi-C interactions in fetal and adult brain (Mah and Won 2019) and significance from SMR method using brain and whole blood eQTL panels from PsychENCODE (Gandal et al. 2018b) and eQTLGen (Võsa et al. 2018), respectively.

Stratified LD score regression (sLDSC)

sLDSC (Finucane et al. 2018) was used to test whether a gene set of interest is enriched for SNP-based heritability in various phenotypes (i.e. diseases and traits) (Bentham et al. 2015; Demontis et al. 2019; Fritsche et al. 2016; Grove et al. 2019; International Multiple Sclerosis Genetics Consortium (IMSGC) et al. 2013; International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS) 2018; Jansen et al. 2019; Lee et al. 2018; Liu et al. 2015; Morris et al. 2012; Nalls et al. 2014; Okada et al. 2014; Okbay et al. 2016; Pardiñas et al. 2018; van Rheenen et al. 2016; Savage et al. 2018; Stahl et al. 2019; Willer et al. 2013; Wray et al. 2018). SNPs were assigned to custom gene categories if they fell within ±100 kb of a gene in the set. For the complement system, we also tested a range of window sizes (±1 kb to 1 Mb) around each gene. These categories were then added to a full baseline model that includes 53 functional categories capturing a broad set of genomic annotations. The MHC region was excluded from all analyses by default. Enrichment was calculated as the proportion of SNP-based heritability accounted for by each category divided by the proportion of total SNPs within the category. Significance was assessed using a block jackknife procedure, followed by Bonferroni correction for the number of phenotypes tested.

MAGMA

MAGMA (v1.07b) (de Leeuw et al. 2015) was used to assess enrichment of SCZ GWAS signals among the complement system. An annotation step was first performed in which SNPs in a specified window surrounding each gene were combined, while accounting for linkage disequilibrium (LD). We tested several window sizes ranging from ±0 kb to 100 kb, and LD was calculated using the European panel of 1000 Genomes Project (1000 Genomes Project Consortium

et al. 2015). A competitive gene-level analysis was then performed using the complement annotations defined above.

Rare variant enrichment

Multiple gene sets were assessed for enrichment of rare variants identified in neurodevelopmental disorders. These included: ~100 high-confidence autism spectrum disorder (ASD) risk genes harboring rare *de novo* variants (Sanders et al. 2015; Satterstrom et al. 2020); ASD risk genes harboring rare inherited variants (Ruzzo et al. 2019); genes harboring recurrent *de novo* copy number variants associated with ASD or SCZ, as compiled in Gandal et al. 2018a; genes harboring an excess of rare exonic variants in ASD, SCZ, intellectual disability (ID), developmental delay (DD), and epilepsy as assessed through an extended version of transmission and *de novo* association test (extTADA) (Nguyen et al. 2017); syndromic and highly ranked (1 and 2) genes from SFARI Gene database; genes harboring disruptive and damaging ultra-rare variants (dURVs) in SCZ cases (Genovese et al. 2016); a list of high-confidence epilepsy risk genes compiled in Polioudakis et al. 2019; risk genes for developmental disorders harboring rare *de novo* variants (Kaplanis et al. 2020); and ten high-confidence SCZ risk genes harboring rare exonic variants as identified by the SCHEMA consortium (Singh et al. 2022). For binary gene sets, statistical enrichment analyses were performed using logistic regression, correcting for linear- and log-transformed gene and transcript lengths as well as GC content. For dURVs, a two-step procedure was used, first creating a logistic regression model for genes harboring dURVs in controls and a second model for those affected in cases and controls. The likelihood ratio test (LRT) was used to assess significance. For SCHEMA and extTADA gene sets, the -$\log_{10}$-transformed P value and

posterior-probability (PP) was used, respectively, in place of binary annotation in the above logistic regression model. All results were FDR-corrected for multiple comparisons.

The PsychENCODE brain genomic dataset

Genotype array and frontal cortex RNA-seq data from Freeze 1 and 2 of PsychENCODE were obtained from www.doi.org/10.7303/syn12080241. This consisted of uniformly processed data from six studies: BipSeq, LIBD_szControl, CMC-HBCC, CommonMind, BrainGVEX, and UCLA-ASD (see Table S1 and Fig. S33 in Wang et al. 2018). Genotype data for these individual studies were previously harmonized (Wang et al. 2018) through phasing and imputation with the Haplotype Reference Consortium (HRC) reference panel. We used post-QC RNA-seq data that were fully processed, filtered, normalized, and extensively corrected for all known biological and technical covariates except the diagnosis status (see Materials/Methods and Fig. S3 in Gandal et al. 2018b). Of note, RNA-seq reads were previously aligned to the hg19 reference genome with STAR 2.4.2a and gene-level quantifications calculated using RSEM v1.2.29. Genes were filtered to include those with TPM > 0.1 in at least 25% of samples (Gandal et al. 2018b). The same expression data were used for all downstream analyses unless otherwise stated.

Imputation of *C4* structural alleles

The *C4* locus harbors multiallelic CNV (mCNV), where human *C4* encoded by two genes (*C4A* and *C4B*) can exist in different combinations of copy numbers. The two paralogs are defined based on four amino acid residues in exon 26, which are thought to alter binding affinities for distinct molecular targets. Either paralog can also contain a human endogenous retroviral insertion (*C4-HERV*) in intron 9, which then functions as an enhancer and preferentially increases *C4A*

expression (Sekar et al. 2016). Recent work demonstrated that four common *C4* structural alleles are in linkage disequilibrium (LD) with nearby SNPs (Sekar et al. 2016) and hence can be accurately imputed from genotype array data. Accordingly, we imputed *C4* alleles in six studies from PsychENCODE separately using Beagle4.1 (Browning and Browning 2016) with a custom HapMap3 CEU reference panel as described (Sekar et al. 2016). We began with the HRC imputed genotype data and filtered for high-quality SNPs by setting the R2 > 0.3 threshold from Minimac3. We restricted imputation and subsequent downstream analyses to samples of European ancestry (N = 812) based on genetic principal component analysis with the 1000 Genomes Project reference panel (1000 Genomes Project Consortium et al. 2015) (**Supplementary Figure 4.1**). There was an overlap of individuals in BipSeq, LIBD_szControl, and CMC_HBCC studies, which used different SNP genotyping platforms (see Table S1 in Wang et al. 2018). For these duplicate samples, the concordance rate of imputation result was high (N = 181/204 individuals with matching result), indicating robust *C4* imputation. For 23 samples with discordant imputation results, we calculated the average dosage for each structural allele and inferred the most likely pair of structural alleles.

Effect of *C4* variation on gene expression

Inferred copy number of *C4* structural elements (*C4A*, *C4B*, and *C4*-HERV) based on the imputed *C4* alleles was associated with *C4A* and *C4B* RNA expression using a linear model. Both best-guess copy number and probabilistic dosage were tested for association, which yielded an analogous result. As shown previously (Gandal et al. 2018b; Sekar et al. 2016; Handsaker et al. 2015), *C4A* expression was strongly associated with *C4A* copy number (R = 0.37, $P = 2.8 \times 10^{-27}$) and *C4*-HERV copy number (R = 0.33, $P = 7.9 \times 10^{-22}$), but not with *C4B* copy number (R = -

0.03, $P = 0.39$; **Supplementary Figure 4.5**). Likewise for *C4B*, expression was associated with corresponding gene dosage (R $= 0.12$, $P = 3.8 \times 10^{-4}$), but not with *C4A* copy number (R $= -0.05$, $P = 0.15$) or *C4*-HERV copy number (R $= -0.05$, $P = 0.17$).

Construction of *C4A*-seeded networks

To ensure imputation quality and thereby draw robust biological inference, we restricted our network analyses to samples with average imputed probabilistic dosage > 0.7 (N $= 552/812$). Most studies had high probabilistic dosage, except BrainGVEX and UCLA-ASD. In the case of BrainGVEX, this was because there were many missing SNPs in the vicinity of *C4* locus. This filtering step hence removed most samples with low-quality imputation from BrainGVEX and UCLA-ASD. Neurotypical control samples with diploid *C4A* CN $= 2$ (N $= 145/552$) (**Supplementary Figure 4.2**) were then used to generate a *C4A*-seeded network by calculating pairwise PCC between *C4A* and 25,774 features, which included 16,541 protein-coding and 9,233 noncoding genes based on Gencode v19 annotations. To test whether this network is enriched for the known complement components than can be expected by chance, we randomly sampled 10,000 seed genes and generated 10,000 seeded networks. For each network, genes positively correlated with the seed gene at FDR < 0.05 were assessed for overlap with the annotated complement system (n $= 57$ genes), while genes negatively correlated with the seed gene at FDR < 0.05 were assessed for overlap with genes annotated within the SynGo database (Koopmans et al. 2019) (n $= 1,103$ genes; **Supplementary Figure 4.4**).

To capture broad genetic effects of *C4A* CNV on *C4A* co-expression, we stratified PsychENCODE samples into three CNV groups (i.e. CN < 2, CN $= 2$, and CN > 2). For control samples, there were

at least 54 samples in each group (**Supplementary Figure 4.2**). To account for uneven sample sizes, we used 10,000 bootstrapping replicates to downsample to 50 samples across each group. We calculated PCC in every iteration as above and eventually took the median PCC and its corresponding P value. Generated using only the control samples, these networks were not influenced by case-control status and disease-associated confounding factors (e.g. medication and RNA degradation effects). Additionally, the control samples were balanced in covariates such as age, RIN, postmortem interval (PMI), brain pH, and sex (**Supplementary Figure 4.16**).

To maximize sample size and hence power to detect significant co-expression, particularly for rarer *C4A* CNV groups (i.e. CN < 2 and CN > 2), we also constructed the seeded networks by using every sample that passed the above quality control (N = 552). Combining all samples irrespective of the diagnosis status led to a minimum of 109 samples in each CNV group (**Supplementary Figure 4.2**), allowing us to generate the networks with bootstrap by downsampling to 100 samples. Such all-sample networks yielded analogous results to control-only networks in terms of the network expansion with respect to *C4A* CNV, effect sizes of *C4A* co-expression, and the patterns of pathway, cell-type, and genetic enrichments (**Supplementary Figure 4.17**). Given the robustness of these network findings, we present results from all-sample networks. For visualization of the *C4A*-seeded networks, a hard-threshold of PCC > 0.5 and FDR < 0.05 was applied. All network plots were drawn using *igraph* and *ggplot2* packages in R.

The GTEx brain genomic dataset

GTEx v7 was used for external replication (GTEx Consortium et al. 2017). We downloaded the GTEx genotype data from dbGaP (accession phs000424.v7.p2) and imputed *C4* alleles in samples

of European ancestry according to genetic principal component analysis. We obtained transcript-level counts from www.gtexportal.org and derived gene-level counts using *tximport* package in R. Briefly, RNA-seq reads were aligned to the hg19 reference genome with STAR 2.4.2a and transcript-level counts quantified with RSEM v1.2.22. We started with samples and features that were used for GTEx eQTL analyses. We then dropped samples from non-brain tissues and tissues with different sample preparation (i.e. cortex and cerebellar hemisphere). We also dropped samples with a history of disease possibly affecting the brain prior to filtering for features with CPM > 0.1 in at least 25% of samples. Gene-level counts were then normalized using TMM normalization in edgeR and $\log_2$-transformed to match PsychENCODE. Each brain region was then assessed for outlier samples, defined as those with standardized sample network connectivity Z scores < -3, which were removed. These quality control steps resulted in 20,765 features based on Gencode v19 annotations and 920 samples across ten brain regions, out of which 540 samples were imputed for *C4* alleles.

We next regressed out biological and technical covariates except region and subject terms using a linear mixed model via *lme4* package in R. We entered region, age, sex, 13 seqPCs (top 13 principal components of sequencing QC metrics from RNA-SeQC), RIN, ischemic time, interval of onset to death for immediate cause, Hardy Scale, body refrigeration status as fixed effects and subject as a random intercept term. To evaluate the relationship between several non-genetic factors and *C4A* gene expression, we added 3 genetic PCs, brain pH, and a covariate of interest (e.g. BMI, height, weight, smoking status, or drinking status) as fixed effects to the above model. Significance was assessed by the likelihood ratio test (LRT) of the full model with the effect in question against the null model without the effect in question.

Due to the relatively limited sample size of GTEx (i.e. less than 10 samples for CN < 2 and CN > 2 in each brain region), we focused on samples with two *C4A* copy number in subsequent analyses. We constructed a *C4A*-seeded network using frontal cortical samples (N = 36) and combined this with the above PsychENCODE control-only network (N = 145) using the Olkin-Pratt (OP) fixed-effect meta-analytical approach as implemented in *metacor* R package.

Interaction of *C4A* copy number with *C4A* expression

The specificity of the *C4A*-seeded network expansion with respect to *C4A* CNV was evaluated statistically via multiple linear regression. We tested for an interaction term between *C4A* copy number variation and *C4A* gene expression on other gene targets transcriptome-wide (i.e. 25,774 brain-expressed genes). Given that *C4A* copy number and *C4B* copy number are negatively correlated with one another (Pearson's R = -0.41, $P = 1.3 \times 10^{-23}$), both terms were included in our regression. The model we tested was: $\text{gene}_j \sim (C4A\ \text{CN} + C4B\ \text{CN}) \times C4A\ \text{expr}$, where the subscript j refers to the expression of gene j (**Supplementary Figure 4.7**). To determine how these results compare to what would be expected by chance, we replaced *C4A* expression in the above model by a randomly selected gene and calculated the number of times the interaction term was significant. We repeated this until we had randomly sampled 10,000 genes, and the empirical *P* values for *C4A* and *C4B* expression were subsequently calculated ($P = 10^{-4}$ and 0.11, respectively).

Pathway enrichment

For pathway enrichment, we focused on genes co-expressed with *C4A* at FDR < 0.05. Enrichment for GO terms was performed using *gProfileR* v0.6.7 package in R with strong hierarchical filtering (**Supplementary Figure 4.8**). Only pathways containing less than 1,000 genes and more than 10

genes were assessed. Background was restricted to brain-expressed genes and an ordered query was used, ranking genes by correlation with *C4A*. Overlap with PsychENCODE WGCNA modules (Gandal et al. 2018b) was assessed using Fisher's exact test, followed by Bonferroni correction for multiple testing (**Supplementary Figure 4.9**). The same gene sets were finally assessed for overlap with differentially expressed genes (DEG) in SCZ brain from PsychENCODE (Gandal et al. 2018b) and LIBD BrainSeq Phase II (Collado-Torres et al. 2019). For PsychENCODE, DEG at FDR < 0.05 were tested, while for LIBD BrainSeq, DEG at FDR < 0.1 were tested.

Expression-weighted cell-type enrichment (EWCE)

We used 10,000 bootstrapping replicates for EWCE with genes co-expressed with *C4A* at various FDR thresholds (**Supplementary Figures 4.10-4.11**). Briefly, EWCE statistically evaluates whether a gene set of interest is expressed highly in a given cell-type than can be expected by chance. Z-score is estimated by the distance of the mean expression of the target gene set from the mean expression of bootstrapping replicates (Skene and Grant 2016). We downloaded pre-computed expression specificity values for several single-cell/nucleus RNA-seq data from http://www.hjerling-leffler-lab.org/data/scz_singlecell/. For independent single-nucleus RNA-seq datasets from Wang et al. 2018 and Hodge et al. 2019, we processed and computed the expression specificity metric of each gene as described (Skene and Grant 2016; Skene et al. 2018).

Sex differences in *C4A* co-expression

As there were fewer female than male samples in PsychENCODE, we combined the control samples with two *C4A* copy number in the 12- to 80-year-old period for each sex separately. The resulting samples were balanced in age (Welch's t-test, $P = 0.70$; Wilcoxon rank-sum test, $P =$

0.54). We then tested for sex differences in *C4A* co-expression using bootstrapping to match the sample size (37 samples + 10,000 iterations). To identify pathways and cell-types differentially co-expressed with *C4A* across sex, we ranked genes by the magnitude of *C4A* co-expression in male and female samples separately. This ranked list was then used for gene set enrichment analysis (GSEA) (Subramanian et al. 2005) using the *clusterProfiler* R package. The union of GO and Hallmark gene sets from the MSigDB collections (C5 + H v7.1) (Liberzon et al. 2015), gene sets from SynGO (Koopmans et al. 2019), and the human brain cell-type markers defined in Lake et al. 2018 were tested for enrichment. To assess significance of GSEA results, we randomly sampled 10,000 seed genes. For each seed gene, we calculated male and female-specific co-expression and performed GSEA as above. The difference in normalized enrichment score (NES) between sexes was used as the test statistic. The empirical *P* value for each gene set was subsequently calculated by comparing the rank of this difference for *C4A* to the empirical null distribution of the test statistic from randomly sampled seed genes (**Supplementary Figure 4.13**).

Spatial resolution of *C4A* co-expression

To ensure the robustness of co-expression results, we focused on eight brain tissues from GTEx that had at least 35 samples with two *C4A* copy number (Ballouz et al. 2015; Iancu et al. 2012). As the number of samples varied across brain regions (i.e. N = 36, 38, 45, 47, 39, 45, 39, and 45 for frontal cortex, anterior cingulate cortex, hippocampus, caudate, putamen, cerebellum, hypothalamus, and nucleus accumbens, respectively), we used 10,000 bootstrapping replicates to downsample to 36 samples. In each iteration, we calculated PCC between *C4A* and every other gene and estimated the number of significantly co-expressed genes at FDR < 0.05. Other threshold metrics were tested as well, which gave similar results (**Supplementary Figure 4.12**). We did not

control for other biological covariates such as age and sex to maximize sample size and also because they were not significantly different across brain regions (one-way ANOVA, $P = 0.99$; Fisher's exact test, $P = 0.95$).

Temporal resolution of *C4A* co-expression

As our analyses suggest that *C4A* copy number variation exhibits strong genetic effects on *C4A* co-expression, we controlled for *C4A* copy number by focusing on samples with two *C4A* copy number in PsychENCODE. In order to reduce other sources of bias such as sex and diagnosis, we only used male samples and performed separate analyses for controls and SCZ cases. We divided the samples by six overlapping time windows and calculated the number of co-expressed genes for *C4A* in each time period with bootstrap (30 samples + 10,000 iterations). Here, we note relatively limited sample size and crude time windows post-stratification of the PsychENCODE dataset in order to control for potential confounding factors.

Differential expression of the complement system

Differential gene expression of the complement was calculated using a linear mixed model via *nlme* package in R as previously reported (Gandal et al. 2018b). We repeated this analysis by randomly downsampling SCZ samples to match the sample size of BD. We additionally performed several conditional analyses by adjusting for *C4A* expression and/or *C4A* copy number (**Supplementary Figure 4.14**). As *C4* alleles were imputed in only the samples of European ancestry, a subset of PsychENCODE was used for such conditional analyses.

Statistics and reproducibility

No statistical methods were used to pre-determine sample sizes, but our study makes use of the largest publicly available genomic dataset of postmortem human brains. Even after stratifying samples by imputed *C4A* copy number, this sample size was sufficient (Ballouz et al. 2015; Iancu et al. 2012) to detect significant gene co-expression, as we observed. Randomization and blinding were not possible due to the study being retrospective and observational. Accordingly, subject-level covariates were used to account for variation in gene expression as well as to remove unwanted confounding effects. We downloaded and uniformly processed the independent data from the GTEx project for external replication of PsychENCODE findings. Overall co-expression pattern and subsequent cell-type, pathway, and genetic enrichment results were replicated. We did not attempt to replicate the network expansion findings due to the small sample size of GTEx for rare copy number variant groups. For differential expression analyses across sex and case-control status, normalized gene expression was assumed to follow normal distribution, but this was not formally tested. Effects of genetic and environmental factors on gene expression were also assessed using a linear model. Additional details for statistical analyses are provided in relevant sub-sections of the Methods.

Data availability

PsychENCODE raw genotype and RNA-seq data that support the findings of this study are available at www.doi.org/10.7303/syn12080241. Processed PsychENCODE summary-level data are available at Resource.PsychENCODE.org. GTEx genotype and RNA-seq data used for the analyses described in this manuscript were obtained from: the GTEx Portal (www.gtexportal.org) and dbGaP accession number phs000424.v7.p2.

Code availability

The code used to perform bioinformatic analyses are available at:

https://github.com/gandallab/C4A-network.

4.6 Supplementary Figures
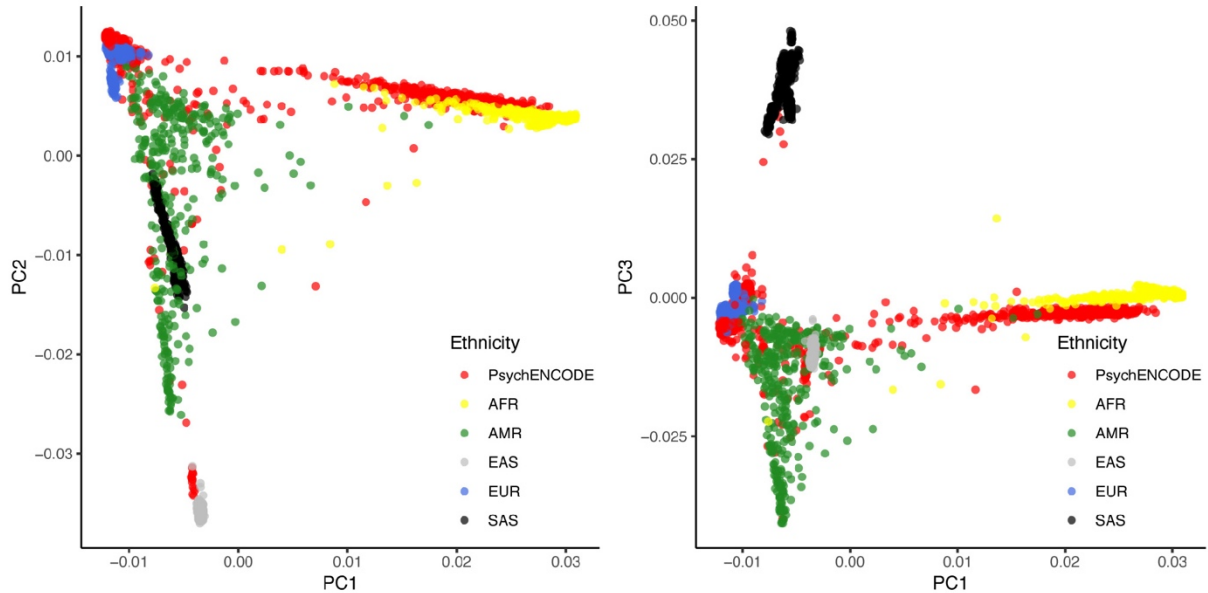
Supplementary Figure 4.1: Ancestry of PsychENCODE subjects. Principal component analysis was performed using PLINK after merging the PsychENCODE genotype data with the 1000 Genomes Project reference panel. The PsychENCODE genotype data was available for a total 1,864 subjects to begin with. Each point represents an individual and points are color-coded by corresponding ethnicity. Global ancestry was inferred by k-nearest neighbors algorithm with the first five principal components. Downstream analyses were restricted to samples of European ancestry (N = 812).

	CN < 2	CN = 2	CN > 2
Autism (ASD)	1	2	4
Bipolar Disorder (BD)	9	44	12
Schizophrenia (SCZ)	31	133	39
Control (CTL)	78	145	54
Total	119	324	109

Supplementary Figure 4.2: Number of PsychENCODE samples with high-quality *C4* imputation. Total 552 samples had average imputed probabilistic dosage > 0.7. These samples were subsequently used to generate *C4A*-seeded networks.

a

Pos

718 | 1151 | 1557

Neg

409 | 743 | 2581

■ PsychENCODE   ■ GTEx

Total # of genes :   3021   6032

b

PsychENCODE vs GTEx (R = 0.68)

GTEx *C4A* correlation

PsychENCODE *C4A* correlation
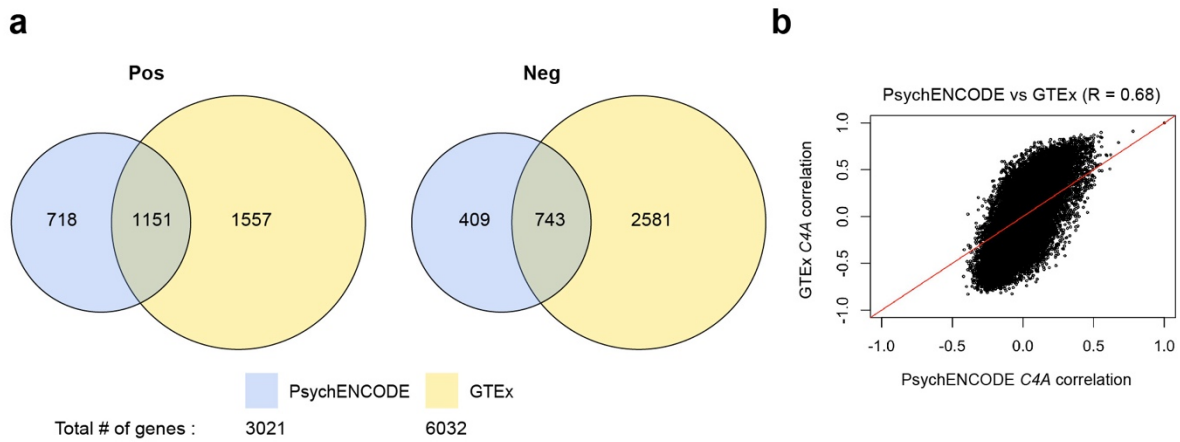
Supplementary Figure 4.3: Replication of PsychENCODE seeded network in GTEx. a, Shown are Venn diagrams of the number of overlapping *C4A*-positive and *C4A*-negative genes in PsychENCODE and GTEx (OR's = 19 and 16, P's $< 10^{-16}$, respectively). These networks were constructed from frontal cortex samples of non-psychiatric controls with *C4A* CN = 2. **b**, Shown is correlation of effect sizes (i.e. PCC) of each gene that is shared between the two networks (R = 0.68, $P < 10^{-16}$).
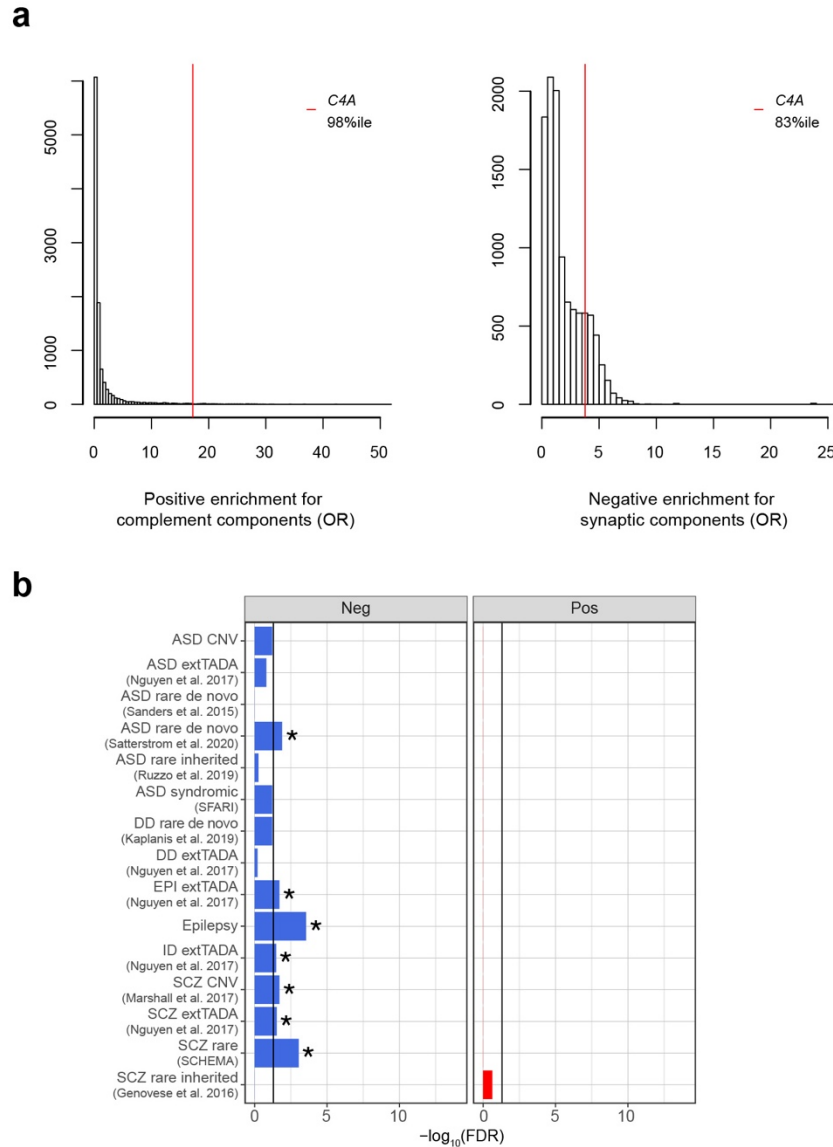
Supplementary Figure 4.4: Enrichment for complement components among *C4A*-positive genes and synaptic components as well as neurodevelopmental risk genes among *C4A*-negative genes. a, Seed genes were permuted 10,000 times and corresponding seeded networks were tested for enrichment of the complement system (n = 57 genes) and synaptic components (n = 1,103 genes) from SynGo. Shown is distribution of the odds ratio from Fisher's exact test. **b**, *C4A*-positive and *C4A*-negative genes at FDR < 0.05 from the meta-analysis of PsychENCODE and GTEx were used for rare variant analyses (logistic regression with significance assessed through likelihood ratio test). The black line denotes FDR-adjusted *P* value at 0.05.
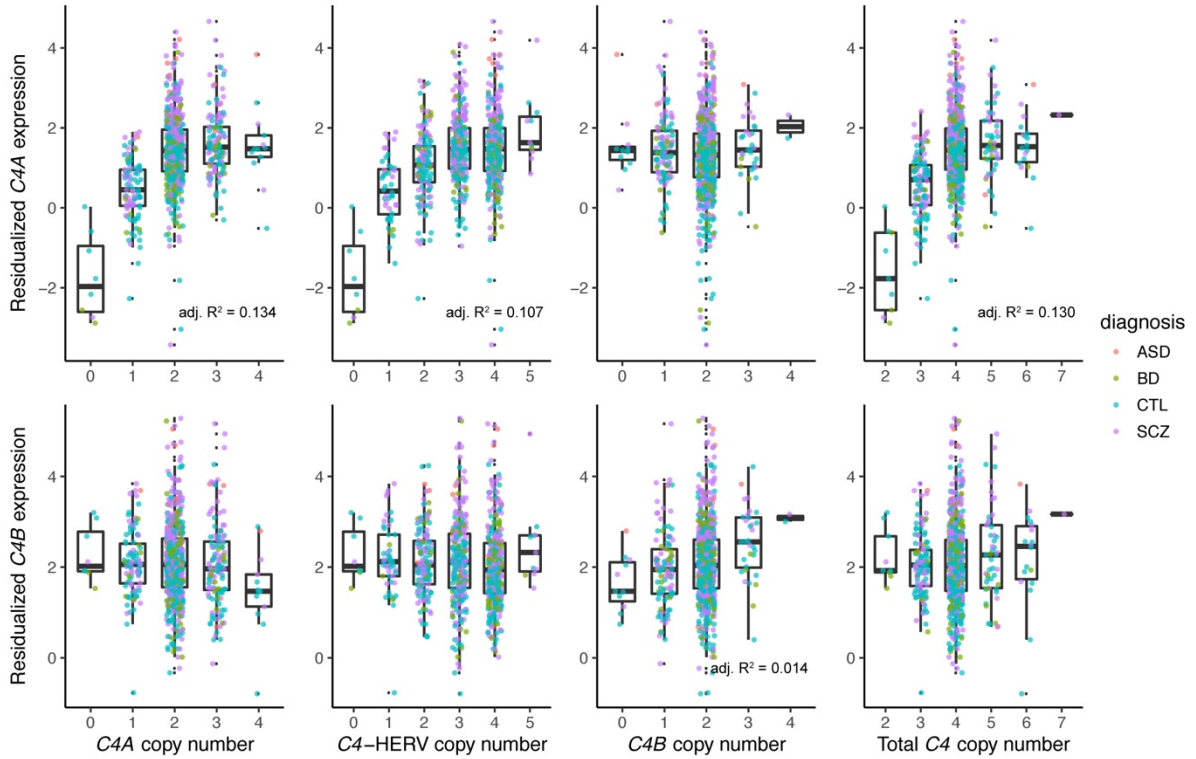
Supplementary Figure 4.5: Relationship between *C4* structural variation and *C4* gene expression. Residualized *C4* gene expression (i.e. normalized and corrected for all known biological and technical covariates except the diagnosis status) was associated strongly with corresponding gene copy number (total N = 812; N = 20, 114, 367, and 311 for ASD, BD, CTL, and SCZ samples, respectively). Adjusted R^2 values are shown for significant correlations. Of note, the best linear models for *C4A* and *C4B* expression explained up to 22% and 2.7% of variation in expression, respectively.
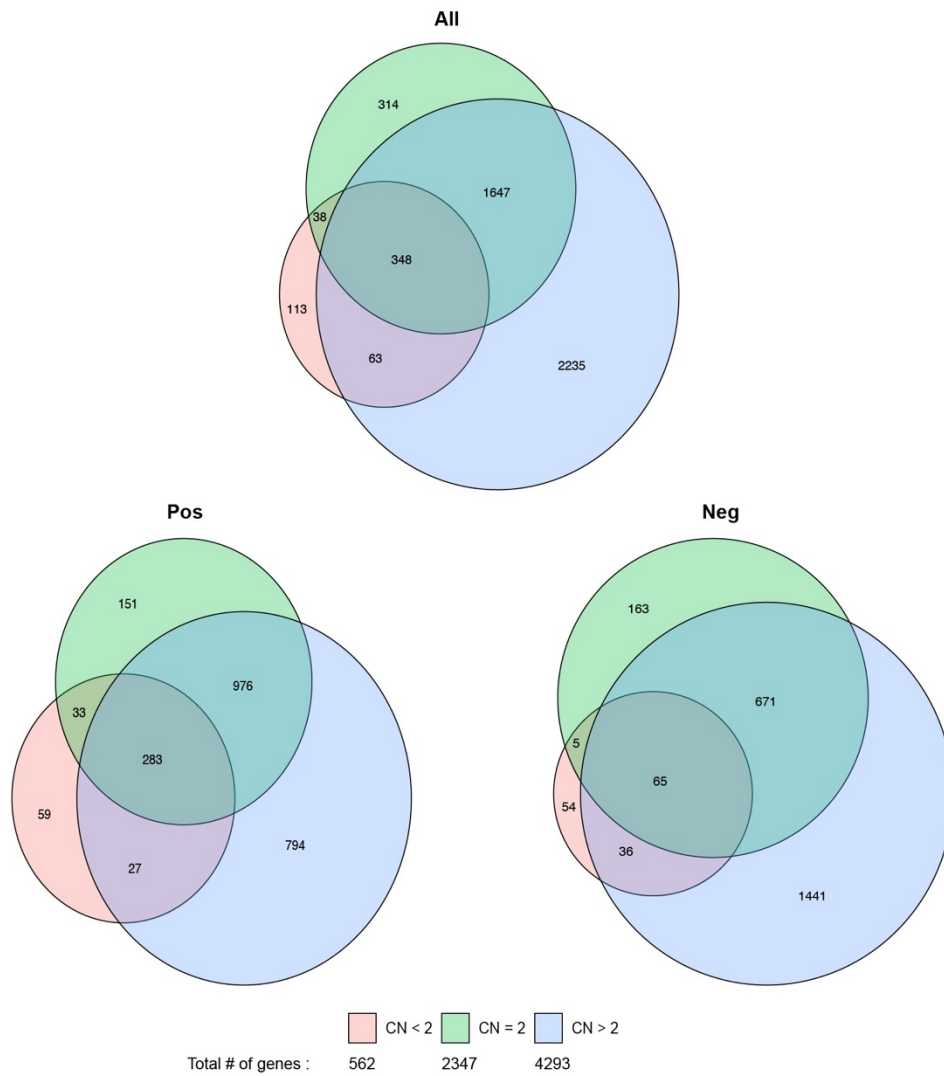
Supplementary Figure 4.6: Larger number of *C4A*-positive and *C4A*-negative genes with increased *C4A* copy number. Shown are Venn diagrams of the number of overlapping *C4A*-positive and *C4A*-negative genes across three CNV groups. Note that the sum of positive and negative genes is equal to the total number of co-expressed genes. The size of the circle is approximately proportional to the number of genes.
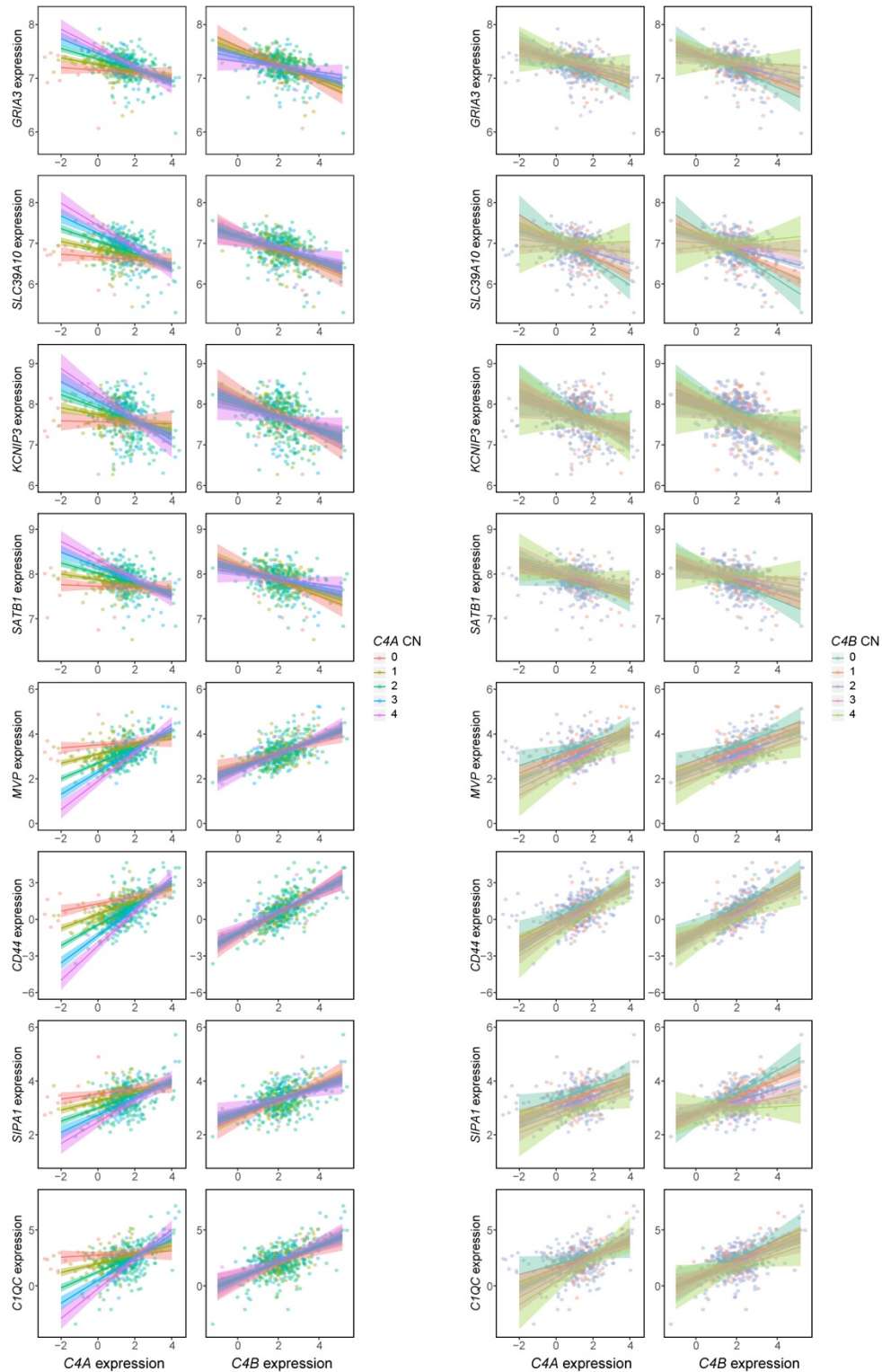
Supplementary Figure 4.7: *C4A*-specific interaction with *C4A* copy number. Multiple regression was performed with interaction terms between *C4* copy numbers and *C4* gene expression. Significant interaction effect was present only between *C4A* copy number and *C4A* expression. Several genes are highlighted to demonstrate this interaction.
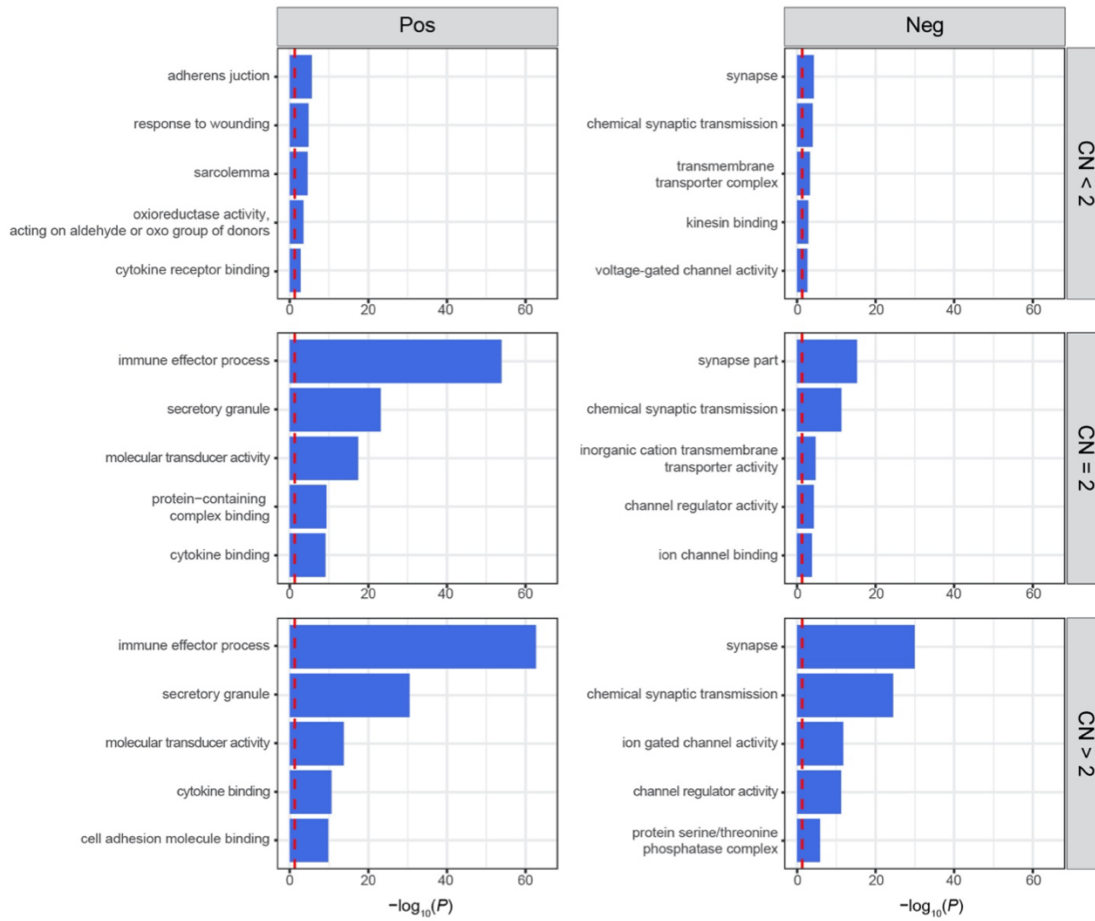
Supplementary Figure 4.8: Enrichment for distinct GO terms among *C4A*-positive and *C4A*-negative genes.
Gene sets obtained from the seeded networks at FDR < 0.05 were used for pathway enrichment analyses. Top five
GO terms with the highest enrichment are shown. The red dotted line denotes FDR-adjusted *P* value at 0.05.

	Pos			Neg		
	CN < 2	CN = 2	CN > 2	CN < 2	CN = 2	CN > 2
geneM1				11.3	11.5	9.7
geneM2						
geneM3	54.3	12.5	10.7			
geneM4						
geneM5	4.34	32.5	44.8			
geneM6		21.4	26.9			
geneM7				9.9		2.79
geneM8					4.68	3.87
geneM9				6.69	6.79	6.16
geneM10						
geneM11						
geneM12					3.42	
geneM13						
geneM14						
geneM15						
geneM16					10.3	10.9
geneM17						
geneM18						
geneM19						
geneM20						
geneM21					8.06	6.1
geneM22						3.67
geneM23						5.64
geneM24						3.51
geneM25						5.01
geneM26						
geneM27						
geneM28						
geneM29						
geneM30						
geneM31						
geneM32		15.5	10.4			
geneM33						
geneM34						
# of genes :	402	1443	2080	160	904	2213

– signed log₁₀(P): 300, 200, 100, 0, −100, −200, −300

Supplementary Figure 4.9: Enrichment for distinct WGCNA modules among *C4A*-positive and *C4A*-negative genes. Gene sets obtained from the seeded networks at FDR < 0.05 were tested for overlap with PsychENCODE WGCNA modules, which capture neurobiological pathways and cell-types. Out of 34 gene-level (geneM) modules, for *C4A*-positive genes, the strongest enrichment was observed for astrocyte module (geneM3) at low copy number and for NFkB module (geneM5) at subsequently higher copy number. Microglial (geneM6) and interferon-response (geneM32) modules also showed stronger enrichment at higher copy number. Meanwhile for *C4A*-negative genes, we observed the strongest enrichment for synapse- and neuron-related modules. Text shows odds ratio from two-sided Fisher's exact test. Bonferroni-significant results are marked with black borders.
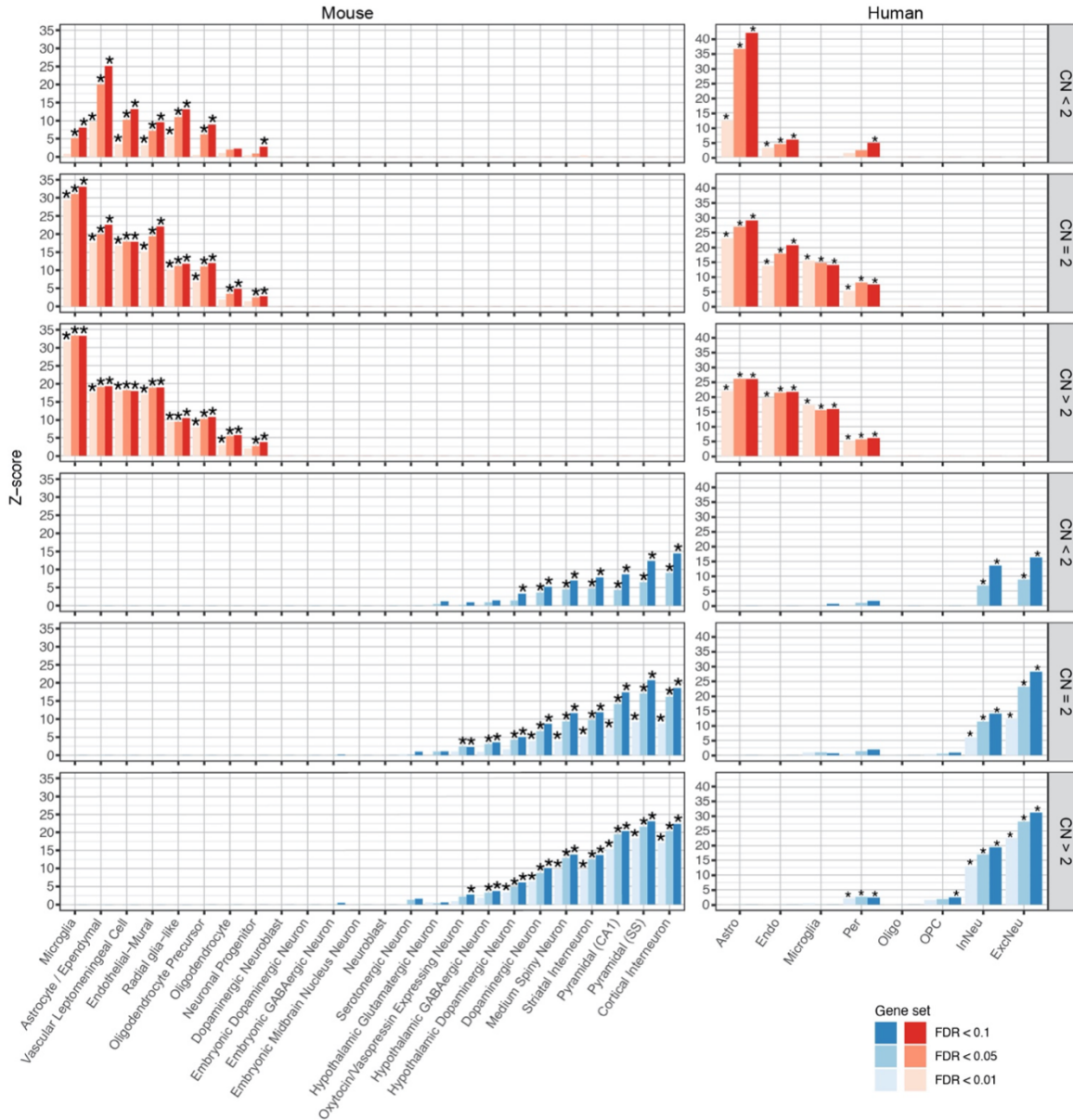
Supplementary Figure 4.10: Expression of *C4A*-positive and *C4A*-negative genes in distinct mouse and human cell-types. Gene sets obtained from the seeded networks were used for EWCE in mouse and human single-cell/nucleus RNA-seq data. All available major cell-types from either mouse or human brain were tested. Note that there were no cell-types from subcortical brain regions in the human dataset. To ensure that the observed enrichment pattern is preserved on a more global and systems-scale, gene sets obtained with more permissive FDR thresholds were also tested. *C4A*-positive and *C4A*-negative genes are shown in red and blue, respectively. Asterisks denote significance at FDR < 0.05.
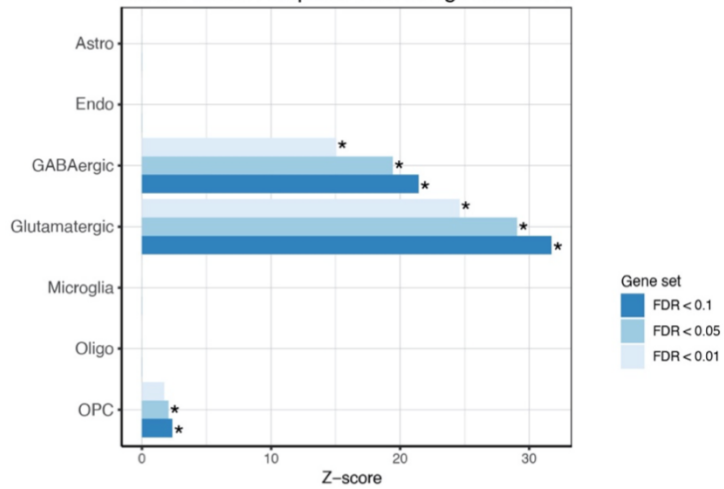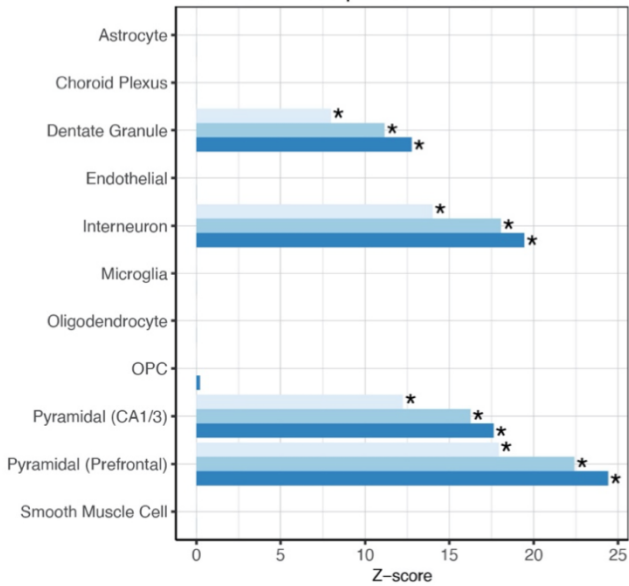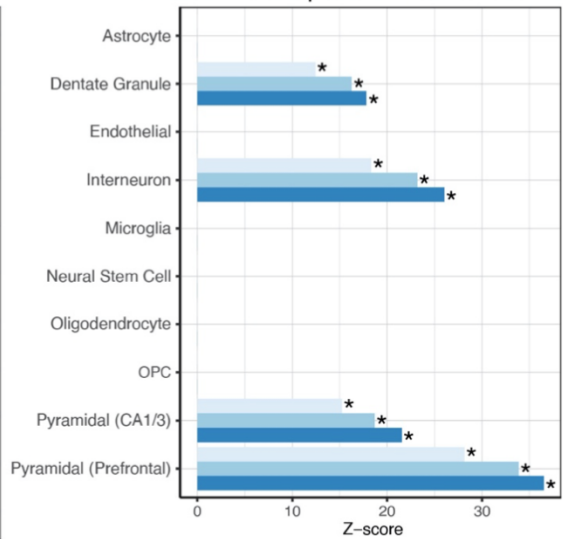
Supplementary Figure 4.11: EWCE analyses in other single-cell/nucleus RNA-seq datasets. *C4A*-negative genes obtained from high CNV group (i.e. CN > 2) at different FDR thresholds were used for EWCE in multiple single-cell/nucleus RNA-seq data. The results were consistent with previous analyses, where *C4A*-negative genes implicate neuronal and synaptic genes. Asterisks denote significance at FDR < 0.05.
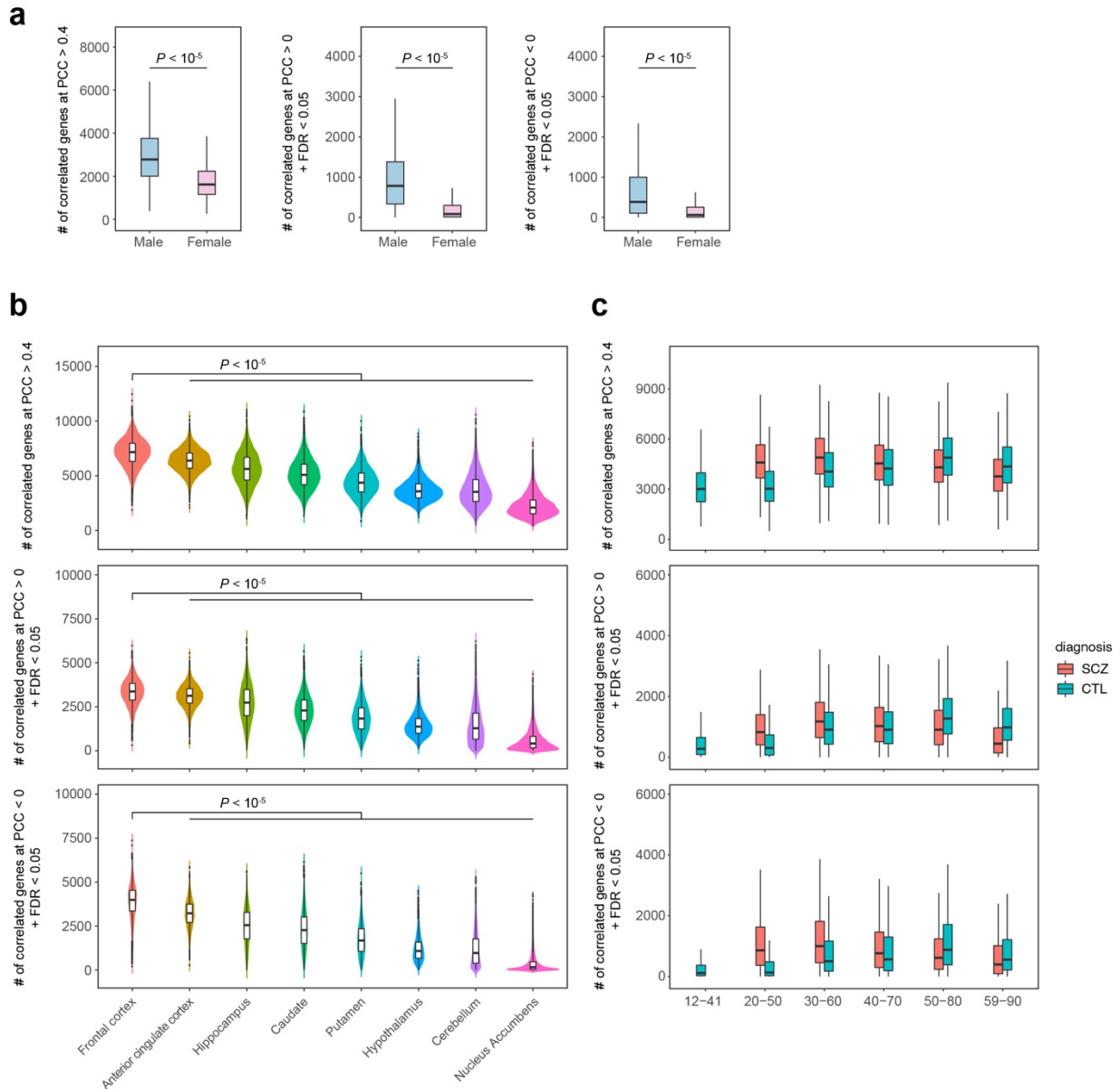
Supplementary Figure 4.12: Sex and spatiotemporal differences in *C4A* co-expression. a, Three different thresholds were tested, namely the number of total co-expressed genes at PCC > 0.4 and the number of *C4A*-positive and *C4A*-negative genes at FDR < 0.05. Males had more co-expressed genes than females regardless of the threshold metric used (permutation test, $P < 10^{-5}$). **b**, Similarly, frontal and anterior cingulate cortex were the two most connected regions for *C4A* regardless of the threshold metric used (permutation test, $P < 10^{-5}$). **c**, Leftward shift in co-expression peak was observed in SCZ cases compared to neurotypical controls across different threshold metrics. All boxplots show median and interquartile range (IQR) with whiskers denoting 1.5 × IQR.
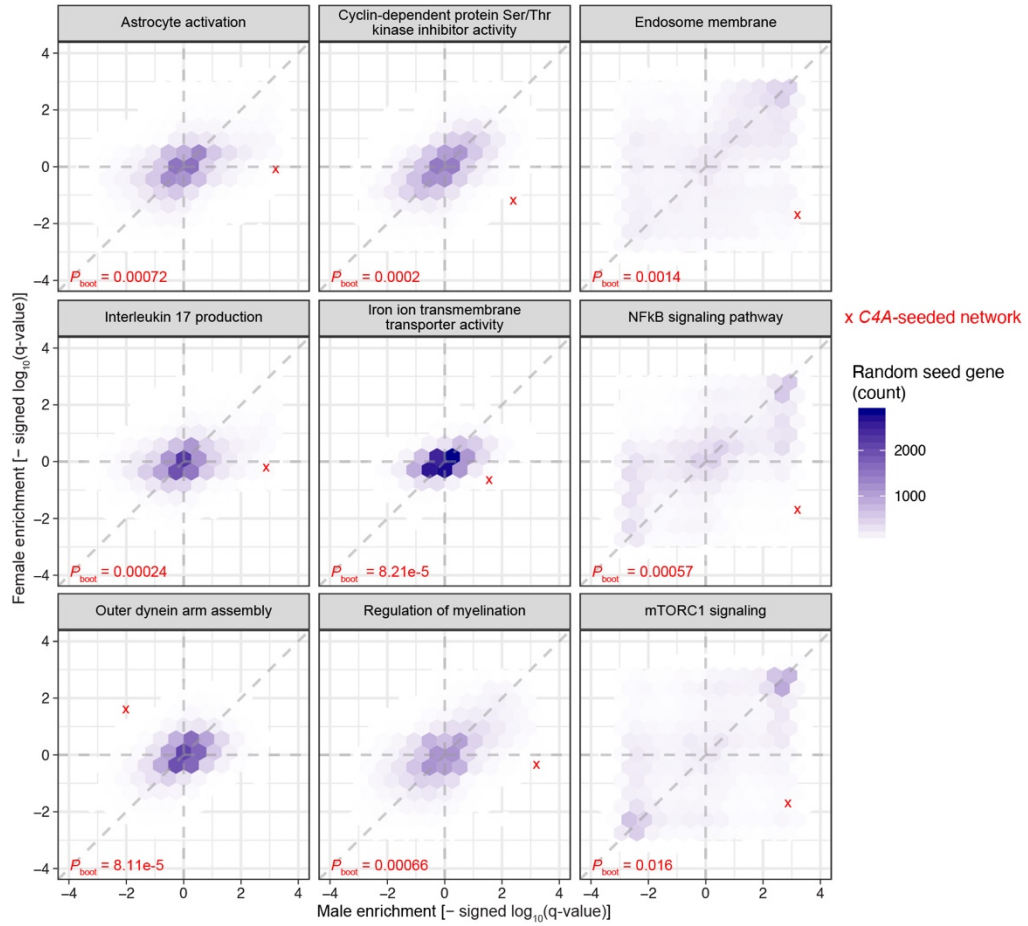
Supplementary Figure 4.13: Pathways exhibiting differential co-expression in males and females. Shown are GSEA enrichments for *C4A* compared to 10,000 random seed genes. Genes were ranked by the magnitude of co-expression in male and female networks separately, and the corresponding gene list was used for GSEA. Several pathways showed the opposite direction of effect.

|	BD (All)	SCZ subsample (BD matched)	SCZ (All)	SCZ (EUR)	SCZ (All) | *C4A* expr	SCZ (EUR) | *C4A* expr	SCZ (EUR) | *C4A* CN	SCZ (EUR) | *C4A* CN + *C4A* expr
C1QA	−0.14	−0.0037	0.013	0.049	−0.11	−0.08	0.041	−0.08
C1QB	−0.043	0.05	0.074	0.1	−0.072	−0.044	0.099	−0.044
C1QC	0.015	0.054	0.073	0.087	−0.075	−0.065	0.081	−0.065
C1R	0.1	0.25 *	0.25 *	0.23 *	0.13 *	0.12 *	0.23 *	0.12 *
C1S	0.039	0.13 *	0.13 *	0.11 *	0.059 *	0.039	0.11 *	0.038
C2	0.061	0.042	0.044	0.025	−0.021	−0.038	0.024	−0.039
C3	−0.18	−0.21 *	−0.19 *	−0.18 *	−0.29 *	−0.29 *	−0.19 *	−0.29 *
C3AR1	−0.18	−0.2 *	−0.18 *	−0.17	−0.26 *	−0.25 *	−0.18	−0.25 *
C4A	0.071	0.4 *	0.39 *	0.44 *	0	0	0.37 *	0
C4B	0.064	0.34 *	0.33 *	0.34 *	0.095 *	0.1	0.35 *	0.1
C5	0.0061	−0.03	−0.023	−0.03	−0.031	−0.039	−0.03	−0.039
C5AR1	0.029	−0.00054	0.02	0.072	−0.061	−0.0087	0.068	−0.0092
C5AR2	−0.079	−0.15 *	−0.13 *	−0.11	−0.15 *	−0.14 *	−0.11	−0.14 *
C7	0.037	0.11	0.094	0.018	0.01	−0.07	0.013	−0.071
C8G	−0.093	−0.078	−0.068	0.042	−0.052	0.065	0.041	0.066
CD46	0.012	−0.024	−0.024	−0.031	−0.021	−0.027	−0.031	−0.027
CD55	−0.0039	−0.018	−0.029	−0.013	−0.046	−0.026	−0.016	−0.026
CD59	0.027	0.054 *	0.055 *	0.053 *	0.032	0.03	0.054 *	0.029
CD93	0.092	0.092	0.11	0.085	−0.024	−0.038	0.08	−0.04
CFB	−0.094	0.047	0.06	0.03	−0.023	−0.039	0.027	−0.04
CFD	−0.17 *	−0.071	−0.077	−0.057	−0.094	−0.081	−0.055	−0.081
CFH	−0.13	0.034	0.035	−0.027	−0.024	−0.093	−0.028	−0.094
CFI	−0.087	0.087	0.091	0.079	−0.042	−0.05	0.077	−0.053
CFP	−0.04	−0.0095	−0.0024	−0.019	0.011	0.0011	−0.0098	0.00053
CLU	0.088 *	0.11 *	0.12 *	0.12 *	0.072 *	0.072 *	0.12 *	0.071 *
COLEC10	0.033	−0.13	−0.11 *	−0.13	−0.09	−0.12	−0.14	−0.12
COLEC11	0.019	0.054	0.058	0.21	0.053	0.19	0.21	0.18
CR1	−0.17	0.19 *	0.14 *	0.26 *	0.022	0.14	0.26 *	0.14
CSMD1	0.0037	0.0028	0.00099	−0.02	0.014	−0.01	−0.019	−0.01
CSMD2	0.038	−0.0019	−0.003	−0.0045	0.0035	0.0053	−0.0022	0.0052
CSMD3	0.0044	0.00025	0.0026	0.0025	0.012	0.013	0.0055	0.013
ELANE	−0.045	−0.13	−0.089	−0.1	−0.11	−0.11	−0.094	−0.11
FCN1	0.017	0.15	0.16	0.34 *	0.071	0.25	0.34 *	0.25
FCN3	−0.055	0.067	0.066	0.031	0.0041	−0.04	0.023	−0.041
ITGAM	−0.14	−0.19 *	−0.18 *	−0.19 *	−0.23 *	−0.25 *	−0.19 *	−0.25 *
ITGAX	−0.14	−0.33 *	−0.31 *	−0.3 *	−0.32 *	−0.32 *	−0.3 *	−0.32 *
ITGB2	−0.15	−0.14	−0.12	−0.14	−0.22 *	−0.24 *	−0.14	−0.24 *
MASP1	0.12 *	0.095	0.11 *	0.049	0.028	−0.032	0.052	−0.033
MASP2	0.011	−0.017	−0.016	−0.02	−0.023	−0.045	−0.024	−0.045
SERPING1	0.13 *	0.16 *	0.16 *	0.13 *	0.098 *	0.072	0.13 *	0.071
VSIG4	0.008	0.11	0.13	0.17	−0.0089	0.03	0.16	0.03
VTN	0.054	−0.048	−0.048	−0.042	−0.03	−0.026	−0.038	−0.026

−log$_{10}$FDR

0.0 2.5 5.0 7.5
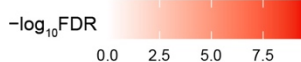
Supplementary Figure 4.14: Differential gene expression of the complement system in SCZ and BD. Differential expression (DE) for brain-expressed complement system genes (n = 42) was assessed in SCZ (N = 531) and BD (N = 217) compared to controls (N = 895). DE was repeated for SCZ after randomly downsampling to match the sample size of BD. DE was also repeated for SCZ while adjusting for *C4A* expression and/or *C4A* copy number. Since *C4A* copy number was only imputed for samples of European ancestry, a subset of PsychENCODE samples was used for such conditional analyses (N = 311 and 367 for SCZ and controls, respectively). Text shows log$_2$FC. Asterisks denote significance at FDR < 0.1.

133
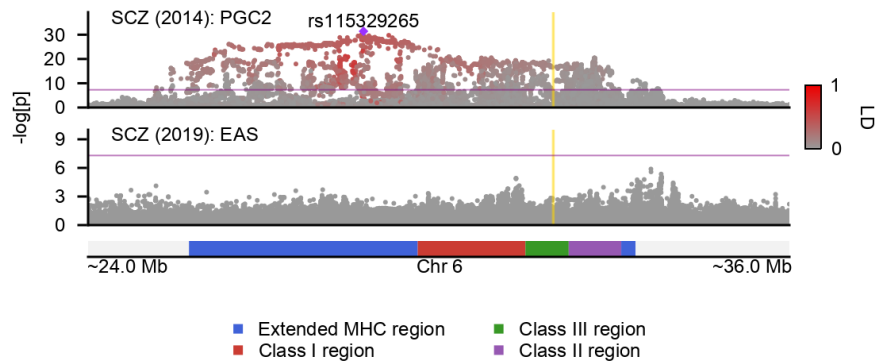
Supplementary Figure 4.15: MHC association for SCZ in European and East Asian samples. There was no sign of MHC association in East Asian population (Lam et al. 2019). Purple line denotes genome-wide significance, and yellow lines denote gene start and end sites for *C4A* gene.
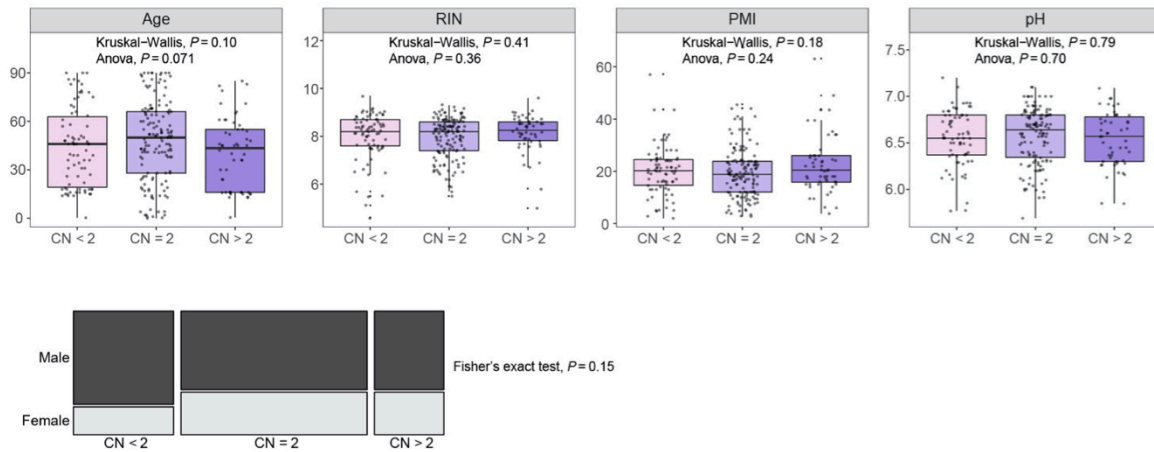
Supplementary Figure 4.16: Baseline characteristics of control samples in PsychENCODE. Age, RIN, postmortem interval (PMI), brain pH, and sex were balanced across the control samples (N = 78, 145, and 54 for CN < 2, CN = 2, and CN > 2, respectively). All boxplots show median and interquartile range (IQR) with whiskers denoting 1.5 × IQR.

Supplementary Figure 4.17: **Characteristics of control-only *C4A*-seeded networks**. **a**, Networks constructed using only the control samples in PsychENCODE expanded in size irrespective of Pearson's correlation coefficient (PCC) threshold. **b**, Correlation of effect sizes (i.e. PCC) from all-sample (All) and control-only (Control) networks were the highest in pairs with equal copy numbers. **c**, Distinct pathways were enriched for *C4A*-positive and *C4A*-negative genes (top 500 genes each ranked by PCC) from control-only networks. **d**, Similarly, distinct cell-types were enriched for *C4A*-positive and *C4A*-negative genes (top 500) from control-only networks.

CHAPTER 5

CONCLUSION

In summary, this dissertation aims to elucidate the neurobiological mechanisms through which established genetic risk factors contribute to the development and progression of mental illnesses. The second chapter shares a high-performance Julia package GeneticsMakie.jl (Kim et al. 2022) for visualizing high-dimensional genetic and genomic data. As the volume and diversity of molecular readouts as well as genetic association results continue to increase, trans-phenotype, trans-tissue, trans-cell-type, trans-ethnic, and trans-omic analyses that consider multiple layers of data will become ever more important. We envision GeneticsMakie.jl facilitating these types of multivariate analyses by enabling fast and seamless data visualization within the larger Julia data science and OpenMendel ecosystems (Zhou et al. 2020). Moreover, Makie.jl (Danisch and Krumbiegel 2021) permits visualization of millions of data points with ease (**Figure 5.1**), and its layout tools allows generation of scientific figures with very complex layouts (**Figures 3.1-3.5**). Reproducible scientific research involves being able to entirely reproduce its figures, which can be easily attained by using GeneticsMakie.jl along with Makie.jl.

The next two chapters tackle the variant-to-gene-to-function problem, specifically the variant-to-gene problem and gene-to-function problem in chapters 3 and 4, respectively. In chapter 3, we systematically dissect the genetic influences on brain gene and isoform expression, partitioning genetic variances and covariances among *cis*- and *trans*-SNPs. Given that statistical tools for fitting multivariate, multiple variance components are currently not available, we implement the minorization-maximization (MM) algorithm in Julia for maximum likelihood (ML) and restricted

maximum likelihood (REML) estimation (Zhou et al. 2019), which benefit from numerical stability and fast convergence while gracefully respecting the positive semidefinite constraint of the variance components parameters. In the future, we plan to improve this Julia package by implementing other optimization algorithms such as Fisher scoring, quasi-Newton, and EM algorithms, dealing with unbalanced or missing data, handling phenotypes with different sets of covariates, and handling phenotypes without residual covariance (Lee et al. 2012; Yang et al. 2011). We could deal with two variance components cases in a more computationally efficient manner as well with the generalized eigenvalue decomposition of kinship matrices (Lee and van der Werf 2016; Zhou et al. 2019). Finally, we can try to impose biologically plausible constraints on the variance components parameters—for example, a low-rank constraint under the assumption that *cis* or *trans* genetic effects on a particular gene and its constituent isoforms are structured. This can aid in estimation step, especially when there are many phenotypes to be modeled jointly, by providing additional information.

In chapter 3, although we quantify the degree of polygenicity and pleiotropy for brain gene and isoform expression, we do not quantify contributions from other sources of variance-covariance such as dominance effects (Hivert et al. 2021; Pazokitoroudi et al. 2021; Zhu et al. 2015), gene-gene interaction (e.g. epistasis), gene-environment interaction, gene-environment correlation, and shared environment, among others. These contributions are assessed by variance components models but with larger sample sizes than PsychENCODE and/or pedigree-based designs. We also do not investigate the relationship between minor allele frequency (MAF) and effect size, which is one of themes the keyword "genetic architecture" is broadly concerned with. On the contrary, by connecting isoform-level eQTL to known GWAS signals, we illustrate the potential of

identifying candidate causal genes with isoform-resolution analyses (i.e. variant-to-gene or more aptly named variant-to-isoform-to-gene problem). By formally comparing these isoform-level and GWAS signals transcriptome-wide—for example, through colocalization (Giambartolomei et al. 2014), transcriptome-wide association studies (TWAS) (Gusev et al. 2016), and summary-data-based Mendelian randomization (SMR) (Zhu et al. 2016)—we hypothesize that we would uncover many more disease genes. In fact, there are multiple high-confidence psychiatric risk genes that are currently missing gene-level eQTL as well as other molecular QTL signals. Such genes include *CACNA1C*, *NRXN1*, *RBFOX1*, *GRIN2A*, *SP4*, *FAM120A*, *STAG1*, *DRD2*, and *KMT2E*, among others. Based on the *XRN2* example in chapter 3, we suspect that such missing eQTL information
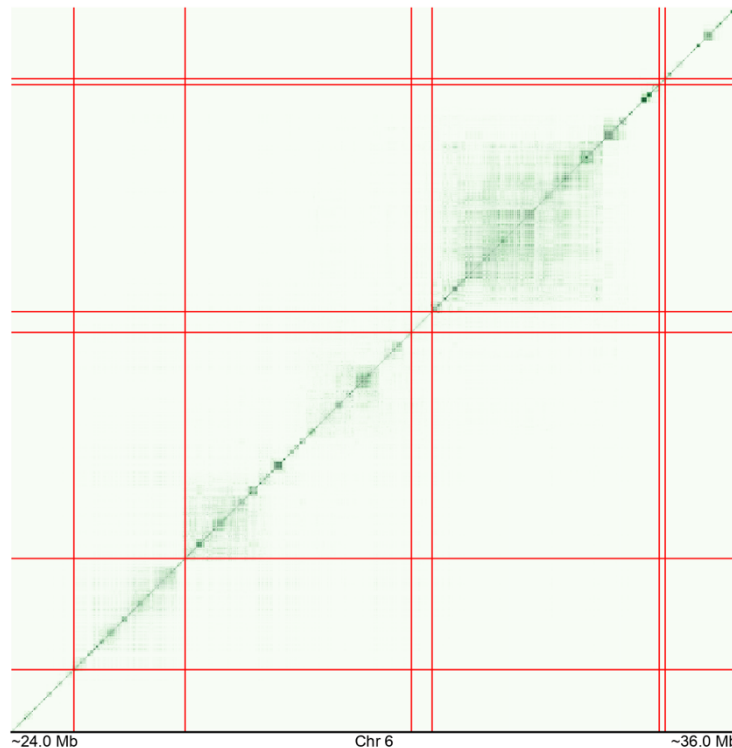
Figure 5.1: LD structure in the MHC region. The strength of LD (r^2) is shown as a heatmap for ~66,000 SNPs in the MHC region. LD is calculated from individuals of European ancestry in the 1000 Genomes Project reference panel. The heatmap contains ~2.2 billion distinct data points. To the best of our knowledge, Makie.jl is the only plotting package where it is possible to visualize the entire LD matrix for the MHC region. The red lines delineate flanking regions, extended MHC regions, and class I, II, III regions (Horton et al. 2004). The two biggest squares on the diagonal correspond to class I and II regions, which are known to be extremely polymorphic.

CSMD1 locus

Figure 5.2: A close look at the *CSMD1* locus using GeneticsMakie.jl. GWAS results for 56 complex phenotypes are shown, which span autoimmune, endocrine, psychiatric, cardiovascular disorders, and cancer. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *CSMD1* gene. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), RBC (red blood cell), SCZ (schizophrenia).

could be retrieved by looking at isoform-level expression with improved transcriptome annotations, such as those generated from long-read RNA sequencing.

In chapter 4, we investigate the functional role of *C4A* in the human brain and its relation to schizophrenia (SCZ) genetic risk factors by characterizing the effect of *C4A* copy number variation (CNV) on gene expression and gene co-expression at the level of bulk tissue (Kim et al. 2021). We find that synaptic pathways rather than the complement system are the key SCZ-relevant pathways. The effect of genetic variation on gene co-expression is conceptually equivalent to co-expression QTL (Lea et al. 2019; van der Wijst et al. 2018), which do not have a clear biological meaning yet, and therefore needs to stand the test of time or undergo experimental validation. Although the complement signaling pathway is not enriched for SCZ genetic risk, there is evidence that several individual genes within the complement system other than *C4A* harbor SCZ GWAS signals, namely *CSMD1*, *CLU*, and *CD46* genes. Interestingly, the *CSMD1* locus exhibits allelic series with distinct GWAS signals from multiple phenotypes such as age at menarche, educational attainment, neuroticism, height, and weight in addition to SCZ. (**Figure 5.2**). The *CLU* locus is a well-known locus that harbors Alzheimer disease association (Pouget 2018), which we find to be distinct from SCZ association (**Figure 5.3**). Despite reaching genome-wide significance in some SCZ GWAS, *CD46* did not pass genome-wide threshold in subsequent GWAS with larger sample sizes (**Figure 5.4**), which indicates that this region may not be a bona fide disease-associated locus.

CLU locus

141

Figure 5.3: A close look at the *CLU* locus using GeneticsMakie.jl. GWAS results for 56 complex phenotypes are shown. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding LD between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *CLU* gene. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), AMD (age-related macular degeneration), BD (bipolar disorder), CAD (coronary artery disease), CKD (chronic kidney disease), IBD (inflammatory bowel disease), RBC (red blood cell), SCZ (schizophrenia).

For *CSMD1* and *CLU* loci, the variant-to-gene-to-function problem needs to be addressed more

thoroughly in the future.

We also find that differential expression of *C4A* between SCZ cases and neurotypical controls

cannot be completely explained by *C4A* CNV, suggesting that other (potentially genetic) factors

contribute to overexpression of *C4A* in SCZ. Elucidating these genetic and/or environmental

factors that influence *C4A* and other differentially expressed genes in SCZ will be crucial for

interpreting differential expression findings. Next, the MHC region remains to be further fine-

mapped, since *C4* structural variation drives only a part of the association signal (**Figure 5.5**). In

fact, there seems to be at least three distinct signals (Sekar et al. 2016), one of which is *C4* structural

alleles. Meanwhile, it is unclear why there is a lack of significant association in the MHC region

for SCZ in East Asian population (Lam et al. 2019). Plausible explanations include differences in

case ascertainment and gene-environment interaction. Aside from gene expression and co-

expression, there have been efforts to identify intermediate phenotypes associated with *C4A* CNV

in biobank-scale data such as UK Biobank and the Adolescent Brain Cognitive Development

(ABCD) study. The focus has been largely on brain imaging measures, given that *C4A* and the

complement system could mediate synaptic pruning in disease-susceptible brain regions and alter

brain structure (e.g. cortical thinning), but the strength of evidence has been so far underwhelming.

Multiple factors can lead to either weak or absence of association: we may not be looking at the

relevant developmental time points, particularly adolescence when mental illnesses usually emerge; we might not be testing the right intermediate phenotypes; we may not have sufficient sample size and hence power to detect the underlying signal; and variation in *C4A* expression may only induce fine-scale changes at the level of local brain circuitry, in which case we may not observe gross changes at the level of brain structure and behavior.
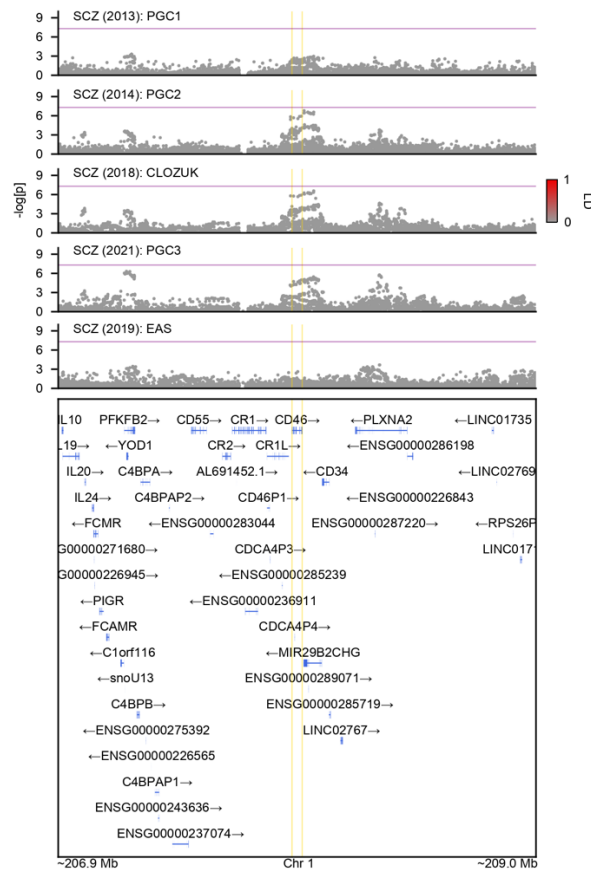
Figure 5.4: *CD46* locus for schizophrenia with increasing sample size. The second to last row shows GWAS result in East Asian population. All GWAS results did not reach genome-wide significance. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for *CD46* gene.

Looking forward, by expanding upon the results and approaches covered in this dissertation, we hope that we can deliver the promise of human genetics, identifying novel disease genes and biological pathways for intervention, and ultimately nominating robust therapeutic targets for psychiatric disorders.
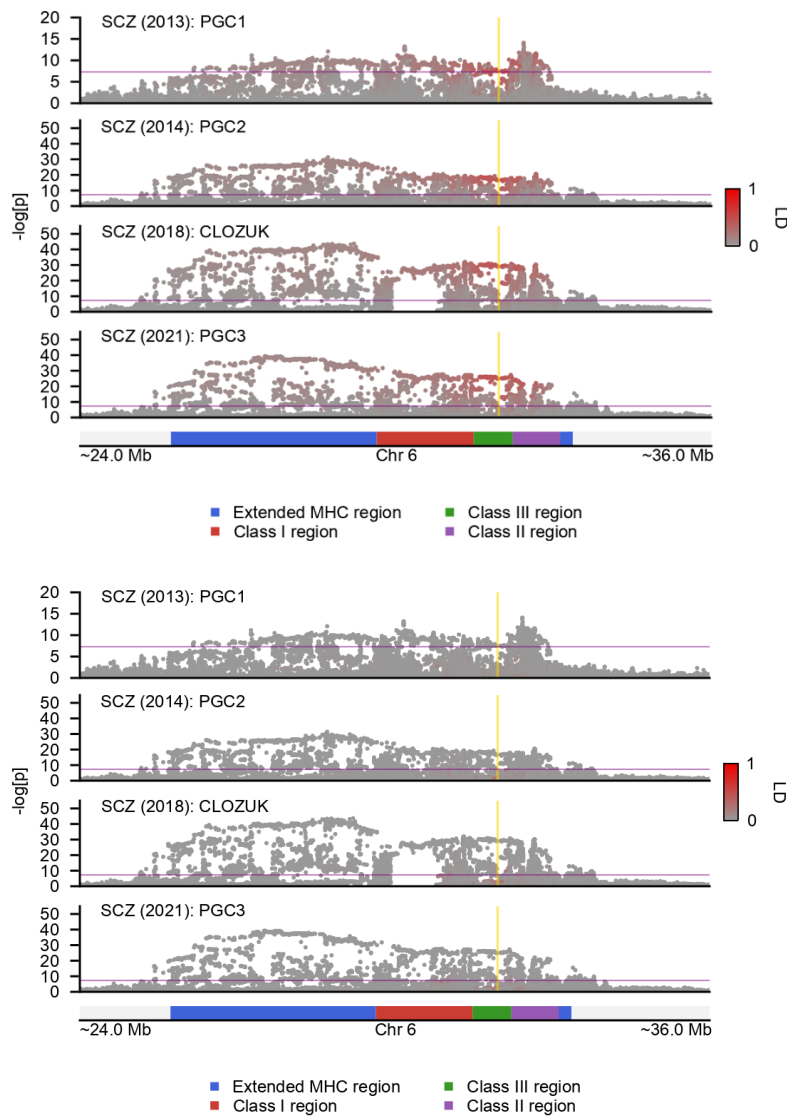
Figure 5.5: MHC association for schizophrenia and its relation to predicted _C4_ expression. Top, squared correlation (r^2) with predicted _C4A_ expression is displayed with the intensity of red color. Bottom, r^2 with predicted _C4B_ expression is displayed. We imputed _C4_ structural alleles for individuals of European ancestry in the 1000 Genomes Project reference panel. _C4_ gene expression was predicted with previously derived expression weights (Sekar et al. 2016). Correlation between _C4_ gene expression and other SNP genotypes in the MHC region was subsequently calculated. Note that GWAS SNPs are mildly correlated with predicted _C4A_ but not predicted _C4B_ expression with the strength of correlation increasing in the vicinity of _C4_ genes. This observation indicates that there are GWAS signals independent from _C4A_ expression that remain to be fine-mapped. Alternative approaches such as conditional analyses using summary statistics (Yang et al. 2012) to adjust for predicted _C4_ expression is also possible, but appropriate _C4_ association statistics are unfortunately not available. Purple line denotes genome-wide significance ($P = 5 \times 10^{-8}$), and yellow lines denote gene start and end sites for _C4A_ gene.

REFERENCES

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., and Abecasis, G. R. (2015), "A global reference for human genetic variation," *Nature*, 526, 68–74.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010), "Data quality control in genetic case-control association studies," *Nature protocols*, 5, 1564–1573.

Aygün, N., Elwell, A. L., Liang, D., Lafferty, M. J., Cheek, K. E., Courtney, K. P., Mory, J., Hadden-Ford, E., Krupa, O., de la Torre-Ubieta, L., Geschwind, D. H., Love, M. I., and Stein, J. L. (2021), "Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis," *American journal of human genetics*, 108, 1647–1668.

Ballouz, S., Verleyen, W., and Gillis, J. (2015), "Guidance for RNA-seq co-expression network construction and analysis: safety in numbers," *Bioinformatics*, 31, 2123–2130.

Bentham, J., Morris, D. L., Graham, D. S. C., Pinder, C. L., Tombleson, P., Behrens, T. W., Martín, J., Fairfax, B. P., Knight, J. C., Chen, L., Replogle, J., Syvänen, A.-C., Rönnblom, L., Graham, R. R., Wither, J. E., Rioux, J. D., Alarcón-Riquelme, M. E., and Vyse, T. J. (2015), "Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus," *Nature genetics*, 47, 1457–1464.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017), "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, 59, 65–98.

Bitner-Mathé, B. C., and Klaczko, L. B. (1999), "Heritability, phenotypic and genetic correlations of size and shape of Drosophila mediopunctata wings," *Heredity*, 83, 688–696.

Boettger, L. M., Handsaker, R. E., Zody, M. C., and McCarroll, S. A. (2012), "Structural haplotypes and recent evolution of the human 17q21.31 region," *Nature genetics*, 44, 881–885.

Boix, C. A., James, B. T., Park, Y. P., Meuleman, W., and Kellis, M. (2021), "Regulatory genomic circuitry of human disease loci by integrative epigenomics," *Nature*, 590, 300–307.

Boughton, A. P., Welch, R. P., Flickinger, M., VandeHaar, P., Taliun, D., Abecasis, G. R., and Boehnke, M. (2021), "LocusZoom.js: Interactive and embeddable visualization of genetic association study results," *Bioinformatics*, 37, 3017-3018.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016), "Near-optimal probabilistic RNA-seq quantification," *Nature biotechnology*, 34, 525–527.

Breen, M. S., Dobbyn, A., Li, Q., Roussos, P., Hoffman, G. E., Stahl, E., Chess, A., Sklar, P., Li, J. B., Devlin, B., Buxbaum, J. D., and CommonMind Consortium (2019), "Global landscape and genetic regulation of RNA editing in cortical samples from individuals with schizophrenia," *Nature neuroscience*, 22, 1402–1412.

Browning, B. L., and Browning, S. R. (2016), "Genotype Imputation with Millions of Reference Samples," *American journal of human genetics*, 98, 116–126.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, 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," *Nature genetics*, 47, 291–295.

Bulik-Sullivan, B. K., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., and Neale, B. M. (2015), "An atlas of genetic correlations across human diseases and traits," *Nature genetics*, 47, 1236–1241.

Carroll, R. J., Bastarache, L., and Denny, J. C. (2014), "R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment," *Bioinformatics*, 30, 2375–2376.

Cheverud, J. M. (1988), "A Comparison of Genetic and Phenotypic Correlations," *Evolution*, 42, 958–968.

Choi, L., and An, J.-Y. (2021), "Genetic architecture of autism spectrum disorder: Lessons from large-scale genomic studies," *Neuroscience and biobehavioral reviews*, 128, 244–257.

Clark, M. B., Wrzesinski, T., Garcia, A. B., Hall, N. A. L., Kleinman, J. E., Hyde, T., Weinberger, D. R., Harrison, P. J., Haerty, W., and Tunbridge, E. M. (2020), "Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain," *Molecular psychiatry*, 25, 37–47.

Collado-Torres, L., Burke, E. E., Peterson, A., Shin, J., Straub, R. E., Rajpurohit, A., Semick, S. A., Ulrich, W. S., BrainSeq Consortium, Price, A. J., Valencia, C., Tao, R., Deep-Soboslay, A., Hyde, T. M., Kleinman, J. E., Weinberger, D. R., and Jaffe, A. E. (2019), "Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia," *Neuron*, 103, 203-216.

Coulthard, L. G., Hawksworth, O. A., and Woodruff, T. M. (2018), "Complement: The Emerging Architect of the Developing Brain," *Trends in neurosciences*, 41, 373–384.

Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., Mowry, B. J., Thapar, A., Goddard, M. E., Witte, J. S., Absher, D., Agartz, I., Akil, H., Amin, F., Andreassen, O. A., Anjorin, A., Anney, R., Anttila, V., Arking, D. E., Asherson, P., Azevedo, M. H., Backlund, L., Badner, J. A., Bailey, A. J., Banaschewski, T., Barchas, J. D., Barnes, M. R., Barrett, T. B., Bass, N., Battaglia, A., Bauer, M., Bayés, M., Bellivier, F., Bergen, S. E., Berrettini, W., Betancur, C., Bettecken, T., Biederman, J., Binder, E. B., Black, D. W., Blackwood, D. H. R., Bloss, C. S., Boehnke, M., Boomsma, D. I., Breen, G., Breuer, R., Bruggeman, R., Cormican, P., Buccola, N. G., Buitelaar, J. K., Bunney, W. E., Buxbaum, J. D., Byerley, W. F., Byrne, E. M., Caesar, S., Cahn, W., Cantor, R. M., Casas, M., Chakravarti, A., Chambert, K., Choudhury, K., Cichon, S., Cloninger, C. R., Collier, D. A., Cook, E. H., Coon, H., Cormand, B., Corvin, A., Coryell, W. H., Craig, D. W., Craig, I. W., Crosbie, J., Cuccaro, M. L., Curtis, D., Czamara, D., Datta, S., Dawson, G., Day, R., De Geus, E. J., Degenhardt, F., Djurovic, S., Donohoe, G. J., Doyle, A. E., Duan, J., Dudbridge, F., Duketis, E., Ebstein, R. P., Edenberg, H. J., Elia, J., Ennis, S., Etain, B., Fanous, A., Farmer, A. E., Ferrier, I. N., Flickinger, M., Fombonne, E., Foroud, T., Frank, J., Franke, B., Fraser, C., Freedman, R., Freimer, N. B., Freitag, C. M., Friedl, M., Frisén, L., Gallagher, L., Gejman, P. V., Georgieva, L., Gershon, E. S., Geschwind, D. H., Giegling, I., Gill, M., Gordon, S. D., Gordon-Smith, K., Green, E. K., Greenwood, T. A.,

Grice, D. E., Gross, M., Grozeva, D., Guan, W., Gurling, H., De Haan, L., Haines, J. L., Hakonarson, H., Hallmayer, J., Hamilton, S. P., Hamshere, M. L., Hansen, T. F., Hartmann, A. M., Hautzinger, M., Heath, A. C., Henders, A. K., Herms, S., Hickie, I. B., Hipolito, M., Hoefels, S., Holmans, P. A., Holsboer, F., Hoogendijk, W. J., Hottenga, J.-J., Hultman, C. M., Hus, V., Ingason, A., Ising, M., Jamain, S., Jones, E. G., Jones, I., Jones, L., Tzeng, J.-Y., Kähler, A. K., Kahn, R. S., Kandaswamy, R., Keller, M. C., Kennedy, J. L., Kenny, E., Kent, L., Kim, Y., Kirov, G. K., Klauck, S. M., Klei, L., Knowles, J. A., Kohli, M. A., Koller, D. L., Konte, B., Korszun, A., Krabbendam, L., Krasucki, R., Kuntsi, J., Kwan, P., Landén, M., Långström, N., Lathrop, M., Lawrence, J., Lawson, W. B., Leboyer, M., Ledbetter, D. H., Lee, P. H., Lencz, T., Lesch, K.-P., Levinson, D. F., Lewis, C. M., Li, J., Lichtenstein, P., Lieberman, J. A., Lin, D.-Y., Linszen, D. H., Liu, C., Lohoff, F. W., Loo, S. K., Lord, C., Lowe, J. K., Lucae, S., MacIntyre, D. J., Madden, P. A. F., Maestrini, E., Magnusson, P. K. E., Mahon, P. B., Maier, W., Malhotra, A. K., Mane, S. M., Martin, C. L., Martin, N. G., Mattheisen, M., Matthews, K., Mattingsdal, M., McCarroll, S. A., McGhee, K. A., McGough, J. J., McGrath, P. J., McGuffin, P., McInnis, M. G., McIntosh, A., McKinney, R., McLean, A. W., McMahon, F. J., McMahon, W. M., McQuillin, A., Medeiros, H., Medland, S. E., Meier, S., Melle, I., Meng, F., Meyer, J., Middeldorp, C. M., Middleton, L., Milanova, V., Miranda, A., Monaco, A. P., Montgomery, G. W., Moran, J. L., Moreno-De-Luca, D., Morken, G., Morris, D. W., Morrow, E. M., Moskvina, V., Muglia, P., Mühleisen, T. W., Muir, W. J., Müller-Myhsok, B., Murtha, M., Myers, R. M., Myin-Germeys, I., Neale, M. C., Nelson, S. F., Nievergelt, C. M., Nikolov, I., Nimgaonkar, V., Nolen, W. A., Nöthen, M. M., Nurnberger, J. I., Nwulia, E. A., Nyholt, D. R., O'Dushlaine, C., Oades, R. D., Olincy, A., Oliveira, G., Olsen, L., Ophoff, R. A., Osby, U., Owen, M. J., Palotie, A., Parr, J. R., Paterson, A. D., Pato, C. N., Pato, M. T., Penninx, B. W., Pergadia, M. L., Pericak-Vance, M. A., Pickard, B. S., Pimm, J., Piven, J., Posthuma, D., Potash, J. B., Poustka, F., Propping, P., Puri, V., Quested, D. J., Quinn, E. M., Ramos-Quiroga, J. A., Rasmussen, H. B., Raychaudhuri, S., Rehnström, K., Reif, A., Ribasés, M., Rice, J. P., Rietschel, M., Roeder, K., Roeyers, H., Rossin, L., Rothenberger, A., Rouleau, G., Ruderfer, D., Rujescu, D., Sanders, A. R., Sanders, S. J., Santangelo, S. L., Sergeant, J. A., Schachar, R., Schalling, M., Schatzberg, A. F., Scheftner, W. A., Schellenberg, G. D., Scherer, S. W., Schork, N. J., Schulze, T. G., Schumacher, J., Schwarz, M., Scolnick, E., Scott, L. J., Shi, J., Shilling, P. D., Shyn, S. I., Silverman, J. M., Slager, S. L., Smalley, S. L., Smit, J. H., Smith, E. N., Sonuga-Barke, E. J. S., St Clair, D., State, M., Steffens, M., Steinhausen, H.-C., Strauss, J. S., Strohmaier, J., Stroup, T. S., Sutcliffe, J. S., Szatmari, P., Szelinger, S., Thirumalai, S., Thompson, R. C., Todorov, A. A., Tozzi, F., Treutlein, J., Uhr, M., van den Oord, E. J. C. G., Van Grootheest, G., Van Os, J., Vicente, A. M., Vieland, V. J., Vincent, J. B., Visscher, P. M., Walsh, C. A., Wassink, T. H., Watson, S. J., Weissman, M. M., Werge, T., Wienker, T. F., Wijsman, E. M., Willemsen, G., Williams, N., Willsey, A. J., Witt, S. H., Xu, W., Young, A. H., Yu, T. W., Zammit, S., Zandi, P. P., Zhang, P., Zitman, F. G., Zöllner, S., Devlin, B., Kelsoe, J. R., Sklar, P., Daly, M. J., O'Donovan, M. C., Craddock, N., Sullivan, P. F., Smoller, J. W., Kendler, K. S., Wray, N. R., and International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) (2013), "Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs," *Nature genetics*, 45, 984–994.

Dadaev, T., Leongamornlert, D. A., Saunders, E. J., Eeles, R., and Kote-Jarai, Z. (2016), "LocusExplorer: a user-friendly tool for integrated visualization of human genetic association data and biological annotations," *Bioinformatics*, 32, 949–951.

Cui, Y., Peng, F., Wang, D., Li, Y., Li, J. S., Li, L., and Li, W. (2021), "3'aQTL-atlas: an atlas of 3'UTR alternative polyadenylation quantitative trait loci across human normal tissues," *Nucleic acids research*, 50, D39–D45.

Dahl, A., Guillemot, V., Mefford, J., Aschard, H., and Zaitlen, N. (2019), "Adjusting for Principal Components of Molecular Phenotypes Induces Replicating False Positives," *Genetics*, 211, 1179–

1189.

Danisch, S., and Krumbiegel, J. (2021), "Makie.jl: Flexible high-performance data visualization for Julia," *Journal of Open Source Software*, 6, 3349.

Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., and Dermitzakis, E. T. (2017), "A complete tool set for molecular QTL discovery and analysis," *Nature communications*, 8, 15452.

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Churchhouse, C., Dumont, A., Eriksson, N., Gandal, M., Goldstein, J. I., Grasby, K. L., Grove, J., Gudmundsson, O. O., Hansen, C. S., Hauberg, M. E., Hollegaard, M. V., Howrigan, D. P., Huang, H., Maller, J. B., Martin, A. R., Martin, N. G., Moran, J., Pallesen, J., Palmer, D. S., Pedersen, C. B., Pedersen, M. G., Poterba, T., Poulsen, J. B., Ripke, S., Robinson, E. B., Satterstrom, F. K., Stefansson, H., Stevens, C., Turley, P., Walters, G. B., Won, H., Wright, M. J., ADHD Working Group of the Psychiatric Genomics Consortium (PGC), Early Lifecourse & Genetic Epidemiology (EAGLE) Consortium, 23andMe Research Team, Andreassen, O. A., Asherson, P., Burton, C. L., Boomsma, D. I., Cormand, B., Dalsgaard, S., Franke, B., Gelernter, J., Geschwind, D., Hakonarson, H., Haavik, J., Kranzler, H. R., Kuntsi, J., Langley, K., Lesch, K.-P., Middeldorp, C., Reif, A., Rohde, L. A., Roussos, P., Schachar, R., Sklar, P., Sonuga-Barke, E. J. S., Sullivan, P. F., Thapar, A., Tung, J. Y., Waldman, I. D., Medland, S. E., Stefansson, K., Nordentoft, M., Hougaard, D. M., Werge, T., Mors, O., Mortensen, P. B., Daly, M. J., Faraone, S. V., Børglum, A. D., and Neale, B. M. (2019), "Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder," *Nature genetics*, 51, 63–75.

van Erp, T. G. M., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., Pearlson, G. D., Yao, N., Fukunaga, M., Hashimoto, R., Okada, N., Yamamori, H., Bustillo, J. R., Clark, V. P., Agartz, I., Mueller, B. A., Cahn, W., de Zwarte, S. M. C., Hulshoff Pol, H. E., Kahn, R. S., Ophoff, R. A., van Haren, N. E. M., Andreassen, O. A., Dale, A. M., Doan, N. T., Gurholt, T. P., Hartberg, C. B., Haukvik, U. K., Jørgensen, K. N., Lagerberg, T. V., Melle, I., Westlye, L. T., Gruber, O., Kraemer, B., Richter, A., Zilles, D., Calhoun, V. D., Crespo-Facorro, B., Roiz-Santiañez, R., Tordesillas-Gutiérrez, D., Loughland, C., Carr, V. J., Catts, S., Cropley, V. L., Fullerton, J. M., Green, M. J., Henskens, F. A., Jablensky, A., Lenroot, R. K., Mowry, B. J., Michie, P. T., Pantelis, C., Quidé, Y., Schall, U., Scott, R. J., Cairns, M. J., Seal, M., Tooney, P. A., Rasser, P. E., Cooper, G., Shannon Weickert, C., Weickert, T. W., Morris, D. W., Hong, E., Kochunov, P., Beard, L. M., Gur, R. E., Gur, R. C., Satterthwaite, T. D., Wolf, D. H., Belger, A., Brown, G. G., Ford, J. M., Macciardi, F., Mathalon, D. H., O'Leary, D. S., Potkin, S. G., Preda, A., Voyvodic, J., Lim, K. O., McEwen, S., Yang, F., Tan, Y., Tan, S., Wang, Z., Fan, F., Chen, J., Xiang, H., Tang, S., Guo, H., Wan, P., Wei, D., Bockholt, H. J., Ehrlich, S., Wolthusen, R. P. F., King, M. D., Shoemaker, J. M., Sponheim, S. R., De Haan, L., Koenders, L., Machielsen, M. W., van Amelsvoort, T., Veltman, D. J., Assogna, F., Banaj, N., de Rossi, P., Iorio, M., Piras, F., Spalletta, G., McKenna, P. J., Pomarol-Clotet, E., Salvador, R., Corvin, A., Donohoe, G., Kelly, S., Whelan, C. D., Dickie, E. W., Rotenberg, D., Voineskos, A. N., Ciufolini, S., Radua, J., Dazzan, P., Murray, R., Reis Marques, T., Simmons, A., Borgwardt, S., Egloff, L., Harrisberger, F., Riecher-Rössler, A., Smieskova, R., Alpert, K. I., Wang, L., Jönsson, E. G., Koops, S., Sommer, I. E. C., Bertolino, A., Bonvino, A., Di Giorgio, A., Neilson, E., Mayer, A. R., Stephen, J. M., Kwon, J. S., Yun, J.-Y., Cannon, D. M., McDonald, C., Lebedeva, I., Tomyshev, A. S., Akhadov, T., Kaleda, V., Fatouros-Bergman, H., Flyckt, L., Karolinska Schizophrenia Project, Busatto, G. F., Rosa, P. G. P., Serpa, M. H., Zanetti, M. V., Hoschl, C., Skoch, A., Spaniel, F., Tomecek, D., Hagenaars, S. P., McIntosh, A. M., Whalley, H. C., Lawrie, S. M., Knöchel, C., Oertel-Knöchel, V., Stäblein, M., Howells, F. M., Stein, D. J., Temmingh, H. S., Uhlmann, A., Lopez-Jaramillo, C., Dima, D., McMahon, A., Faskowitz, J. I., Gutman, B. A.,

Jahanshad, N., Thompson, P. M., and Turner, J. A. (2018), "Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium," *Biological psychiatry*, 84, 644–654.

Feinberg, I. (1982), "Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence?," *Journal of psychiatric research*, 17, 319–334.

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., Genovese, G., Saunders, A., Macosko, E., Pollack, S., Brainstorm Consortium, Perry, J. R. B., Buenrostro, J. D., Bernstein, B. E., Raychaudhuri, S., McCarroll, S., Neale, B. M., and Price, A. L. (2018), "Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types," *Nature genetics*, 50, 621–629.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Howe, K. L., Hunt, T., Izuogu, O. G., Johnson, R., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Riera, F. C., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Wolf, M. Y., Xu, J., Yang, Y. T., Yates, A., Zerbino, D., Zhang, Y., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Tress, M. L., and Flicek, P. (2021), "GENCODE 2021," *Nucleic acids research*, 49, D916–D923.

Flaherty, E., Zhu, S., Barretto, N., Cheng, E., Deans, P. J. M., Fernando, M. B., Schrode, N., Francoeur, N., Antoine, A., Alganem, K., Halpern, M., Deikus, G., Shah, H., Fitzgerald, M., Ladran, I., Gochman, P., Rapoport, J., Tsankova, N. M., McCullumsmith, R., Hoffman, G. E., Sebra, R., Fang, G., and Brennand, K. J. (2019), "Neuronal impact of patient-specific aberrant NRXN1α splicing," *Nature genetics*, 51, 1679–1690.

Foerster, P., Daclin, M., Asm, S., Faucourt, M., Boletta, A., Genovesio, A., and Spassky, N. (2017), "mTORC1 signaling and primary cilia are required for brain ventricle morphogenesis," *Development* , 144, 201–210.

Forsingdal, A., Jørgensen, T. N., Olsen, L., Werge, T., Didriksen, M., and Nielsen, J. (2018), "Can Animal Models of Copy Number Variants That Predispose to Schizophrenia Elucidate Underlying Biology?," *Biological psychiatry*, 85, 13–24.

Fritsche, L. G., Igl, W., Bailey, J. N. C., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., Burdon, K. P., Hebbring, S. J., Wen, C., Gorski, M., Kim, I. K., Cho, D., Zack, D., Souied, E., Scholl, H. P. N., Bala, E., Lee, K. E., Hunter, D. J., Sardell, R. J., Mitchell, P., Merriam, J. E., Cipriani, V., Hoffman, J. D., Schick, T., Lechanteur, Y. T. E., Guymer, R. H., Johnson, M. P., Jiang, Y., Stanton, C. M., Buitendijk, G. H. S., Zhan, X., Kwong, A. M., Boleda, A., Brooks, M., Gieser, L., Ratnapriya, R., Branham, K. E., Foerster, J. R., Heckenlively, J. R., Othman, M. I., Vote, B. J., Liang, H. H., Souzeau, E., McAllister, I. L., Isaacs, T., Hall, J., Lake, S., Mackey, D. A., Constable, I. J., Craig, J. E., Kitchner, T. E., Yang, Z., Su, Z., Luo, H., Chen, D., Ouyang, H., Flagg, K., Lin, D., Mao, G., Ferreyra, H., Stark, K., von Strachwitz, C. N., Wolf, A., Brandl, C., Rudolph, G., Olden, M., Morrison, M. A., Morgan, D. J., Schu, M., Ahn, J., Silvestri, G., Tsironi, E. E., Park, K. H., Farrer, L. A., Orlin, A., Brucker, A., Li, M., Curcio, C. A., Mohand-Saïd, S., Sahel, J.-A., Audo, I., Benchaboune, M., Cree, A. J., Rennie, C. A., Goverdhan, S. V., Grunin, M., Hagbi-Levi, S., Campochiaro, P., Katsanis, N., Holz, F. G., Blond, F., Blanché, H., Deleuze, J.-F., Igo, R. P., Jr, Truitt, B., Peachey, N. S., Meuer, S. M., Myers, C. E., Moore, E. L., Klein, R., Hauser, M. A.,

Postel, E. A., Courtenay, M. D., Schwartz, S. G., Kovach, J. L., Scott, W. K., Liew, G., Tan, A. G., Gopinath, B., Merriam, J. C., Smith, R. T., Khan, J. C., Shahid, H., Moore, A. T., McGrath, J. A., Laux, R., Brantley, M. A., Jr, Agarwal, A., Ersoy, L., Caramoy, A., Langmann, T., Saksens, N. T. M., de Jong, E. K., Hoyng, C. B., Cain, M. S., Richardson, A. J., Martin, T. M., Blangero, J., Weeks, D. E., Dhillon, B., van Duijn, C. M., Doheny, K. F., Romm, J., Klaver, C. C. W., Hayward, C., Gorin, M. B., Klein, M. L., Baird, P. N., den Hollander, A. I., Fauser, S., Yates, J. R. W., Allikmets, R., Wang, J. J., Schaumberg, D. A., Klein, B. E. K., Hagstrom, S. A., Chowers, I., Lotery, A. J., Léveillard, T., Zhang, K., Brilliant, M. H., Hewitt, A. W., Swaroop, A., Chew, E. Y., Pericak-Vance, M. A., DeAngelis, M., Stambolian, D., Haines, J. L., Iyengar, S. K., Weber, B. H. F., Abecasis, G. R., and Heid, I. M. (2016), "A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants," *Nature genetics*, 48, 134–143.

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., Klei, L. L., Kramer, R., Pinto, D., Gümüş, Z. H., Cicek, A. E., Dang, K. K., Browne, A., Lu, C., Xie, L., Readhead, B., Stahl, E. A., Xiao, J., Parvizi, M., Hamamsy, T., Fullard, J. F., Wang, Y.-C., Mahajan, M. C., Derry, J. M. J., Dudley, J. T., Hemby, S. E., Logsdon, B. A., Talbot, K., Raj, T., Bennett, D. A., De Jager, P. L., Zhu, J., Zhang, B., Sullivan, P. F., Chess, A., Purcell, S. M., Shinobu, L. A., Mangravite, L. M., Toyoshiba, H., Gur, R. E., Hahn, C.-G., Lewis, D. A., Haroutunian, V., Peters, M. A., Lipska, B. K., Buxbaum, J. D., Schadt, E. E., Hirai, K., Roeder, K., Brennand, K. J., Katsanis, N., Domenici, E., Devlin, B., and Sklar, P. (2016), "Gene expression elucidates functional impact of polygenic risk for schizophrenia," *Nature neuroscience*, 19, 1442–1453.

Gagliano Taliun, S. A., VandeHaar, P., Boughton, A. P., Welch, R. P., Taliun, D., Schmidt, E. M., Zhou, W., Nielsen, J. B., Willer, C. J., Lee, S., Fritsche, L. G., Boehnke, M., and Abecasis, G. R. (2020), "Exploring and visualizing large-scale genetic associations by using PheWeb," *Nature genetics*, 52, 550–552.

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., Liu, C., White, K. P., CommonMind Consortium, PsychENCODE Consortium, iPSYCH-BROAD Working Group, Horvath, S., and Geschwind, D. H. (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," *Nature neuroscience*, 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., Shieh, A. W., Haney, J., Parhami, S., Belmont, J., Kim, M., Moran Losada, P., Khan, Z., Mleczko, J., Xia, Y., Dai, R., Wang, D., Yang, Y. T., Xu, M., Fish, K., Hof, P. R., Warrell, J., Fitzgerald, D., White, K., Jaffe, A. E., PsychENCODE Consortium, Peters, M. A., Gerstein, M., Liu, C., Iakoucheva, L. M., Pinto, D., and Geschwind, D. H. (2018b), "Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder," *Science*, 362. https://doi.org/10.1126/science.aat8127.

Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F., and Guigó, R. (2021), "Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome," *Nature communications*, 12, 727.

Geihs, M., Yan, Y., Walter, K., Huang, J., Memari, Y., Min, J. L., Mead, D., UK10K Consortium,

Hubbard, T. J., Timpson, N. J., Down, T. A., and Soranzo, N. (2015), "An interactive genome browser of association results from the UK10K cohorts project," *Bioinformatics*, 31, 4029–4031.

Genovese, G., Fromer, M., Stahl, E. A., Ruderfer, D. M., Chambert, K., Landén, M., Moran, J. L., Purcell, S. M., Sklar, P., Sullivan, P. F., Hultman, C. M., and McCarroll, S. A. (2016), "Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia," *Nature neuroscience*, 19, 1433–1441.

Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R., and Smoller, J. W. (2017), "Phenome-wide heritability analysis of the UK Biobank," *PLoS genetics*, 13, e1006711.

Ge, T., Reuter, M., Winkler, A. M., Holmes, A. J., Lee, P. H., Tirrell, L. S., Roffman, J. L., Buckner, R. L., Smoller, J. W., and Sabuncu, M. R. (2016), "Multidimensional heritability analysis of neuroanatomical shape," *Nature communications*, 7, 13291.

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 genetics*, 10, e1004383.

Glantz, L. A., and Lewis, D. A. (2000), "Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia," *Archives of general psychiatry*, 57, 65–73.

Glinos, D. A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K. L., Garimella, K., Bowers, T., Costello, M., Ardlie, K., Jian, R., Tucker, N. R., Ellinor, P. T., Harrington, E. D., Tang, H., Snyder, M., Juul, S., Mohammadi, P., MacArthur, D. G., Lappalainen, T., and Cummings, B. (2021), "Transcriptome variation in human tissues revealed by long-read sequencing," *bioRxiv*. https://doi.org/10.1101/2021.01.22.427687.

Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., and Greenleaf, W. J. (2021), "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis," *Nature genetics*, 53, 403–411.

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O. A., Anney, R., Awashti, S., Belliveau, R., Bettella, F., Buxbaum, J. D., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Christensen, J. H., Churchhouse, C., Dellenvall, K., Demontis, D., De Rubeis, S., Devlin, B., Djurovic, S., Dumont, A. L., Goldstein, J. I., Hansen, C. S., Hauberg, M. E., Hollegaard, M. V., Hope, S., Howrigan, D. P., Huang, H., Hultman, C. M., Klei, L., Maller, J., Martin, J., Martin, A. R., Moran, J. L., Nyegaard, M., Nærland, T., Palmer, D. S., Palotie, A., Pedersen, C. B., Pedersen, M. G., dPoterba, T., Poulsen, J. B., Pourcain, B. S., Qvist, P., Rehnström, K., Reichenberg, A., Reichert, J., Robinson, E. B., Roeder, K., Roussos, P., Saemundsen, E., Sandin, S., Satterstrom, F. K., Davey Smith, G., Stefansson, H., Steinberg, S., Stevens, C. R., Sullivan, P. F., Turley, P., Walters, G. B., Xu, X., Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium, BUPGEN, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, 23andMe Research Team, Stefansson, K., Geschwind, D. H., Nordentoft, M., Hougaard, D. M., Werge, T., Mors, O., Mortensen, P. B., Neale, B. M., Daly, M. J., and Børglum, A. D. (2019), "Identification of common genetic risk variants for autism spectrum disorder," *Nature genetics*, 51, 431–444.

GTEx Consortium, Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., Huang, K. H., Lek, M., Li, X., Nedzel, J. L., Nguyen, D. Y., Noble, M. S., Sullivan, T.

J., Tukiainen, T., MacArthur, D. G., Getz, G., Addington, A., Guan, P., Koester, S., Little, A. R., Lockhart, N. C., Moore, H. M., Rao, A., Struewing, J. P., Volpi, S., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B. A., Gillard, B. M., Karasik, E., Kumar, R., Miklos, M., Moser, M. T., Jewell, S. D., Montroy, R. G., Rohrer, D. C., Valley, D., Mash, D. C., Davis, D. A., Sobin, L., Barcus, M. E., Branton, P. A., Abell, N. S., Balliu, B., Delaneau, O., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Han, B., He, A. Z., Hormozdiari, F., Li, X., Liu, B., Kang, E. Y., McDowell, I. C., Ongen, H., Palowitch, J. J., Peterson, C. B., Quon, G., Ripke, S., Saha, A., Shabalin, A. A., Shimko, T. C., Sul, J. H., Teran, N. A., Tsang, E. K., Zhang, H., Zhou, Y.-H., Bustamante, C. D., Cox, N. J., Guigó, R., Kellis, M., McCarthy, M. I., Conrad, D. F., Eskin, E., Li, G., Nobel, A. B., Sabatti, C., Stranger, B. E., Wen, X., Wright, F. A., Ardlie, K. G., Dermitzakis, E. T., Lappalainen, T., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Karczewski, K. J., Lek, M., Li, X., MacArthur, D. G., Nedzel, J. L., Nguyen, D. T., Noble, M. S., Segrè, A. V., Trowbridge, C. A., Tukiainen, T., Abell, N. S., Balliu, B., Barshir, R., Basha, O., Battle, A., Bogu, G. K., Brown, A., Brown, C. D., Castel, S. E., Chen, L. S., Chiang, C., Conrad, D. F., Cox, N. J., Damani, F. N., Davis, J. R., Delaneau, O., Dermitzakis, E. T., Engelhardt, B. E., Eskin, E., Ferreira, P. G., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Guigo, R., Hall, I. M., Han, B., He, Y., Hormozdiari, F., Howald, C., Kyung Im, H., Jo, B., Yong Kang, E., Kim, Y., Kim-Hellmuth, S., Lappalainen, T., Li, G., Li, X., Liu, B., Mangul, S., McCarthy, M. I., McDowell, I. C., Mohammadi, P., Monlong, J., Montgomery, S. B., Muñoz-Aguirre, M., Ndungu, A. W., Nicolae, D. L., Nobel, A. B., Oliva, M., Ongen, H., Palowitch, J. J., Panousis, N., Papasaikas, P., Park, Y., Parsana, P., Payne, A. J., Peterson, C. B., Quan, J., Reverter, F., Sabatti, C., Saha, A., Sammeth, M., Scott, A. J., Shabalin, A. A., Sodaei, R., Stephens, M., Stranger, B. E., Strober, B. J., Sul, J. H., Tsang, E. K., Urbut, S., van de Bunt, M., Wang, G., Wen, X., Wright, F. A., Xi, H. S., Yeger-Lotem, E., Zappala, Z., Zaugg, J. B., Zhou, Y.-H., Akey, J. M., Bates, D., Chan, J., Chen, L. S., Claussnitzer, M., Demanelis, K., Diegel, M., Doherty, J. A., Feinberg, A. P., Fernando, M. S., Halow, J., Hansen, K. D., Haugen, E., Hickey, P. F., Hou, L., Jasmine, F., Jian, R., Jiang, L., Johnson, A., Kaul, R., Kellis, M., Kibriya, M. G., Lee, K., Billy Li, J., Li, Q., Li, X., Lin, J., Lin, S., Linder, S., Linke, C., Liu, Y., Maurano, M. T., Molinie, B., Montgomery, S. B., Nelson, J., Neri, F. J., Oliva, M., Park, Y., Pierce, B. L., Rinaldi, N. J., Rizzardi, L. F., Sandstrom, R., Skol, A., Smith, K. S., Snyder, M. P., Stamatoyannopoulos, J., Stranger, B. E., Tang, H., Tsang, E. K., Wang, L., Wang, M., Van Wittenberghe, N., Wu, F., Zhang, R., Nierras, C. R., Branton, P. A., Carithers, L. J., Guan, P., Moore, H. M., Rao, A., Vaught, J. B., Gould, S. E., Lockart, N. C., Martin, C., Struewing, J. P., Volpi, S., Addington, A. M., Koester, S. E., Little, A. R., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B. A., Gillard, B. M., Karasik, E., Kumar, R., Miklos, M., Moser, M. T., Jewell, S. D., Montroy, R. G., Rohrer, D. C., Valley, D. R., Davis, D. A., Mash, D. C., Undale, A. H., Smith, A. M., Tabor, D. E., Roche, N. V., McLean, J. A., Vatanian, N., Robinson, K. L., Sobin, L., Barcus, M. E., Valentino, K. M., Qi, L., Hunter, S., Hariharan, P., Singh, S., Um, K. S., Matose, T., Tomaszewski, M. M., Barker, L. K., Mosavel, M., Siminoff, L. A., Traino, H. M., Flicek, P., Juettemann, T., Ruffier, M., Sheppard, D., Taylor, K., Trevanion, S. J., Zerbino, D. R., Craft, B., Goldman, M., Haeussler, M., Kent, W. J., Lee, C. M., Paten, B., Rosenbloom, K. R., Vivian, J., and Zhu, J. (2017), "Genetic effects on gene expression across human tissues," *Nature*, 550, 204–213.

GTEx Consortium (2020), "The GTEx Consortium atlas of genetic regulatory effects across human tissues," *Science*, 369, 1318–1330.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P., and Pasaniuc, B. (2016), "Integrative approaches for large-scale transcriptome-wide association studies," *Nature genetics*, 48, 245–252.

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., and McCarroll, S. A. (2015), "Large multiallelic copy number variations in humans," *Nature genetics*, 47, 296–303.

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., Troakes, C., Turecki, G., O'Donovan, M. C., Schalkwyk, L. C., Bray, N. J., and Mill, J. (2016), "Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci," *Nature neuroscience*, 19, 48–54.

Han, Y.-G., Spassky, N., Romaguera-Ros, M., Garcia-Verdugo, J.-M., Aguilar, A., Schneider-Maunoury, S., and Alvarez-Buylla, A. (2008), "Hedgehog signaling and primary cilia are required for the formation of adult neural stem cells," *Nature neuroscience*, 11, 277–284.

Hernandez, L. M., Kim, M., Hoftman, G. D., Haney, J. R., de la Torre-Ubieta, L., Pasaniuc, B., and Gandal, M. J. (2021), "Transcriptomic Insight Into the Polygenic Mechanisms Underlying Psychiatric Disorders," *Biological psychiatry*, 89, 54–64.

Hivert, V., Sidorenko, J., Rohart, F., Goddard, M. E., Yang, J., Wray, N. R., Yengo, L., and Visscher, P. M. (2021), "Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals," *American journal of human genetics*, 108, 786–798.

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., Casper, T., Cobbs, C., Dalley, R., Dee, N., Ding, S.-L., Ellenbogen, R. G., Fong, O., Garren, E., Goldy, J., Gwinn, R. P., Hirschstein, D., Keene, C. D., Keshk, M., Ko, A. L., Lathia, K., Mahfouz, A., Maltzer, Z., McGraw, M., Nguyen, T. N., Nyhus, J., Ojemann, J. G., Oldre, A., Parry, S., Reynolds, S., Rimorin, C., Shapovalova, N. V., Somasundaram, S., Szafer, A., Thomsen, E. R., Tieu, M., Quon, G., Scheuermann, R. H., Yuste, R., Sunkin, S. M., Lelieveldt, B., Feng, D., Ng, L., Bernard, A., Hawrylycz, M., Phillips, J. W., Tasic, B., Zeng, H., Jones, A. R., Koch, C., and Lein, E. S. (2019), "Conserved cell types with divergent features in human versus mouse cortex," *Nature*, 573, 61–68.

Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., Lush, M. J., Povey, S., Talbot, C. C., Jr, Wright, M. W., Wain, H. M., Trowsdale, J., Ziegler, A., and Beck, S. (2004), "Gene map of the extended human MHC," *Nature reviews genetics*, 5, 889–899.

Hyman, S. E. (2018), "The daunting polygenicity of mental illness: making a new map," *Philosophical transactions of the royal society b,* 373. https://doi.org/10.1098/rstb.2017.0031.

Iancu, O. D., Darakjian, P., Malmanger, B., Walter, N. A. R., McWeeney, S., and Hitzemann, R. (2012), "Gene networks and haloperidol-induced catalepsy," *Genes, brain, and behavior*, 11, 29–37.

International Multiple Sclerosis Genetics Consortium (IMSGC), Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., Cotsapas, C., Shah, T. S., Spencer, C., Booth, D., Goris, A., Oturai, A., Saarela, J., Fontaine, B., Hemmer, B., Martin, C., Zipp, F., D'Alfonso, S., Martinelli-Boneschi, F., Taylor, B., Harbo, H. F., Kockum, I., Hillert, J., Olsson, T., Ban, M., Oksenberg, J. R.,

Hintzen, R., Barcellos, L. F., Wellcome Trust Case Control Consortium 2 (WTCCC2), International IBD Genetics Consortium (IIBDGC), Agliardi, C., Alfredsson, L., Alizadeh, M., Anderson, C., Andrews, R., Søndergaard, H. B., Baker, A., Band, G., Baranzini, S. E., Barizzone, N., Barrett, J., Bellenguez, C., Bergamaschi, L., Bernardinelli, L., Berthele, A., Biberacher, V., Binder, T. M. C., Blackburn, H., Bomfim, I. L., Brambilla, P., Broadley, S., Brochet, B., Brundin, L., Buck, D., Butzkueven, H., Caillier, S. J., Camu, W., Carpentier, W., Cavalla, P., Celius, E. G., Coman, I., Comi, G., Corrado, L., Cosemans, L., Cournu-Rebeix, I., Cree, B. A. C., Cusi, D., Damotte, V., Defer, G., Delgado, S. R., Deloukas, P., di Sapio, A., Dilthey, A. T., Donnelly, P., Dubois, B., Duddy, M., Edkins, S., Elovaara, I., Esposito, F., Evangelou, N., Fiddes, B., Field, J., Franke, A., Freeman, C., Frohlich, I. Y., Galimberti, D., Gieger, C., Gourraud, P.-A., Graetz, C., Graham, A., Grummel, V., Guaschino, C., Hadjixenofontos, A., Hakonarson, H., Halfpenny, C., Hall, G., Hall, P., Hamsten, A., Harley, J., Harrower, T., Hawkins, C., Hellenthal, G., Hillier, C., Hobart, J., Hoshi, M., Hunt, S. E., Jagodic, M., Jelčić, I., Jochim, A., Kendall, B., Kermode, A., Kilpatrick, T., Koivisto, K., Konidari, I., Korn, T., Kronsbein, H., Langford, C., Larsson, M., Lathrop, M., Lebrun-Frenay, C., Lechner-Scott, J., Lee, M. H., Leone, M. A., Leppä, V., Liberatore, G., Lie, B. A., Lill, C. M., Lindén, M., Link, J., Luessi, F., Lycke, J., Macciardi, F., Männistö, S., Manrique, C. P., Martin, R., Martinelli, V., Mason, D., Mazibrada, G., McCabe, C., Mero, I.-L., Mescheriakova, J., Moutsianas, L., Myhr, K.-M., Nagels, G., Nicholas, R., Nilsson, P., Piehl, F., Pirinen, M., Price, S. E., Quach, H., Reunanen, M., Robberecht, W., Robertson, N. P., Rodegher, M., Rog, D., Salvetti, M., Schnetz-Boutaud, N. C., Sellebjerg, F., Selter, R. C., Schaefer, C., Shaunak, S., Shen, L., Shields, S., Siffrin, V., Slee, M., Sorensen, P. S., Sorosina, M., Sospedra, M., Spurkland, A., Strange, A., Sundqvist, E., Thijs, V., Thorpe, J., Ticca, A., Tienari, P., van Duijn, C., Visser, E. M., Vucic, S., Westerlind, H., Wiley, J. S., Wilkins, A., Wilson, J. F., Winkelmann, J., Zajicek, J., Zindler, E., Haines, J. L., Pericak-Vance, M. A., Ivinson, A. J., Stewart, G., Hafler, D., Hauser, S. L., Compston, A., McVean, G., De Jager, P., Sawcer, S. J., and McCauley, J. L. (2013), "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis," *Nature genetics*, 45, 1353–1360.

International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS) (2018), "Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis," *Molecular psychiatry*, 23, 1181–1188.

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., Colantuoni, C., Ulrich, W. S., Maher, B. J., Deep-Soboslay, A., BrainSeq Consortium, Cross, A. J., Brandon, N. J., Leek, J. T., Hyde, T. M., Kleinman, J. E., and Weinberger, D. R. (2018), "Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis," *Nature neuroscience*, 21, 1117–1125.

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witoelar, A., Stringer, S., Aarsland, D., Almdahl, I. S., Andersen, F., Bergh, S., Bettella, F., Bjornsson, S., Brækhus, A., Bråthen, G., de Leeuw, C., Desikan, R. S., Djurovic, S., Dumitrescu, L., Fladby, T., Hohman, T. J., Jonsson, P. V., Kiddle, S. J., Rongve, A., Saltvedt, I., Sando, S. B., Selbæk, G., Shoai, M., Skene, N. G., Snaedal, J., Stordal, E., Ulstein, I. D., Wang, Y., White, L. R., Hardy, J., Hjerling-Leffler, J., Sullivan, P. F., van der Flier, W. M., Dobson, R., Davis, L. K., Stefansson, H., Stefansson, K., Pedersen, N. L., Ripke, S., Andreassen, O. A., and Posthuma, D. (2019), "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk," *Nature genetics*, 51, 404–413.

Jones, H. J., Gage, S. H., Heron, J., Hickman, M., Lewis, G., Munafò, M. R., and Zammit, S. (2018), "Association of Combined Patterns of Tobacco and Cannabis Use in Adolescence With Psychotic

Experiences," *JAMA psychiatry*, 75, 240–246.

Jorgenson, E., Kvale, M., and Witte, J. S. (2009), "VALID: visualization of association study results and linkage disequilibrium," *Genetic epidemiology*, 33, 599–603.

Juliusdottir, T., Banasik, K., Robertson, N. R., Mott, R., and McCarthy, M. I. (2018), "Toppar: an interactive browser for viewing association study results," *Bioinformatics*, 34, 1922–1924.

Kamitaki, N., Sekar, A., Handsaker, R. E., de Rivera, H., Tooley, K., Morris, D. L., Taylor, K. E., Whelan, C. W., Tombleson, P., Olde Loohuis, L. M., Boehnke, M., Kimberly, R. P., Kaufman, K. M., Harley, J. B., Langefeld, C. D., Seidman, C. E., Pato, M. T., Pato, C. N., Ophoff, R. A., Graham, R. R., Criswell, L. A., Vyse, T. J., and McCarroll, S. A. (2020), "Complement genes contribute sex-biased vulnerability in diverse disorders," *Nature*, 582, 577–581.

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., McRae, J. F., Short, P. J., Torene, R. I., de Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., Reijnders, M. R. F., Yeung, A., Yntema, H. G., Borras, S., Clark, C., Dean, J., Miedzybrodzka, Z., Ross, A., Tennant, S., Dabir, T., Donnelly, D., Humphreys, M., Magee, A., McConnell, V., McKee, S., McNerlan, S., Morrison, P. J., Rea, G., Stewart, F., Cole, T., Cooper, N., Cooper-Charles, L., Cox, H., Islam, L., Jarvis, J., Keelagher, R., Lim, D., McMullan, D., Morton, J., Naik, S., O'Driscoll, M., Ong, K.-R., Osio, D., Ragge, N., Turton, S., Vogt, J., Williams, D., Bodek, S., Donaldson, A., Hills, A., Low, K., Newbury-Ecob, R., Norman, A. M., Roberts, E., Scurr, I., Smithson, S., Tooley, M., Abbs, S., Armstrong, R., Dunn, C., Holden, S., Park, S.-M., Paterson, J., Raymond, L., Reid, E., Sandford, R., Simonic, I., Tischkowitz, M., Woods, G., Bradley, L., Comerford, J., Green, A., Lynch, S., McQuaid, S., Mullaney, B., Berg, J., Goudie, D., Mavrak, E., McLean, J., McWilliam, C., Reavey, E., Azam, T., Cleary, E., Jackson, A., Lam, W., Lampe, A., Moore, D., Porteous, M., Baple, E., Baptista, J., Brewer, C., Castle, B., Kivuva, E., Owens, M., Rankin, J., Shaw-Smith, C., Turner, C., Turnpenny, P., Tysoe, C., Bradley, T., Davidson, R., Gardiner, C., Joss, S., Kinning, E., Longman, C., McGowan, R., Murday, V., Pilz, D., Tobias, E., Whiteford, M., Williams, N., Barnicoat, A., Clement, E., Faravelli, F., Hurst, J., Jenkins, L., Jones, W., Kumar, V. K. A., Lees, M., Loughlin, S., Male, A., Morrogh, D., Rosser, E., Scott, R., Wilson, L., Beleza, A., Deshpande, C., Flinter, F., Holder, M., Irving, M., Izatt, L., Josifova, D., Mohammed, S., Molenda, A., Robert, L., Roworth, W., Ruddy, D., Ryten, M., Yau, S., Bennett, C., Blyth, M., Campbell, J., Coates, A., Dobbie, A., Hewitt, S., Hobson, E., Jackson, E., Jewell, R., Kraus, A., Prescott, K., Sheridan, E., Thomson, J., Bradshaw, K., Dixit, A., Eason, J., Haines, R., Harrison, R., Mutch, S., Sarkar, A., Searle, C., Shannon, N., Sharif, A., Suri, M., Vasudevan, P., Canham, N., Ellis, I., Greenhalgh, L., Howard, E., Stinton, V., Swale, A., Weber, A., Banka, S., Breen, C., Briggs, T., Burkitt-Wright, E., Chandler, K., Clayton-Smith, J., Donnai, D., Douzgou, S., Gaunt, L., Jones, E., Kerr, B., Langley, C., Metcalfe, K., Smith, A., Wright, R., Bourn, D., Burn, J., Fisher, R., Hellens, S., Henderson, A., Montgomery, T., Splitt, M., Straub, V., Wright, M., Zwolinski, S., Allen, Z., Bernhard, B., Brady, A., Brooks, C., Busby, L., Clowes, V., Ghali, N., Holder, S., Ibitoye, R., Wakeling, E., Blair, E., Carmichael, J., Cilliers, D., Clasper, S., Gibbons, R., Kini, U., Lester, T., Nemeth, A., Poulton, J., Price, S., Shears, D., Stewart, H., Wilkie, A., Albaba, S., Baker, D., Balasubramanian, M., Johnson, D., Parker, M., Quarrell, O., Stewart, A., Willoughby, J., Crosby, C., Elmslie, F., Homfray, T., Jin, H., Lahiri, N., Mansour, S., Marks, K., McEntagart, M., Saggar, A., Tatton-Brown, K., Butler, R., Clarke, A., Corrin, S., Fry, A., Kamath, A., McCann, E., Mugalaasi, H., Pottinger, C., Procter, A., Sampson, J., Sansbury, F., Varghese, V., Baralle, D., Callaway, A., Cassidy, E. J., Daniels, S., Douglas, A., Foulds, N., Hunt, D., Kharbanda, M., Lachlan, K., Mercer, C., Side, L., Temple, I. K., Wellesley, D., Vissers, L. E. L. M., Juusola, J., Wright, C. F., Brunner, H. G., Firth, H. V., FitzPatrick, D. R., Barrett, J. C., Hurles, M. E., Gilissen, C., Retterer, K., and Deciphering Developmental Disorders Study (2020), "Evidence for 28 genetic disorders

discovered by combining healthcare and research data," *Nature*, 586, 757–762.

Karama, S., Ducharme, S., Corley, J., Chouinard-Decorte, F., Starr, J. M., Wardlaw, J. M., Bastin, M. E., and Deary, I. J. (2015), "Cigarette smoking and thinning of the brain's cortex," *Molecular psychiatry*, 20, 778–785.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010), "Analysis and design of RNA sequencing experiments for identifying isoform regulation," *Nature methods*, 7, 1009–1015.

Kendler, K. S., Lönn, S. L., Sundquist, J., and Sundquist, K. (2015), "Smoking and schizophrenia in population cohorts of Swedish women and men: a prospective co-relative control study," *The American journal of psychiatry*, 172, 1092–1100.

Keshavan, M. S., Anderson, S., and Pettegrew, J. W. (1994), "Is schizophrenia due to excessive synaptic pruning in the prefrontal cortex? The Feinberg hypothesis revisited," *Journal of psychiatric research*, 28, 239–265.

Kim, J., Shen, J., Wang, A., Mehrotra, D. V., Ko, S., Zhou, J. J., and Zhou, H. (2021), "VCSEL: Prioritizing SNP-set by penalized, variance component selection," *Annals of applied statistics*, 15, 1652–1672.

Kim, M., Haney, J. R., Zhang, P., Hernandez, L. M., Wang, L.-K., Perez-Cano, L., Loohuis, L. M. O., de la Torre-Ubieta, L., and Gandal, M. J. (2021), "Brain gene co-expression networks link complement signaling with convergent synaptic pathology in schizophrenia," *Nature neuroscience*, 24, 799–809.

Kim, M., Jops, C. T., and Gandal, M. J. (2022), "GeneticsMakie.jl: A versatile and scalable toolkit for visualizing locus-level genetic and genomic data," *bioRxiv*. https://doi.org/10.1101/2022.04.18.488573.

Koopmans, F., van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijsouw, T., Coba, M. P., Niels Cornelisse, L., Farrell, R. J., Goldschmidt, H. L., Howrigan, D. P., Hussain, N. K., Imig, C., de Jong, A. P. H., Jung, H., Kohansalnodehi, M., Kramarz, B., Lipstein, N., Lovering, R. C., MacGillavry, H., Mariano, V., Mi, H., Ninov, M., Osumi-Sutherland, D., Pielot, R., Smalla, K.-H., Tang, H., Tashman, K., Toonen, R. F. G., Verpelli, C., Reig-Viader, R., Watanabe, K., van Weering, J., Achsel, T., Ashrafi, G., Asi, N., Brown, T. C., De Camilli, P., Feuermann, M., Foulger, R. E., Gaudet, P., Joglekar, A., Kanellopoulos, A., Malenka, R., Nicoll, R. A., Pulido, C., de Juan-Sanz, J., Sheng, M., Südhof, T. C., Tilgner, H. U., Bagni, C., Bayés, À., Biederer, T., Brose, N., Chua, J. J. E., Dieterich, D. C., Gundelfinger, E. D., Hoogenraad, C., Huganir, R. L., Jahn, R., Kaeser, P. S., Kim, E., Kreutz, M. R., McPherson, P. S., Neale, B. M., O'Connor, V., Posthuma, D., Ryan, T. A., Sala, C., Feng, G., Hyman, S. E., Thomas, P. D., Smit, A. B., and Verhage, M. (2019), "SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse," *Neuron*, 103, 217–234.

Kramer, N. E., Davis, E. S., Wenger, C. D., Deoudes, E. M., Parker, S. M., Love, M. I., and Phanstiel, D. H. (2022), "Plotgardener: Cultivating precise multi-panel figures in R," *Bioinformatics*, 38, 2042–2045.

Kwong, A., Boughton, A. P., Wang, M., VandeHaar, P., Boehnke, M., Abecasis, G., and Kang, H. M. (2021), "FIVEx: an interactive eQTL browser across public datasets," *Bioinformatics*, 38, 559–561.

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., and Zhang, K. (2018), "Integrative single-cell analysis of transcriptional and

epigenetic states in the human adult brain," *Nature biotechnology*, 36, 70–80.

Lam, M., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C., Liu, R., Zhou, W., Guan, L., Kamatani, Y., Kim, S.-W., Kubo, M., Kusumawardhani, A. A. A. A., Liu, C.-M., Ma, H., Periyasamy, S., Takahashi, A., Xu, Z., Yu, H., Zhu, F., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Indonesia Schizophrenia Consortium, Genetic REsearch on schizophreniA neTwork-China and the Netherlands (GREAT-CN), Chen, W. J., Faraone, S., Glatt, S. J., He, L., Hyman, S. E., Hwu, H.-G., McCarroll, S. A., Neale, B. M., Sklar, P., Wildenauer, D. B., Yu, X., Zhang, D., Mowry, B. J., Lee, J., Holmans, P., Xu, S., Sullivan, P. F., Ripke, S., O'Donovan, M. C., Daly, M. J., Qin, S., Sham, P., Iwata, N., Hong, K. S., Schwab, S. G., Yue, W., Tsuang, M., Liu, J., Ma, X., Kahn, R. S., Shi, Y., and Huang, H. (2019), "Comparative genetic architectures of schizophrenia in East Asian and European populations," *Nature genetics*, 51, 1670–1678.

Lea, A., Subramaniam, M., Ko, A., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Mononen, N., Raitakari, O. T., Ala-Korpela, M., Pajukanta, P., Zaitlen, N., and Ayroles, J. F. (2019), "Genetic and environmental perturbations lead to regulatory decoherence," *eLife*, 8. https://doi.org/10.7554/eLife.40538.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., Yengo, L., 23andMe Research Team, COGENT (Cognitive Genomics Consortium), Social Science Genetic Association Consortium, Alver, M., Bao, Y., Clark, D. W., Day, F. R., Furlotte, N. A., Joshi, P. K., Kemper, K. E., Kleinman, A., Langenberg, C., Mägi, R., Trampush, J. W., Verma, S. S., Wu, Y., Lam, M., Zhao, J. H., Zheng, Z., Boardman, J. D., Campbell, H., Freese, J., Harris, K. M., Hayward, C., Herd, P., Kumari, M., Lencz, T., Luan, J. 'an, Malhotra, A. K., Metspalu, A., Milani, L., Ong, K. K., Perry, J. R. B., Porteous, D. J., Ritchie, M. D., Smart, M. C., Smith, B. H., Tung, J. Y., Wareham, N. J., Wilson, J. F., Beauchamp, J. P., Conley, D. C., Esko, T., Lehrer, S. F., Magnusson, P. K. E., Oskarsson, S., Pers, T. H., Robinson, M. R., Thom, K., Watson, C., Chabris, C. F., Meyer, M. N., Laibson, D. I., Yang, J., Johannesson, M., Koellinger, P. D., Turley, P., Visscher, P. M., Benjamin, D. J., and Cesarini, D. (2018), "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals," *Nature genetics*, 50, 1112–1121.

Lee, S. H., and van der Werf, J. H. J. (2016), "MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information," *Bioinformatics*, 32, 1420–1422.

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012), "Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood," *Bioinformatics*, 28, 2540–2542.

de Leeuw, C. A., Mooij, J. M., Heskes, T., and Posthuma, D. (2015), "MAGMA: generalized gene-set analysis of GWAS data," *PLoS computational biology*, 11, e1004219.

Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., Dempster, E. L., Bray, N. J., O'Neill, P., Tseng, E., Ahmed, Z., Collier, D. A., Jeffery, E. D., Prabhakar, S., Schalkwyk, L., Jops, C., Gandal, M. J., Sheynkman, G. M., Hannon, E., and Mill, J. (2021), "Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing," *Cell reports*, 37, 110022.

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.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015), "The Molecular Signatures Database (MSigDB) hallmark gene set collection," *Cell systems*, 1, 417–425.

Lieberman, J. (2018), "Unveiling the RNA World," *The New England journal of medicine*, 379, 1278–1280.

Li, L., Huang, K.-L., Gao, Y., Cui, Y., Wang, G., Elrod, N. D., Li, Y., Chen, Y. E., Ji, P., Peng, F., Russell, W. K., Wagner, E. J., and Li, W. (2021), "An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability," *Nature genetics*, 53, 994–1005.

Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., Sunkin, S. M., Li, Z., Shin, Y., Zhu, Y., Sousa, A. M. M., Werling, D. M., Kitchen, R. R., Kang, H. J., Pletikos, M., Choi, J., Muchnik, S., Xu, X., Wang, D., Lorente-Galdos, B., Liu, S., Giusti-Rodríguez, P., Won, H., de Leeuw, C. A., Pardiñas, A. F., BrainSpan Consortium, PsychENCODE Consortium, PsychENCODE Developmental Subgroup, Hu, M., Jin, F., Li, Y., Owen, M. J., O'Donovan, M. C., Walters, J. T. R., Posthuma, D., Reimers, M. A., Levitt, P., Weinberger, D. R., Hyde, T. M., Kleinman, J. E., Geschwind, D. H., Hawrylycz, M. J., State, M. W., Sanders, S. J., Sullivan, P. F., Gerstein, M. B., Lein, E. S., Knowles, J. A., and Sestan, N. (2018), "Integrative functional genomic analysis of human brain development and neuropsychiatric risks," *Science*, 362. https://doi.org/10.1126/science.aat7615.

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C. T., Rigina, O., Rapacki, K., Stærfeldt, H. H., Brunak, S., Jensen, T. S., and Lage, K. (2017), "A scored human protein-protein interaction network to catalyze genomic interpretation," *Nature methods*, 14, 61–64.

Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E., and Montgomery, S. B. (2019), "Abundant associations with gene expression complicate GWAS follow-up," *Nature genetics*, 51, 768–769.

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., Abedian, S., Cheon, J. H., Cho, J., Dayani, N. E., Franke, L., Fuyuno, Y., Hart, A., Juyal, R. C., Juyal, G., Kim, W. H., Morris, A. P., Poustchi, H., Newman, W. G., Midha, V., Orchard, T. R., Vahedi, H., Sood, A., Sung, J. Y., Malekzadeh, R., Westra, H.-J., Yamazaki, K., Yang, S.-K., International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, Barrett, J. C., Alizadeh, B. Z., Parkes, M., Bk, T., Daly, M. J., Kubo, M., Anderson, C. A., and Weersma, R. K. (2015), "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations," *Nature genetics*, 47, 979–986.

Liu, X., Finucane, H. K., Gusev, A., Bhatia, G., Gazal, S., O'Connor, L., Bulik-Sullivan, B., Wright, F. A., Sullivan, P. F., Neale, B. M., and Price, A. L. (2017), "Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues," *American journal of human genetics*, 100, 605–616.

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

Li, Y. I., Wong, G., Humphrey, J., and Raj, T. (2019), "Prioritizing Parkinson's disease genes using

population-scale transcriptomic data," *Nature communications*, 10, 994.

Lloyd-Jones, L. R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., Gibson, G., Spector, T., Montgomery, G., Esko, T., Visscher, P. M., and Powell, J. E. (2017), "The Genetic Architecture of Gene Expression in Peripheral Blood," *American journal of human genetics*, 100, 228–237.

Lopes, K. de P., Snijders, G. J. L., Humphrey, J., Allan, A., Sneeboer, M. A. M., Navarro, E., Schilder, B. M., Vialle, R. A., Parks, M., Missall, R., van Zuiden, W., Gigase, F. A. J., Kübler, R., van Berlekom, A. B., Hicks, E. M., Böttcher, C., Priller, J., Kahn, R. S., de Witte, L. D., and Raj, T. (2022), "Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies," *Nature genetics*, 54, 4–17.

Lukowski, S. W., Lloyd-Jones, L. R., Holloway, A., Kirsten, H., Hemani, G., Yang, J., Small, K., Zhao, J., Metspalu, A., Dermitzakis, E. T., Gibson, G., Spector, T. D., Thiery, J., Scholz, M., Montgomery, G. W., Esko, T., Visscher, P. M., and Powell, J. E. (2017), "Genetic correlations reveal the shared genetic architecture of transcription in human peripheral blood," *Nature communications*, 8, 483.

Lyon, M. S., Andrews, S. J., Elsworth, B., Gaunt, T. R., Hemani, G., and Marcora, E. (2021), "The variant call format provides efficient and robust storage of GWAS summary statistics," *Genome biology*, 22, 32.

MacDonald, M. L., Alhassan, J., Newman, J. T., Richard, M., Gu, H., Kelly, R. M., Sampson, A. R., Fish, K. N., Penzes, P., Wills, Z. P., Lewis, D. A., and Sweet, R. A. (2017), "Selective Loss of Smaller Spines in Schizophrenia," *The American journal of psychiatry*, 174, 586–594.

Machiela, M. J., and Chanock, S. J. (2015), "LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants," *Bioinformatics*, 31, 3555–3557.

Machiela, M. J., and Chanock, S. J. (2018), "LDassoc: an online tool for interactively exploring genome-wide association study results and prioritizing variants for functional investigation," *Bioinformatics*, 34, 887–889.

Mah, W., and Won, H. (2019), "The three-dimensional landscape of the genome in human brain tissue unveils regulatory mechanisms leading to schizophrenia risk," *Schizophrenia research*, 217, 17–25.

Marley, A., and von Zastrow, M. (2012), "A simple cell-based assay reveals that diverse neuropsychiatric risk genes converge on primary cilia," *PloS one*, 7, e46647.

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fajarado, K. V. F., Maile, M. S., Ripke, S., Agartz, I., Albus, M., Alexander, M., Amin, F., Atkins, J., Bacanu, S. A., Belliveau, R. A., Jr, Bergen, S. E., Bertalan, M., Bevilacqua, E., Bigdeli, T. B., Black, D. W., Bruggeman, R., Buccola, N. G., Buckner, R. L., Bulik-Sullivan, B., Byerley, W., Cahn, W., Cai, G., Cairns, M. J., Campion, D., Cantor, R. M., Carr, V. J., Carrera, N., Catts, S. V., Chambert, K. D., Cheng, W., Cloninger, C. R., Cohen, D., Cormican, P., Craddock, N., Crespo-Facorro, B., Crowley, J. J., Curtis, D., Davidson, M., Davis, K. L., Degenhardt, F., Del Favero, J., DeLisi, L. E., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A. H., Farh, K.-H., Farrell, M. S., Frank, J., Franke, L., Freedman, R., Freimer, N. B., Friedman, J. I., Forstner, A. J., Fromer,

M., Genovese, G., Georgieva, L., Gershon, E. S., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J. I., Gratten, J., de Haan, L., Hamshere, M. L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A. M., Henskens, F. A., Herms, S., Hirschhorn, J. N., Hoffmann, P., Hofman, A., Huang, H., Ikeda, M., Joa, I., Kähler, A. K., Kahn, R. S., Kalaydjieva, L., Karjalainen, J., Kavanagh, D., Keller, M. C., Kelly, B. J., Kennedy, J. L., Kim, Y., Knowles, J. A., Konte, B., Laurent, C., Lee, P., Lee, S. H., Legge, S. E., Lerer, B., Levy, D. L., Liang, K.-Y., Lieberman, J., Lönnqvist, J., Loughland, C. M., Magnusson, P. K. E., Maher, B. S., Maier, W., Mallet, J., Mattheisen, M., Mattingsdal, M., McCarley, R. W., McDonald, C., McIntosh, A. M., Meier, S., Meijer, C. J., Melle, I., Mesholam-Gately, R. I., Metspalu, A., Michie, P. T., Milani, L., Milanova, V., Mokrab, Y., Morris, D. W., Müller-Myhsok, B., Murphy, K. C., Murray, R. M., Myin-Germeys, I., Nenadic, I., Nertney, D. A., Nestadt, G., Nicodemus, K. K., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., Oh, S.-Y., Olincy, A., Olsen, L., O'Neill, F. A., Van Os, J., Pantelis, C., Papadimitriou, G. N., Parkhomenko, E., Pato, M. T., Paunio, T., Psychosis Endophenotypes International Consortium, Perkins, D. O., Pers, T. H., Pietiläinen, O., Pimm, J., Pocklington, A. J., Powell, J., Price, A., Pulver, A. E., Purcell, S. M., Quested, D., Rasmussen, H. B., Reichenberg, A., Reimers, M. A., Richards, A. L., Roffman, J. L., Roussos, P., Ruderfer, D. M., Salomaa, V., Sanders, A. R., Savitz, A., Schall, U., Schulze, T. G., Schwab, S. G., Scolnick, E. M., Scott, R. J., Seidman, L. J., Shi, J., Silverman, J. M., Smoller, J. W., Söderman, E., Spencer, C. C. A., Stahl, E. A., Strengman, E., Strohmaier, J., Stroup, T. S., Suvisaari, J., Svrakic, D. M., Szatkiewicz, J. P., Thirumalai, S., Tooney, P. A., Veijola, J., Visscher, P. M., Waddington, J., Walsh, D., Webb, B. T., Weiser, M., Wildenauer, D. B., Williams, N. M., Williams, S., Witt, S. H., Wolen, A. R., Wormley, B. K., Wray, N. R., Wu, J. Q., Zai, C. C., Adolfsson, R., Andreassen, O. A., Blackwood, D. H. R., Bramon, E., Buxbaum, J. D., Cichon, S., Collier, D. A., Corvin, A., Daly, M. J., Darvasi, A., Domenici, E., Esko, T., Gejman, P. V., Gill, M., Gurling, H., Hultman, C. M., Iwata, N., Jablensky, A. V., Jönsson, E. G., Kendler, K. S., Kirov, G., Knight, J., Levinson, D. F., Li, Q. S., McCarroll, S. A., McQuillin, A., Moran, J. L., Mowry, B. J., Nöthen, M. M., Ophoff, R. A., Owen, M. J., Palotie, A., Pato, C. N., Petryshen, T. L., Posthuma, D., Rietschel, M., Riley, B. P., Rujescu, D., Sklar, P., St Clair, D., Walters, J. T. R., Werge, T., Sullivan, P. F., O'Donovan, M. C., Scherer, S. W., Neale, B. M., Sebat, J., and CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium (2017), "Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects," *Nature genetics*, 49, 27–35.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004), "The fine-scale structure of recombination rate variation in the human genome," *Science*, 304, 581–584.

Mittleman, B. E., Pott, S., Warland, S., Zeng, T., Mu, Z., Kaur, M., Gilad, Y., and Li, Y. (2020), "Alternative polyadenylation mediates genetic regulation of gene expression," *eLife*, 9. https://doi.org/10.7554/eLife.57492.

Molenberghs, G., and Verbeke, G. (2007), "Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space," *The American statistician*, 61, 22–27.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., Luan, J. 'an, Lindgren, C. M., Müller-Nurasyid, M., Pechlivanis, S., Rayner, N. W., Scott, L. J., Wiltshire, S., Yengo, L., Kinnunen, L., Rossin, E. J., Raychaudhuri, S., Johnson, A. D., Dimas, A. S., Loos, R. J. F., Vedantam, S., Chen, H., Florez, J. C., Fox, C., Liu, C.-T., Rybin, D., Couper, D. J., Kao, W. H. L., Li, M., Cornelis, M. C., Kraft, P., Sun, Q., van Dam, R. M., Stringham, H. M., Chines, P. S., Fischer, K., Fontanillas, P., Holmen, O. L., Hunt, S. E., Jackson, A. U., Kong, A., Lawrence, R., Meyer, J., Perry, J. R. B., Platou, C. G. P., Potter, S., Rehnberg, E., Robertson, N., Sivapalaratnam, S., Stančáková, A.,

Stirrups, K., Thorleifsson, G., Tikkanen, E., Wood, A. R., Almgren, P., Atalay, M., Benediktsson, R., Bonnycastle, L. L., Burtt, N., Carey, J., Charpentier, G., Crenshaw, A. T., Doney, A. S. F., Dorkhan, M., Edkins, S., Emilsson, V., Eury, E., Forsen, T., Gertow, K., Gigante, B., Grant, G. B., Groves, C. J., Guiducci, C., Herder, C., Hreidarsson, A. B., Hui, J., James, A., Jonsson, A., Rathmann, W., Klopp, N., Kravic, J., Krjutškov, K., Langford, C., Leander, K., Lindholm, E., Lobbens, S., Männistö, S., Mirza, G., Mühleisen, T. W., Musk, B., Parkin, M., Rallidis, L., Saramies, J., Sennblad, B., Shah, S., Sigurðsson, G., Silveira, A., Steinbach, G., Thorand, B., Trakalo, J., Veglia, F., Wennauer, R., Winckler, W., Zabaneh, D., Campbell, H., van Duijn, C., Uitterlinden, A. G., Hofman, A., Sijbrands, E., Abecasis, G. R., Owen, K. R., Zeggini, E., Trip, M. D., Forouhi, N. G., Syvänen, A.-C., Eriksson, J. G., Peltonen, L., Nöthen, M. M., Balkau, B., Palmer, C. N. A., Lyssenko, V., Tuomi, T., Isomaa, B., Hunter, D. J., Qi, L., Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner, A. R., Roden, M., Barroso, I., Wilsgaard, T., Beilby, J., Hovingh, K., Price, J. F., Wilson, J. F., Rauramaa, R., Lakka, T. A., Lind, L., Dedoussis, G., Njølstad, I., Pedersen, N. L., Khaw, K.-T., Wareham, N. J., Keinanen-Kiukaanniemi, S. M., Saaristo, T. E., Korpi-Hyövälti, E., Saltevo, J., Laakso, M., Kuusisto, J., Metspalu, A., Collins, F. S., Mohlke, K. L., Bergman, R. N., Tuomilehto, J., Boehm, B. O., Gieger, C., Hveem, K., Cauchi, S., Froguel, P., Baldassarre, D., Tremoli, E., Humphries, S. E., Saleheen, D., Danesh, J., Ingelsson, E., Ripatti, S., Salomaa, V., Erbel, R., Jöckel, K.-H., Moebus, S., Peters, A., Illig, T., de Faire, U., Hamsten, A., Morris, A. D., Donnelly, P. J., Frayling, T. M., Hattersley, A. T., Boerwinkle, E., Melander, O., Kathiresan, S., Nilsson, P. M., Deloukas, P., Thorsteinsdottir, U., Groop, L. C., Stefansson, K., Hu, F., Pankow, J. S., Dupuis, J., Meigs, J. B., Altshuler, D., Boehnke, M., McCarthy, M. I., and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012), "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes," *Nature genetics*, 44, 981–990.

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.

Mullins, N., Forstner, A. J., O'Connell, K. S., Coombes, B., Coleman, J. R. I., Qiao, Z., Als, T. D., Bigdeli, T. B., Børte, S., Bryois, J., Charney, A. W., Drange, O. K., Gandal, M. J., Hagenaars, S. P., Ikeda, M., Kamitaki, N., Kim, M., Krebs, K., Panagiotaropoulou, G., Schilder, B. M., Sloofman, L. G., Steinberg, S., Trubetskoy, V., Winsvold, B. S., Won, H.-H., Abramova, L., Adorjan, K., Agerbo, E., Al Eissa, M., Albani, D., Alliey-Rodriguez, N., Anjorin, A., Antilla, V., Antoniou, A., Awasthi, S., Baek, J. H., Bækvad-Hansen, M., Bass, N., Bauer, M., Beins, E. C., Bergen, S. E., Birner, A., Bøcker Pedersen, C., Bøen, E., Boks, M. P., Bosch, R., Brum, M., Brumpton, B. M., Brunkhorst-Kanaan, N., Budde, M., Bybjerg-Grauholm, J., Byerley, W., Cairns, M., Casas, M., Cervantes, P., Clarke, T.-K., Cruceanu, C., Cuellar-Barboza, A., Cunningham, J., Curtis, D., Czerski, P. M., Dale, A. M., Dalkner, N., David, F. S., Degenhardt, F., Djurovic, S., Dobbyn, A. L., Douzenis, A., Elvsåshagen, T., Escott-Price, V., Ferrier, I. N., Fiorentino, A., Foroud, T. M., Forty, L., Frank, J., Frei, O., Freimer, N. B., Frisén, L., Gade, K., Garnham, J., Gelernter, J., Giørtz Pedersen, M., Gizer, I. R., Gordon, S. D., Gordon-Smith, K., Greenwood, T. A., Grove, J., Guzman-Parra, J., Ha, K., Haraldsson, M., Hautzinger, M., Heilbronner, U., Hellgren, D., Herms, S., Hoffmann, P., Holmans, P. A., Huckins, L., Jamain, S., Johnson, J. S., Kalman, J. L., Kamatani, Y., Kennedy, J. L., Kittel-Schneider, S., Knowles, J. A., Kogevinas, M., Koromina, M., Kranz, T. M., Kranzler, H. R., Kubo, M., Kupka, R., Kushner, S. A., Lavebratt, C., Lawrence, J., Leber, M., Lee, H.-J., Lee, P. H., Levy, S. E., Lewis, C., Liao, C., Lucae, S., Lundberg, M., MacIntyre, D. J., Magnusson, S. H., Maier, W., Maihofer, A., Malaspina, D., Maratou, E., Martinsson, L., Mattheisen, M., McCarroll, S. A.,

161

McGregor, N. W., McGuffin, P., McKay, J. D., Medeiros, H., Medland, S. E., Millischer, V., Montgomery, G. W., Moran, J. L., Morris, D. W., Mühleisen, T. W., O'Brien, N., O'Donovan, C., Olde Loohuis, L. M., Oruc, L., Papiol, S., Pardiñas, A. F., Perry, A., Pfennig, A., Porichi, E., Potash, J. B., Quested, D., Raj, T., Rapaport, M. H., DePaulo, J. R., Regeer, E. J., Rice, J. P., Rivas, F., Rivera, M., Roth, J., Roussos, P., Ruderfer, D. M., Sánchez-Mora, C., Schulte, E. C., Senner, F., Sharp, S., Shilling, P. D., Sigurdsson, E., Sirignano, L., Slaney, C., Smeland, O. B., Smith, D. J., Sobell, J. L., Søholm Hansen, C., Soler Artigas, M., Spijker, A. T., Stein, D. J., Strauss, J. S., Świątkowska, B., Terao, C., Thorgeirsson, T. E., Toma, C., Tooney, P., Tsermpini, E.-E., Vawter, M. P., Vedder, H., Walters, J. T. R., Witt, S. H., Xi, S., Xu, W., Yang, J. M. K., Young, A. H., Young, H., Zandi, P. P., Zhou, H., Zillich, L., HUNT All-In Psychiatry, Adolfsson, R., Agartz, I., Alda, M., Alfredsson, L., Babadjanova, G., Backlund, L., Baune, B. T., Bellivier, F., Bengesser, S., Berrettini, W. H., Blackwood, D. H. R., Boehnke, M., Børglum, A. D., Breen, G., Carr, V. J., Catts, S., Corvin, A., Craddock, N., Dannlowski, U., Dikeos, D., Esko, T., Etain, B., Ferentinos, P., Frye, M., Fullerton, J. M., Gawlik, M., Gershon, E. S., Goes, F. S., Green, M. J., Grigoroiu-Serbanescu, M., Hauser, J., Henskens, F., Hillert, J., Hong, K. S., Hougaard, D. M., Hultman, C. M., Hveem, K., Iwata, N., Jablensky, A. V., Jones, I., Jones, L. A., Kahn, R. S., Kelsoe, J. R., Kirov, G., Landén, M., Leboyer, M., Lewis, C. M., Li, Q. S., Lissowska, J., Lochner, C., Loughland, C., Martin, N. G., Mathews, C. A., Mayoral, F., McElroy, S. L., McIntosh, A. M., McMahon, F. J., Melle, I., Michie, P., Milani, L., Mitchell, P. B., Morken, G., Mors, O., Mortensen, P. B., Mowry, B., Müller-Myhsok, B., Myers, R. M., Neale, B. M., Nievergelt, C. M., Nordentoft, M., Nöthen, M. M., O'Donovan, M. C., Oedegaard, K. J., Olsson, T., Owen, M. J., Paciga, S. A., Pantelis, C., Pato, C., Pato, M. T., Patrinos, G. P., Perlis, R. H., Posthuma, D., Ramos-Quiroga, J. A., Reif, A., Reininghaus, E. Z., Ribasés, M., Rietschel, M., Ripke, S., Rouleau, G. A., Saito, T., Schall, U., Schalling, M., Schofield, P. R., Schulze, T. G., Scott, L. J., Scott, R. J., Serretti, A., Shannon Weickert, C., Smoller, J. W., Stefansson, H., Stefansson, K., Stordal, E., Streit, F., Sullivan, P. F., Turecki, G., Vaaler, A. E., Vieta, E., Vincent, J. B., Waldman, I. D., Weickert, T. W., Werge, T., Wray, N. R., Zwart, J.-A., Biernacka, J. M., Nurnberger, J. I., Cichon, S., Edenberg, H. J., Stahl, E. A., McQuillin, A., Di Florio, A., Ophoff, R. A., and Andreassen, O. A. (2021), "Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology," *Nature genetics*, 53, 817–829.

Murphy, A. E., Schilder, B. M., and Skene, N. G. (2021), "MungeSumstats: A Bioconductor package for the standardisation and quality control of many GWAS summary statistics," *Bioinformatics*, 37, 4593–4596.

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., DeStefano, A. L., Kara, E., Bras, J., Sharma, M., Schulte, C., Keller, M. F., Arepalli, S., Letson, C., Edsall, C., Stefansson, H., Liu, X., Pliner, H., Lee, J. H., Cheng, R., International Parkinson's Disease Genomics Consortium (IPDGC), Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI), 23andMe, GenePD, NeuroGenetics Research Consortium (NGRC), Hussman Institute of Human Genomics (HIHG), Ashkenazi Jewish Dataset Investigator, Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE), North American Brain Expression Consortium (NABEC), United Kingdom Brain Expression Consortium (UKBEC), Greek Parkinson's Disease Consortium, Alzheimer Genetic Analysis Group, Ikram, M. A., Ioannidis, J. P. A., Hadjigeorgiou, G. M., Bis, J. C., Martinez, M., Perlmutter, J. S., Goate, A., Marder, K., Fiske, B., Sutherland, M., Xiromerisiou, G., Myers, R. H., Clark, L. N., Stefansson, K., Hardy, J. A., Heutink, P., Chen, H., Wood, N. W., Houlden, H., Payami, H., Brice, A., Scott, W. K., Gasser, T., Bertram, L., Eriksson, N., Foroud, T., and Singleton, A. B. (2014), "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease," *Nature genetics*, 46, 989–993.

Neniskyte, U., and Gross, C. T. (2017), "Errant gardeners: glial-cell-dependent synaptic pruning and

neurodevelopmental disorders," *Nature reviews neuroscience*, 18, 658–670.

Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium (2015), "Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways," *Nature neuroscience*, 18, 199–209.

Ng, B., White, C. C., Klein, H.-U., Sieberts, S. K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D. A., Mostafavi, S., and De Jager, P. L. (2017), "An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome," *Nature neuroscience*, 20, 1418–1426.

Nguyen, H. T., Bryois, J., Kim, A., Dobbyn, A., Huckins, L. M., Munoz-Manchado, A. B., Ruderfer, D. M., Genovese, G., Fromer, M., Xu, X., Pinto, D., Linnarsson, S., Verhage, M., Smit, A. B., Hjerling-Leffler, J., Buxbaum, J. D., Hultman, C., Sklar, P., Purcell, S. M., Lage, K., He, X., Sullivan, P. F., and Stahl, E. A. (2017), "Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders," *Genome medicine*, 9, 114.

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., Pardiñas, A. F., Owen, M. J., O'Donovan, M. C., Mill, J., and Bray, N. J. (2018), "Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders," *Genome biology*, 19, 194.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., Pappas, D. A., Jiang, L., Yin, J., Ye, L., Su, D.-F., Yang, J., Xie, G., Keystone, E., Westra, H.-J., Esko, T., Metspalu, A., Zhou, X., Gupta, N., Mirel, D., Stahl, E. A., Diogo, D., Cui, J., Liao, K., Guo, M. H., Myouzen, K., Kawaguchi, T., Coenen, M. J. H., van Riel, P. L. C. M., van de Laar, M. A. F. J., Guchelaar, H.-J., Huizinga, T. W. J., Dieudé, P., Mariette, X., Bridges, S. L., Jr, Zhernakova, A., Toes, R. E. M., Tak, P. P., Miceli-Richard, C., Bang, S.-Y., Lee, H.-S., Martin, J., Gonzalez-Gay, M. A., Rodriguez-Rodriguez, L., Rantapää-Dahlqvist, S., Arlestig, L., Choi, H. K., Kamatani, Y., Galan, P., Lathrop, M., RACI consortium, GARNET consortium, Eyre, S., Bowes, J., Barton, A., de Vries, N., Moreland, L. W., Criswell, L. A., Karlson, E. W., Taniguchi, A., Yamada, R., Kubo, M., Liu, J. S., Bae, S.-C., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Raychaudhuri, S., Stranger, B. E., De Jager, P. L., Franke, L., Visscher, P. M., Brown, M. A., Yamanaka, H., Mimori, T., Takahashi, A., Xu, H., Behrens, T. W., Siminovitch, K. A., Momohara, S., Matsuda, F., Yamamoto, K., and Plenge, R. M. (2014), "Genetics of rheumatoid arthritis contributes to biology and drug discovery," *Nature*, 506, 376–381.

Okbay, A., Baselmans, B. M. L., De Neve, J.-E., Turley, P., Nivard, M. G., Fontana, M. A., Meddens, S. F. W., Linnér, R. K., Rietveld, C. A., Derringer, J., Gratten, J., Lee, J. J., Liu, J. Z., de Vlaming, R., Ahluwalia, T. S., Buchwald, J., Cavadino, A., Frazier-Wood, A. C., Furlotte, N. A., Garfield, V., Geisel, M. H., Gonzalez, J. R., Haitjema, S., Karlsson, R., van der Laan, S. W., Ladwig, K.-H., Lahti, J., van der Lee, S. J., Lind, P. A., Liu, T., Matteson, L., Mihailov, E., Miller, M. B., Minica, C. C., Nolte, I. M., Mook-Kanamori, D., van der Most, P. J., Oldmeadow, C., Qian, Y., Raitakari, O., Rawal, R., Realo, A., Rueedi, R., Schmidt, B., Smith, A. V., Stergiakouli, E., Tanaka, T., Taylor, K., Thorleifsson, G., Wedenoja, J., Wellmann, J., Westra, H.-J., Willems, S. M., Zhao, W., LifeLines Cohort Study, Amin, N., Bakshi, A., Bergmann, S., Bjornsdottir, G., Boyle, P. A., Cherney, S., Cox, S. R., Davies, G., Davis, O. S. P., Ding, J., Direk, N., Eibich, P., Emeny, R. T., Fatemifar, G., Faul, J. D., Ferrucci, L., Forstner, A. J., Gieger, C., Gupta, R., Harris, T. B., Harris, J. M., Holliday, E. G., Hottenga, J.-J., De Jager, P. L., Kaakinen, M. A., Kajantie, E., Karhunen, V., Kolcic, I., Kumari, M., Launer, L. J., Franke, L., Li-Gao, R., Liewald, D. C., Koini, M., Loukola, A., Marques-Vidal, P.,

Montgomery, G. W., Mosing, M. A., Paternoster, L., Pattie, A., Petrovic, K. E., Pulkki-Råback, L., Quaye, L., Räikkönen, K., Rudan, I., Scott, R. J., Smith, J. A., Sutin, A. R., Trzaskowski, M., Vinkhuyzen, A. E., Yu, L., Zabaneh, D., Attia, J. R., Bennett, D. A., Berger, K., Bertram, L., Boomsma, D. I., Snieder, H., Chang, S.-C., Cucca, F., Deary, I. J., van Duijn, C. M., Eriksson, J. G., Bültmann, U., de Geus, E. J. C., Groenen, P. J. F., Gudnason, V., Hansen, T., Hartman, C. A., Haworth, C. M. A., Hayward, C., Heath, A. C., Hinds, D. A., Hyppönen, E., Iacono, W. G., Järvelin, M.-R., Jöckel, K.-H., Kaprio, J., Kardia, S. L. R., Keltikangas-Järvinen, L., Kraft, P., Kubzansky, L. D., Lehtimäki, T., Magnusson, P. K. E., Martin, N. G., McGue, M., Metspalu, A., Mills, M., de Mutsert, R., Oldehinkel, A. J., Pasterkamp, G., Pedersen, N. L., Plomin, R., Polasek, O., Power, C., Rich, S. S., Rosendaal, F. R., den Ruijter, H. M., Schlessinger, D., Schmidt, H., Svento, R., Schmidt, R., Alizadeh, B. Z., Sørensen, T. I. A., Spector, T. D., Starr, J. M., Stefansson, K., Steptoe, A., Terracciano, A., Thorsteinsdottir, U., Thurik, A. R., Timpson, N. J., Tiemeier, H., Uitterlinden, A. G., Vollenweider, P., Wagner, G. G., Weir, D. R., Yang, J., Conley, D. C., Smith, G. D., Hofman, A., Johannesson, M., Laibson, D. I., Medland, S. E., Meyer, M. N., Pickrell, J. K., Esko, T., Krueger, R. F., Beauchamp, J. P., Koellinger, P. D., Benjamin, D. J., Bartels, M., and Cesarini, D. (2016), "Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses," *Nature genetics*, 48, 624–633.

Ouwens, K. G., Jansen, R., Nivard, M. G., van Dongen, J., Frieser, M. J., Hottenga, J.-J., Arindrarto, W., Claringbould, A., van Iterson, M., Mei, H., Franke, L., Heijmans, B. T., A C 't Hoen, P., van Meurs, J., Brooks, A. I., BIOS Consortium, Penninx, B. W. J. H., and Boomsma, D. I. (2020), "A characterization of cis- and trans-heritability of RNA-Seq-based gene expression," *European journal of human genetics*, 28, 253–263.

Palmer, C. R., Liu, C. S., Romanow, W. J., Lee, M.-H., and Chun, J. (2021), "Altered cell and RNA isoform diversity in aging Down syndrome brains," *Proceedings of the National Academy of Sciences of the United States of America*, 118, e2114326118. https://doi.org/10.1073/pnas.2114326118.

Panyard, D. J., Kim, K. M., Darst, B. F., Deming, Y. K., Zhong, X., Wu, Y., Kang, H., Carlsson, C. M., Johnson, S. C., Asthana, S., Engelman, C. D., and Lu, Q. (2021), "Cerebrospinal fluid metabolomics identifies 19 brain-related phenotype associations," *Communications biology*, 4, 63.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., Han, J., Hubbard, L., Lynham, A., Mantripragada, K., Rees, E., MacCabe, J. H., McCarroll, S. A., Baune, B. T., Breen, G., Byrne, E. M., Dannlowski, U., Eley, T. C., Hayward, C., Martin, N. G., McIntosh, A. M., Plomin, R., Porteous, D. J., Wray, N. R., Caballero, A., Geschwind, D. H., Huckins, L. M., Ruderfer, D. M., Santiago, E., Sklar, P., Stahl, E. A., Won, H., Agerbo, E., Als, T. D., Andreassen, O. A., Bækvad-Hansen, M., Mortensen, P. B., Pedersen, C. B., Børglum, A. D., Bybjerg-Grauholm, J., Djurovic, S., Durmishi, N., Pedersen, M. G., Golimbet, V., Grove, J., Hougaard, D. M., Mattheisen, M., Molden, E., Mors, O., Nordentoft, M., Pejovic-Milovancevic, M., Sigurdsson, E., Silagadze, T., Hansen, C. S., Stefansson, K., Stefansson, H., Steinberg, S., Tosato, S., Werge, T., GERAD1 Consortium:, CRESTAR Consortium:, Collier, D. A., Rujescu, D., Kirov, G., Owen, M. J., O'Donovan, M. C., Walters, J. T. R., GERAD1 Consortium, CRESTAR Consortium, GERAD1 Consortium, and CRESTAR Consortium (2018), "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection," *Nature genetics*, 50, 381–389.

Parikshak, N. N., Gandal, M. J., and Geschwind, D. H. (2015), "Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders," *Nature reviews genetics*, 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.

Park, E., Jiang, Y., Hao, L., Hui, J., and Xing, Y. (2021), "Genetic variation and microRNA targeting of A-to-I RNA editing fine tune human tissue transcriptomes," *Genome biology*, 22, 77.

Park, S. M., Jang, H. J., and Lee, J. H. (2019), "Roles of Primary Cilia in the Developing Brain," *Frontiers in cellular neuroscience*, 13, 218.

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

Pazokitoroudi, A., Chiu, A. M., Burch, K. S., Pasaniuc, B., and Sankararaman, S. (2021), "Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data," *American journal of human genetics*, 108, 799–808.

Pazokitoroudi, A., Wu, Y., Burch, K. S., Hou, K., Zhou, A., Pasaniuc, B., and Sankararaman, S. (2020), "Efficient variance components analysis across millions of genomes," *Nature communications*, 11, 4020.

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., Lu, D., Connell, W., Ruzzo, E. K., Lowe, J. K., Hadzic, T., Hinz, F. I., Sabri, S., Lowry, W. E., Gerstein, M. B., Plath, K., and Geschwind, D. H. (2019), "A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation," *Neuron*, 103, 785–801.e8.

Pouget, J. G. (2018), "The Emerging Immunogenetic Architecture of Schizophrenia," *Schizophrenia bulletin*, 44, 993–1004.

Price, A. L., Helgason, A., Thorleifsson, G., McCarroll, S. A., Kong, A., and Stefansson, K. (2011), "Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals," *PLoS genetics*, 7, e1001317.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., and Willer, C. J. (2010), "LocusZoom: regional visualization of genome-wide association scan results," *Bioinformatics*, 26, 2336–2337.

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, T., Li, Y. I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V., Schadt, E. E., Young-Pearse, T., Mostafavi, S., Zhang, B., Sklar, P., Bennett, D. A., and De Jager, P. L. (2018), "Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility," *Nature genetics*, 50, 1584–1592.

van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019), "Genetic correlations of polygenic disease traits: from theory to practice," *Nature reviews genetics*, 20, 567–581.

van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., van der Spek, R. A. A., Võsa, U., de Jong, S., Robinson, M. R., Yang, J., Fogh, I., van Doormaal, P. T.,

Tazelaar, G. H. P., Koppers, M., Blokhuis, A. M., Sproviero, W., Jones, A. R., Kenna, K. P., van Eijk, K. R., Harschnitz, O., Schellevis, R. D., Brands, W. J., Medic, J., Menelaou, A., Vajda, A., Ticozzi, N., Lin, K., Rogelj, B., Vrabec, K., Ravnik-Glavač, M., Koritnik, B., Zidar, J., Leonardis, L., Grošelj, L. D., Millecamps, S., Salachas, F., Meininger, V., de Carvalho, M., Pinto, S., Mora, J. S., Rojas-García, R., Polak, M., Chandran, S., Colville, S., Swingler, R., Morrison, K. E., Shaw, P. J., Hardy, J., Orrell, R. W., Pittman, A., Sidle, K., Fratta, P., Malaspina, A., Topp, S., Petri, S., Abdulla, S., Drepper, C., Sendtner, M., Meyer, T., Ophoff, R. A., Staats, K. A., Wiedau-Pazos, M., Lomen-Hoerth, C., Van Deerlin, V. M., Trojanowski, J. Q., Elman, L., McCluskey, L., Basak, A. N., Tunca, C., Hamzeiy, H., Parman, Y., Meitinger, T., Lichtner, P., Radivojkov-Blagojevic, M., Andres, C. R., Maurel, C., Bensimon, G., Landwehrmeyer, B., Brice, A., Payan, C. A. M., Saker-Delye, S., Dürr, A., Wood, N. W., Tittmann, L., Lieb, W., Franke, A., Rietschel, M., Cichon, S., Nöthen, M. M., Amouyel, P., Tzourio, C., Dartigues, J.-F., Uitterlinden, A. G., Rivadeneira, F., Estrada, K., Hofman, A., Curtis, C., Blauw, H. M., van der Kooi, A. J., de Visser, M., Goris, A., Weber, M., Shaw, C. E., Smith, B. N., Pansarasa, O., Cereda, C., Del Bo, R., Comi, G. P., D'Alfonso, S., Bertolin, C., Sorarù, G., Mazzini, L., Pensato, V., Gellera, C., Tiloca, C., Ratti, A., Calvo, A., Moglia, C., Brunetti, M., Arcuti, S., Capozzo, R., Zecca, C., Lunetta, C., Penco, S., Riva, N., Padovani, A., Filosto, M., Muller, B., Stuit, R. J., PARALS Registry, SLALOM Group, SLAP Registry, FALS Sequencing Consortium, SLAGEN Consortium, NNIPPS Study Group, Blair, I., Zhang, K., McCann, E. P., Fifita, J. A., Nicholson, G. A., Rowe, D. B., Pamphlett, R., Kiernan, M. C., Grosskreutz, J., Witte, O. W., Ringer, T., Prell, T., Stubendorff, B., Kurth, I., Hübner, C. A., Leigh, P. N., Casale, F., Chio, A., Beghi, E., Pupillo, E., Tortelli, R., Logroscino, G., Powell, J., Ludolph, A. C., Weishaupt, J. H., Robberecht, W., Van Damme, P., Franke, L., Pers, T. H., Brown, R. H., Glass, J. D., Landers, J. E., Hardiman, O., Andersen, P. M., Corcia, P., Vourc'h, P., Silani, V., Wray, N. R., Visscher, P. M., de Bakker, P. I. W., van Es, M. A., Pasterkamp, R. J., Lewis, C. M., Breen, G., Al-Chalabi, A., van den Berg, L. H., and Veldink, J. H. (2016), "Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis," *Nature genetics*, 48, 1043–1048.

Ribeiro, D. M., Rubinacci, S., Ramisch, A., Hofmeister, R. J., Dermitzakis, E. T., and Delaneau, O. (2021), "The molecular basis, genetic control and pleiotropic effects of local gene co-expression," *Nature communications*, 12, 4842.

Ruzzo, E. K., Pérez-Cano, L., Jung, J.-Y., Wang, L.-K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J. N., Leventhal, O., Leppä, V. M., Gandal, M. J., Paskov, K., Stockham, N., Polioudakis, D., Lowe, J. K., Prober, D. A., Geschwind, D. H., and Wall, D. P. (2019), "Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks," *Cell*, 178, 850–866.e26.

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., Murtha, M. T., Bal, V. H., Bishop, S. L., Dong, S., Goldberg, A. P., Jinlu, C., Keaney, J. F., 3rd, Klei, L., Mandell, J. D., Moreno-De-Luca, D., Poultney, C. S., Robinson, E. B., Smith, L., Solli-Nowlan, T., Su, M. Y., Teran, N. A., Walker, M. F., Werling, D. M., Beaudet, A. L., Cantor, R. M., Fombonne, E., Geschwind, D. H., Grice, D. E., Lord, C., Lowe, J. K., Mane, S. M., Martin, D. M., Morrow, E. M., Talkowski, M. E., Sutcliffe, J. S., Walsh, C. A., Yu, T. W., Autism Sequencing Consortium, Ledbetter, D. H., Martin, C. L., Cook, E. H., Buxbaum, J. D., Daly, M. J., Devlin, B., Roeder, K., and State, M. W. (2015), "Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci," *Neuron*, 87, 1215–1233.

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., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., Brand, H., Schwartz, G., Nguyen, R., Guerrero, E. E., Dias, C., Autism Sequencing Consortium, iPSYCH-Broad Consortium, Betancur, C., Cook, E. H., Gallagher,

L., Gill, M., Sutcliffe, J. S., Thurm, A., Zwick, M. E., Børglum, A. D., State, M. W., Cicek, A. E., Talkowski, M. E., Cutler, D. J., Devlin, B., Sanders, S. J., Roeder, K., Daly, M. J., and Buxbaum, J. D. (2020), "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism," *Cell*, 180, 568–584.e23.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., Nagel, M., Awasthi, S., Barr, P. B., Coleman, J. R. I., Grasby, K. L., Hammerschlag, A. R., Kaminski, J. A., Karlsson, R., Krapohl, E., Lam, M., Nygaard, M., Reynolds, C. A., Trampush, J. W., Young, H., Zabaneh, D., Hägg, S., Hansell, N. K., Karlsson, I. K., Linnarsson, S., Montgomery, G. W., Muñoz-Manchado, A. B., Quinlan, E. B., Schumann, G., Skene, N. G., Webb, B. T., White, T., Arking, D. E., Avramopoulos, D., Bilder, R. M., Bitsios, P., Burdick, K. E., Cannon, T. D., Chiba-Falek, O., Christoforou, A., Cirulli, E. T., Congdon, E., Corvin, A., Davies, G., Deary, I. J., DeRosse, P., Dickinson, D., Djurovic, S., Donohoe, G., Conley, E. D., Eriksson, J. G., Espeseth, T., Freimer, N. A., Giakoumaki, S., Giegling, I., Gill, M., Glahn, D. C., Hariri, A. R., Hatzimanolis, A., Keller, M. C., Knowles, E., Koltai, D., Konte, B., Lahti, J., Le Hellard, S., Lencz, T., Liewald, D. C., London, E., Lundervold, A. J., Malhotra, A. K., Melle, I., Morris, D., Need, A. C., Ollier, W., Palotie, A., Payton, A., Pendleton, N., Poldrack, R. A., Räikkönen, K., Reinvang, I., Roussos, P., Rujescu, D., Sabb, F. W., Scult, M. A., Smeland, O. B., Smyrnis, N., Starr, J. M., Steen, V. M., Stefanis, N. C., Straub, R. E., Sundet, K., Tiemeier, H., Voineskos, A. N., Weinberger, D. R., Widen, E., Yu, J., Abecasis, G., Andreassen, O. A., Breen, G., Christiansen, L., Debrabant, B., Dick, D. M., Heinz, A., Hjerling-Leffler, J., Ikram, M. A., Kendler, K. S., Martin, N. G., Medland, S. E., Pedersen, N. L., Plomin, R., Polderman, T. J. C., Ripke, S., van der Sluis, S., Sullivan, P. F., Vrieze, S. I., Wright, M. J., and Posthuma, D. (2018), "Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence," *Nature genetics*, 50, 912–919.

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

Schilder, B. M., Humphrey, J., and Raj, T. (2021), "echolocatoR: an automated end-to-end statistical and functional genomic fine-mapping pipeline," *Bioinformatics*, 38, 536–539.

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

Scott, A. J., Chiang, C., and Hall, I. M. (2021), "Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes," *Genome research*, 31, 2249–2257.

Searle, S. R. (1961), "Phenotypic, Genetic and Environmental Correlations," *Biometrics*, 17, 474–480.

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly, M. J., Carroll, M. C., Stevens, B., and McCarroll, S. A. (2016), "Schizophrenia risk from complex variation of complement component 4," *Nature*, 530, 177–183.

Sellgren, C. M., Gracias, J., Watmuff, B., Biag, J. D., Thanos, J. M., Whittredge, P. B., Fu, T., Worringer, K., Brown, H. E., Wang, J., Kaykas, A., Karmacharya, R., Goold, C. P., Sheridan, S. D., and Perlis, R. H. (2019), "Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning," *Nature neuroscience*, 22, 374–385.

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," *Proceedings of the National Academy of Sciences of the United States of America*, 111, E5593–601.

Singh, R. N., Seo, J., and Singh, N. N. (2020), "RNA in spinal muscular atrophy: therapeutic implications of targeting," *Expert opinion on therapeutic targets*, 24, 731–743.

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., Bromet, E. J., Buckley, P. F., Bunney, W. E., Bybjerg-Grauholm, J., Byerley, W. F., Chapman, S. B., Chen, W. J., Churchhouse, C., Craddock, N., Cusick, C. M., DeLisi, L., Dodge, S., Escamilla, M. A., Eskelinen, S., Fanous, A. H., Faraone, S. V., Fiorentino, A., Francioli, L., Gabriel, S. B., Gage, D., Gagliano Taliun, S. A., Ganna, A., Genovese, G., Glahn, D. C., Grove, J., Hall, M.-H., Hämäläinen, E., Heyne, H. O., Holi, M., Hougaard, D. M., Howrigan, D. P., Huang, H., Hwu, H.-G., Kahn, R. S., Kang, H. M., Karczewski, K. J., Kirov, G., Knowles, J. A., Lee, F. S., Lehrer, D. S., Lescai, F., Malaspina, D., Marder, S. R., McCarroll, S. A., McIntosh, A. M., Medeiros, H., Milani, L., Morley, C. P., Morris, D. W., Mortensen, P. B., Myers, R. M., Nordentoft, M., O'Brien, N. L., Olivares, A. M., Ongur, D., Ouwehand, W. H., Palmer, D. S., Paunio, T., Quested, D., Rapaport, M. H., Rees, E., Rollins, B., Satterstrom, F. K., Schatzberg, A., Scolnick, E., Scott, L. J., Sharp, S. I., Sklar, P., Smoller, J. W., Sobell, J. L., Solomonson, M., Stahl, E. A., Stevens, C. R., Suvisaari, J., Tiao, G., Watson, S. J., Watts, N. A., Blackwood, D. H., Børglum, A. D., Cohen, B. M., Corvin, A. P., Esko, T., Freimer, N. B., Glatt, S. J., Hultman, C. M., McQuillin, A., Palotie, A., Pato, C. N., Pato, M. T., Pulver, A. E., St Clair, D., Tsuang, M. T., Vawter, M. P., Walters, J. T., Werge, T. M., Ophoff, R. A., Sullivan, P. F., Owen, M. J., Boehnke, M., O'Donovan, M. C., Neale, B. M., and Daly, M. J. (2022), "Rare coding variants in ten genes confer substantial risk for schizophrenia," *Nature*, 604, 509–516.

Skene, N. G., Bryois, J., Bakken, T. E., Breen, G., Crowley, J. J., Gaspar, H. A., Giusti-Rodriguez, P., Hodge, R. D., Miller, J. A., Muñoz-Manchado, A. B., O'Donovan, M. C., Owen, M. J., Pardiñas, A. F., Ryge, J., Walters, J. T. R., Linnarsson, S., Lein, E. S., Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Sullivan, P. F., and Hjerling-Leffler, J. (2018), "Genetic identification of brain cell types underlying schizophrenia," *Nature genetics*, 50, 825–833.

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

Sng, L. M. F., Thomson, P. C., and Trabzuni, D. (2019), "Genome-wide human brain eQTLs: In-depth analysis and insights using the UKBEC dataset," *Scientific reports*, 9, 19201.

Sodini, S. M., Kemper, K. E., Wray, N. R., and Trzaskowski, M. (2018), "Comparison of Genotypic and Phenotypic Correlations: Cheverud's Conjecture in Humans," *Genetics*, 209, 941–948.

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., Mattheisen, M., Wang, Y., Coleman, J. R. I., Gaspar, H. A., de Leeuw, C. A., Steinberg, S., Pavlides, J. M. W., Trzaskowski, M., Byrne, E. M., Pers, T. H., Holmans, P. A., Richards, A. L., Abbott, L., Agerbo, E., Akil, H., Albani, D., Alliey-Rodriguez, N., Als, T. D., Anjorin, A., Antilla, V., Awasthi, S., Badner, J. A., Bækvad-Hansen, M., Barchas, J. D., Bass, N., Bauer, M., Belliveau, R., Bergen, S. E., Pedersen, C. B., Bøen, E., Boks, M. P., Boocock, J., Budde, M., Bunney, W., Burmeister, M., Bybjerg-Grauholm, J., Byerley, W., Casas, M., Cerrato, F., Cervantes, P., Chambert, K., Charney, A. W., Chen, D., Churchhouse, C., Clarke, T.-K., Coryell, W., Craig, D. W., Cruceanu, C., Curtis, D., Czerski, P. M., Dale, A. M., de Jong, S., Degenhardt, F., Del-Favero, J., DePaulo, J. R., Djurovic, S.,

Dobbyn, A. L., Dumont, A., Elvsåshagen, T., Escott-Price, V., Fan, C. C., Fischer, S. B., Flickinger, M., Foroud, T. M., Forty, L., Frank, J., Fraser, C., Freimer, N. B., Frisén, L., Gade, K., Gage, D., Garnham, J., Giambartolomei, C., Pedersen, M. G., Goldstein, J., Gordon, S. D., Gordon-Smith, K., Green, E. K., Green, M. J., Greenwood, T. A., Grove, J., Guan, W., Guzman-Parra, J., Hamshere, M. L., Hautzinger, M., Heilbronner, U., Herms, S., Hipolito, M., Hoffmann, P., Holland, D., Huckins, L., Jamain, S., Johnson, J. S., Juréus, A., Kandaswamy, R., Karlsson, R., Kennedy, J. L., Kittel-Schneider, S., Knowles, J. A., Kogevinas, M., Koller, A. C., Kupka, R., Lavebratt, C., Lawrence, J., Lawson, W. B., Leber, M., Lee, P. H., Levy, S. E., Li, J. Z., Liu, C., Lucae, S., Maaser, A., MacIntyre, D. J., Mahon, P. B., Maier, W., Martinsson, L., McCarroll, S., McGuffin, P., McInnis, M. G., McKay, J. D., Medeiros, H., Medland, S. E., Meng, F., Milani, L., Montgomery, G. W., Morris, D. W., Mühleisen, T. W., Mullins, N., Nguyen, H., Nievergelt, C. M., Adolfsson, A. N., Nwulia, E. A., O'Donovan, C., Loohuis, L. M. O., Ori, A. P. S., Oruc, L., Össby, U., Perlis, R. H., Perry, A., Pfennig, A., Potash, J. B., Purcell, S. M., Regeer, E. J., Reif, A., Reinbold, C. S., Rice, J. P., Rivas, F., Rivera, M., Roussos, P., Ruderfer, D. M., Ryu, E., Sánchez-Mora, C., Schatzberg, A. F., Scheftner, W. A., Schork, N. J., Shannon Weickert, C., Shehktman, T., Shilling, P. D., Sigurdsson, E., Slaney, C., Smeland, O. B., Sobell, J. L., Søholm Hansen, C., Spijker, A. T., St Clair, D., Steffens, M., Strauss, J. S., Streit, F., Strohmaier, J., Szelinger, S., Thompson, R. C., Thorgeirsson, T. E., Treutlein, J., Vedder, H., Wang, W., Watson, S. J., Weickert, T. W., Witt, S. H., Xi, S., Xu, W., Young, A. H., Zandi, P., Zhang, P., Zöllner, S., eQTLGen Consortium, BIOS Consortium, Adolfsson, R., Agartz, I., Alda, M., Backlund, L., Baune, B. T., Bellivier, F., Berrettini, W. H., Biernacka, J. M., Blackwood, D. H. R., Boehnke, M., Børglum, A. D., Corvin, A., Craddock, N., Daly, M. J., Dannlowski, U., Esko, T., Etain, B., Frye, M., Fullerton, J. M., Gershon, E. S., Gill, M., Goes, F., Grigoroiu-Serbanescu, M., Hauser, J., Hougaard, D. M., Hultman, C. M., Jones, I., Jones, L. A., Kahn, R. S., Kirov, G., Landén, M., Leboyer, M., Lewis, C. M., Li, Q. S., Lissowska, J., Martin, N. G., Mayoral, F., McElroy, S. L., McIntosh, A. M., McMahon, F. J., Melle, I., Metspalu, A., Mitchell, P. B., Morken, G., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Myers, R. M., Neale, B. M., Nimgaonkar, V., Nordentoft, M., Nöthen, M. M., O'Donovan, M. C., Oedegaard, K. J., Owen, M. J., Paciga, S. A., Pato, C., Pato, M. T., Posthuma, D., Ramos-Quiroga, J. A., Ribasés, M., Rietschel, M., Rouleau, G. A., Schalling, M., Schofield, P. R., Schulze, T. G., Serretti, A., Smoller, J. W., Stefansson, H., Stefansson, K., Stordal, E., Sullivan, P. F., Turecki, G., Vaaler, A. E., Vieta, E., Vincent, J. B., Werge, T., Nurnberger, J. I., Wray, N. R., Di Florio, A., Edenberg, H. J., Cichon, S., Ophoff, R. A., Scott, L. J., Andreassen, O. A., Kelsoe, J., Sklar, P., and Bipolar Disorder Working Group of the Psychiatric Genomics Consortium (2019), "Genome-wide association study identifies 30 loci associated with bipolar disorder," *Nature genetics*, 51, 793–803.

Stein, J. L., de la Torre-Ubieta, L., Tian, Y., Parikshak, N. N., Hernández, I. A., Marchetto, M. C., Baker, D. K., Lu, D., Hinman, C. R., Lowe, J. K., Wexler, E. M., Muotri, A. R., Gage, F. H., Kosik, K. S., and Geschwind, D. H. (2014), "A quantitative framework to evaluate modeling of cortical development by neural stem cells," *Neuron*, 83, 69–86.

Stephan, A. H., Barres, B. A., and Stevens, B. (2012), "The complement system: an unexpected role in synaptic pruning during development and disease," *Annual review of neuroscience*, 35, 369–389.

Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. H., and Blencowe, B. J. (2018), "Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop," *Molecular cell*, 72, 187–200.e6.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005), "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15545–

15550.

Südhof, T. C. (2018), "Towards an Understanding of Synapse Formation," *Neuron*, 100, 276–293.

Sullivan, P. F., Kendler, K. S., and Neale, M. C. (2003), "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies," *Archives of general psychiatry*, 60, 1187–1192.

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," *Nature communications*, 8, 14519.

Tang, G., Gudsnuk, K., Kuo, S.-H., Cotrina, M. L., Rosoklija, G., Sosunov, A., Sonders, M. S., Kanter, E., Castagna, C., Yamamoto, A., Yue, Z., Arancio, O., Peterson, B. S., Champagne, F., Dwork, A. J., Goldman, J., and Sulzer, D. (2014), "Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits," *Neuron*, 83, 1131–1143.

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C.-Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., Wu, Y., Zeng, J., Grove, J., Kim, M., Li, Z., Voloudakis, G., Zhang, W., Adams, M., Agartz, I., Atkinson, E. G., Agerbo, E., Al Eissa, M., Albus, M., Alexander, M., Alizadeh, B. Z., Alptekin, K., Als, T. D., Amin, F., Arolt, V., Arrojo, M., Athanasiu, L., Azevedo, M. H., Bacanu, S. A., Bass, N. J., Begemann, M., Belliveau, R. A., Bene, J., Benyamin, B., Bergen, S. E., Blasi, G., Bobes, J., Bonassi, S., Braun, A., Bressan, R. A., Bromet, E. J., Bruggeman, R., Buckley, P. F., Buckner, R. L., Bybjerg-Grauholm, J., Cahn, W., Cairns, M. J., Calkins, M. E., Carr, V. J., Castle, D., Catts, S. V., Chambert, K. D., Chan, R. C. K., Chaumette, B., Cheng, W., Cheung, E. F. C., Chong, S. A., Cohen, D., Consoli, A., Cordeiro, Q., Costas, J., Curtis, C., Davidson, M., Davis, K. L., de Haan, L., Degenhardt, F., DeLisi, L. E., Demontis, D., Dickerson, F., Dikeos, D., Dinan, T., Djurovic, S., Duan, J., Ducci, G., Dudbridge, F., Eriksson, J. G., Fañanás, L., Faraone, S. V., Fiorentino, A., Forstner, A., Frank, J., Freimer, N. B., Fromer, M., Frustaci, A., Gadelha, A., Genovese, G., Gershon, E. S., Giannitelli, M., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J. I., González Peñas, J., González-Pinto, A., Gopal, S., Gratten, J., Green, M. F., Greenwood, T. A., Guillin, O., Gülöksüz, S., Gur, R. E., Gur, R. C., Gutiérrez, B., Hahn, E., Hakonarson, H., Haroutunian, V., Hartmann, A. M., Harvey, C., Hayward, C., Henskens, F. A., Herms, S., Hoffmann, P., Howrigan, D. P., Ikeda, M., Iyegbe, C., Joa, I., Julià, A., Kähler, A. K., Kam-Thong, T., Kamatani, Y., Karachanak-Yankova, S., Kebir, O., Keller, M. C., Kelly, B. J., Khrunin, A., Kim, S.-W., Klovins, J., Kondratiev, N., Konte, B., Kraft, J., Kubo, M., Kučinskas, V., Kučinskiene, Z. A., Kusumawardhani, A., Kuzelova-Ptackova, H., Landi, S., Lazzeroni, L. C., Lee, P. H., Legge, S. E., Lehrer, D. S., Lencer, R., Lerer, B., Li, M., Lieberman, J., Light, G. A., Limborska, S., Liu, C.-M., Lönnqvist, J., Loughland, C. M., Lubinski, J., Luykx, J. J., Lynham, A., Macek, M., Jr, Mackinnon, A., Magnusson, P. K. E., Maher, B. S., Maier, W., Malaspina, D., Mallet, J., Marder, S. R., Marsal, S., Martin, A. R., Martorell, L., Mattheisen, M., McCarley, R. W., McDonald, C., McGrath, J. J., Medeiros, H., Meier, S., Melegh, B., Melle, I., Mesholam-Gately, R. I., Metspalu, A., Michie, P. T., Milani, L., Milanova, V., Mitjans, M., Molden, E., Molina, E., Molto, M. D., Mondelli, V., Moreno, C., Morley, C. P., Muntané, G., Murphy, K. C., Myin-Germeys, I., Nenadić, I., Nestadt, G., Nikitina-Zake, L., Noto, C., Nuechterlein, K. H., O'Brien, N. L., O'Neill, F. A., Oh, S.-Y., Olincy, A., Ota, V. K., Pantelis, C., Papadimitriou, G. N., Parellada, M., Paunio, T., Pellegrino, R., Periyasamy, S., Perkins, D. O., Pfuhlmann, B., Pietiläinen, O., Pimm, J., Porteous, D., Powell, J., Quattrone, D., Quested, D., Radant, A. D., Rampino, A., Rapaport, M. H., Rautanen, A., Reichenberg, A., Roe, C., Roffman, J. L., Roth, J., Rothermundt, M., Rutten, B. P. F., Saker-Delye, S., Salomaa, V., Sanjuan, J., Santoro, M. L., Savitz, A., Schall, U., Scott, R. J., Seidman, L. J., Sharp, S. I., Shi, J., Siever, L. J., Sigurdsson, E., Sim, K., Skarabis, N., Slominsky, P., So, H.-C.,

Sobell, J. L., Söderman, E., Stain, H. J., Steen, N. E., Steixner-Kumar, A. A., Stögmann, E., Stone, W. S., Straub, R. E., Streit, F., Strengman, E., Stroup, T. S., Subramaniam, M., Sugar, C. A., Suvisaari, J., Svrakic, D. M., Swerdlow, N. R., Szatkiewicz, J. P., Ta, T. M. T., Takahashi, A., Terao, C., Thibaut, F., Toncheva, D., Tooney, P. A., Torretta, S., Tosato, S., Tura, G. B., Turetsky, B. I., Üçok, A., Vaaler, A., van Amelsvoort, T., van Winkel, R., Veijola, J., Waddington, J., Walter, H., Waterreus, A., Webb, B. T., Weiser, M., Williams, N. M., Witt, S. H., Wormley, B. K., Wu, J. Q., Xu, Z., Yolken, R., Zai, C. C., Zhou, W., Zhu, F., Zimprich, F., Atbaşoğlu, E. C., Ayub, M., Benner, C., Bertolino, A., Black, D. W., Bray, N. J., Breen, G., Buccola, N. G., Byerley, W. F., Chen, W. J., Cloninger, C. R., Crespo-Facorro, B., Donohoe, G., Freedman, R., Galletly, C., Gandal, M. J., Gennarelli, M., Hougaard, D. M., Hwu, H.-G., Jablensky, A. V., McCarroll, S. A., Moran, J. L., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Neil, A. L., Nordentoft, M., Pato, M. T., Petryshen, T. L., Pirinen, M., Pulver, A. E., Schulze, T. G., Silverman, J. M., Smoller, J. W., Stahl, E. A., Tsuang, D. W., Vilella, E., Wang, S.-H., Xu, S., Indonesia Schizophrenia Consortium, PsychENCODE, Psychosis Endophenotypes International Consortium, SynGO Consortium, Adolfsson, R., Arango, C., Baune, B. T., Belangero, S. I., Børglum, A. D., Braff, D., Bramon, E., Buxbaum, J. D., Campion, D., Cervilla, J. A., Cichon, S., Collier, D. A., Corvin, A., Curtis, D., Forti, M. D., Domenici, E., Ehrenreich, H., Escott-Price, V., Esko, T., Fanous, A. H., Gareeva, A., Gawlik, M., Gejman, P. V., Gill, M., Glatt, S. J., Golimbet, V., Hong, K. S., Hultman, C. M., Hyman, S. E., Iwata, N., Jönsson, E. G., Kahn, R. S., Kennedy, J. L., Khusnutdinova, E., Kirov, G., Knowles, J. A., Krebs, M.-O., Laurent-Levinson, C., Lee, J., Lencz, T., Levinson, D. F., Li, Q. S., Liu, J., Malhotra, A. K., Malhotra, D., McIntosh, A., McQuillin, A., Menezes, P. R., Morgan, V. A., Morris, D. W., Mowry, B. J., Murray, R. M., Nimgaonkar, V., Nöthen, M. M., Ophoff, R. A., Paciga, S. A., Palotie, A., Pato, C. N., Qin, S., Rietschel, M., Riley, B. P., Rivera, M., Rujescu, D., Saka, M. C., Sanders, A. R., Schwab, S. G., Serretti, A., Sham, P. C., Shi, Y., St Clair, D., Stefánsson, H., Stefansson, K., Tsuang, M. T., van Os, J., Vawter, M. P., Weinberger, D. R., Werge, T., Wildenauer, D. B., Yu, X., Yue, W., Holmans, P. A., Pocklington, A. J., Roussos, P., Vassos, E., Verhage, M., Visscher, P. M., Yang, J., Posthuma, D., Andreassen, O. A., Kendler, K. S., Owen, M. J., Wray, N. R., Daly, M. J., Huang, H., Neale, B. M., Sullivan, P. F., Ripke, S., Walters, J. T. R., O'Donovan, M. C., and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2022), "Mapping genomic loci implicates genes and synaptic biology in schizophrenia," *Nature*, 604, 502–508.

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., González-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., and Barash, Y. (2016), "A new view of transcriptome complexity and regulation through the lens of local splicing variations," *eLife*, 5, e11752.

Vattikuti, S., Guo, J., and Chow, C. C. (2012), "Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits," *PLoS genetics*, 8, e1002637.

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G.-B., Lee, S. H., Wray, N. R., Goddard, M. E., and Yang, J. (2014), "Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples," *PLoS genetics*, 10, e1004269.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017), "10 Years of GWAS Discovery: Biology, Function, and Translation," *American journal of human genetics*, 101, 5–22.

de Vlaming, R., Slob, E. A. W., Jansen, P. R., Dagher, A., Koellinger, P. D., Groenen, P. J. F., and Rietveld, C. A. (2021), "Multivariate analysis reveals shared genetic architecture of brain morphology and human behavior," *Communications biology*, 4, 1180.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A.,

Kreuzhuber, R., Kasela, S., Pervjakova, N., Alvaes, I., Fave, M.-J., Agbessi, M., Christiansen, M., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg-Guzman, J., Kettunen, J., Powell, J., Lee, B., Zhang, F., Arindrarto, W., Beutner, F., BIOS Consortium, Brugge, H., i2QTL Consortium, Dmitreva, J., Elansary, M., Fairfax, B. P., Georges, M., Heijmans, B. T., Kähönen, M., Kim, Y., Knight, J. C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U. M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M., Penninx, B., Pritchard, J., Raitakari, O., Rotzchke, O., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., 't, P. A., Thiery, J., Tönjes, A., van Dongen, J., van Iterson, M., Veldink, J., Völker, U., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G. W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., van Meurs, J., Prokisch, H., Ahsan, H., Pierce, B., Lehtimäki, T., Boomsma, D., Psaty, B. M., Gharib, S. A., Awadalla, P., Milani, L., Ouwehand, W., Downes, K., Stegle, O., Battle, A., Yang, J., Visscher, P. M., Scholz, M., Gibson, G., Esko, T., and Franke, L. (2018), "Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis," *bioRxiv*. https://doi.org/10.1101/447367.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, J. L. M., Im, H. K., Pasaniuc, B., Rivas, M. A., and Kundaje, A. (2019), "Opportunities and challenges for transcriptome-wide association studies," *Nature genetics*, 51, 592–599.

Waitt, D. E., and Levin, D. A. (1998), "Genetic and phenotypic correlations in plants: a botanical test of Cheverud's conjecture," *Heredity*, 80, 310–319.

Walker, R. L., Ramaswami, G., Hartl, C., Mancuso, N., Gandal, M. J., de la Torre-Ubieta, L., Pasaniuc, B., Stein, J. L., and Geschwind, D. H. (2019), "Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms," *Cell*, 179, 750–771.e22.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., Clarke, D., Gu, M., Emani, P., Yang, Y. T., Xu, M., Gandal, M. J., Lou, S., Zhang, J., Park, J. J., Yan, C., Rhie, S. K., Manakongtreecheep, K., Zhou, H., Nathan, A., Peters, M., Mattei, E., Fitzgerald, D., Brunetti, T., Moore, J., Jiang, Y., Girdhar, K., Hoffman, G. E., Kalayci, S., Gümüş, Z. H., Crawford, G. E., PsychENCODE Consortium, Roussos, P., Akbarian, S., Jaffe, A. E., White, K. P., Weng, Z., Sestan, N., Geschwind, D. H., Knowles, J. A., and Gerstein, M. B. (2018), "Comprehensive functional genomic resource and integrative model for the human brain," *Science*, 362. https://doi.org/10.1126/science.aat8464.

Wang, S., Livingston, M. J., Su, Y., and Dong, Z. (2015), "Reciprocal regulation of cilia and autophagy via the MTOR and proteasome pathways," *Autophagy*, 11, 607–616.

Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., Aquino-Michaels, K., GTEx Consortium, Cox, N. J., Nicolae, D. L., and Im, H. K. (2016), "Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues," *PLoS genetics*, 12, e1006423.

Wiegering, A., Rüther, U., and Gerhardt, C. (2019), "The Role of Primary Cilia in the Crosstalk between the Ubiquitin–Proteasome System and Autophagy," *Cells*, 8, 241.

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., LifeLines Cohort Study, BIOS Consortium, and Franke, L. (2018), "Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs," *Nature genetics*, 50, 493–497.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann, J. S., Bragg-Gresham, J. L., Chang, H.-Y., Demirkan, A., Den Hertog, H. M., Do, R., Donnelly, L. A., Ehret, G. B., Esko, T., Feitosa, M. F., Ferreira, T., Fischer, K., Fontanillas, P., Fraser, R. M., Freitag, D. F., Gurdasani, D., Heikkilä, K., Hyppönen, E., Isaacs, A., Jackson, A. U., Johansson, Å., Johnson, T., Kaakinen, M., Kettunen, J., Kleber, M. E., Li, X., Luan, J. 'an, Lyytikäinen, L.-P., Magnusson, P. K. E., Mangino, M., Mihailov, E., Montasser, M. E., Müller-Nurasyid, M., Nolte, I. M., O'Connell, J. R., Palmer, C. D., Perola, M., Petersen, A.-K., Sanna, S., Saxena, R., Service, S. K., Shah, S., Shungin, D., Sidore, C., Song, C., Strawbridge, R. J., Surakka, I., Tanaka, T., Teslovich, T. M., Thorleifsson, G., Van den Herik, E. G., Voight, B. F., Volcik, K. A., Waite, L. L., Wong, A., Wu, Y., Zhang, W., Absher, D., Asiki, G., Barroso, I., Been, L. F., Bolton, J. L., Bonnycastle, L. L., Brambilla, P., Burnett, M. S., Cesana, G., Dimitriou, M., Doney, A. S. F., Döring, A., Elliott, P., Epstein, S. E., Ingi Eyjolfsson, G., Gigante, B., Goodarzi, M. O., Grallert, H., Gravito, M. L., Groves, C. J., Hallmans, G., Hartikainen, A.-L., Hayward, C., Hernandez, D., Hicks, A. A., Holm, H., Hung, Y.-J., Illig, T., Jones, M. R., Kaleebu, P., Kastelein, J. J. P., Khaw, K.-T., Kim, E., Klopp, N., Komulainen, P., Kumari, M., Langenberg, C., Lehtimäki, T., Lin, S.-Y., Lindström, J., Loos, R. J. F., Mach, F., McArdle, W. L., Meisinger, C., Mitchell, B. D., Müller, G., Nagaraja, R., Narisu, N., Nieminen, T. V. M., Nsubuga, R. N., Olafsson, I., Ong, K. K., Palotie, A., Papamarkou, T., Pomilla, C., Pouta, A., Rader, D. J., Reilly, M. P., Ridker, P. M., Rivadeneira, F., Rudan, I., Ruokonen, A., Samani, N., Scharnagl, H., Seeley, J., Silander, K., Stančáková, A., Stirrups, K., Swift, A. J., Tiret, L., Uitterlinden, A. G., van Pelt, L. J., Vedantam, S., Wainwright, N., Wijmenga, C., Wild, S. H., Willemsen, G., Wilsgaard, T., Wilson, J. F., Young, E. H., Zhao, J. H., Adair, L. S., Arveiler, D., Assimes, T. L., Bandinelli, S., Bennett, F., Bochud, M., Boehm, B. O., Boomsma, D. I., Borecki, I. B., Bornstein, S. R., Bovet, P., Burnier, M., Campbell, H., Chakravarti, A., Chambers, J. C., Chen, Y.-D. I., Collins, F. S., Cooper, R. S., Danesh, J., Dedoussis, G., de Faire, U., Feranil, A. B., Ferrières, J., Ferrucci, L., Freimer, N. B., Gieger, C., Groop, L. C., Gudnason, V., Gyllensten, U., Hamsten, A., Harris, T. B., Hingorani, A., Hirschhorn, J. N., Hofman, A., Hovingh, G. K., Hsiung, C. A., Humphries, S. E., Hunt, S. C., Hveem, K., Iribarren, C., Järvelin, M.-R., Jula, A., Kähönen, M., Kaprio, J., Kesäniemi, A., Kivimaki, M., Kooner, J. S., Koudstaal, P. J., Krauss, R. M., Kuh, D., Kuusisto, J., Kyvik, K. O., Laakso, M., Lakka, T. A., Lind, L., Lindgren, C. M., Martin, N. G., März, W., McCarthy, M. I., McKenzie, C. A., Meneton, P., Metspalu, A., Moilanen, L., Morris, A. D., Munroe, P. B., Njølstad, I., Pedersen, N. L., Power, C., Pramstaller, P. P., Price, J. F., Psaty, B. M., Quertermous, T., Rauramaa, R., Saleheen, D., Salomaa, V., Sanghera, D. K., Saramies, J., Schwarz, P. E. H., Sheu, W. H.-H., Shuldiner, A. R., Siegbahn, A., Spector, T. D., Stefansson, K., Strachan, D. P., Tayo, B. O., Tremoli, E., Tuomilehto, J., Uusitupa, M., van Duijn, C. M., Vollenweider, P., Wallentin, L., Wareham, N. J., Whitfield, J. B., Wolffenbuttel, B. H. R., Ordovas, J. M., Boerwinkle, E., Palmer, C. N. A., Thorsteinsdottir, U., Chasman, D. I., Rotter, J. I., Franks, P. W., Ripatti, S., Cupples, L. A., Sandhu, M. S., Rich, S. S., Boehnke, M., Deloukas, P., Kathiresan, S., Mohlke, K. L., Ingelsson, E., Abecasis, G. R., and Global Lipids Genetics Consortium (2013), "Discovery and refinement of loci associated with lipid levels," *Nature genetics*, 45, 1274–1283.

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., Murtha, M. T., Bichsel, C., Niu, W., Cotney, J., Ercan-Sencicek, A. G., Gockley, J., Gupta, A. R., Han, W., He, X., Hoffman, E. J., Klei, L., Lei, J., Liu, W., Liu, L., Lu, C., Xu, X., Zhu, Y., Mane, S. M., Lein, E. S., Wei, L., Noonan, J. P., Roeder, K., Devlin, B., Sestan, N., and State, M. W. (2013), "Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism," *Cell*, 155, 997–1007.

Wingo, A. P., Liu, Y., Gerasimov, E. S., Gockley, J., Logsdon, B. A., Duong, D. M., Dammer, E. B., Robins, C., Beach, T. G., Reiman, E. M., Epstein, M. P., De Jager, P. L., Lah, J. J., Bennett, D. A., Seyfried, N. T., Levey, A. I., and Wingo, T. S. (2021), "Integrating human brain proteomes with

genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis," *Nature genetics*, 53, 143–146.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., Christensen, J. H., Clarke, T.-K., Coleman, J. I. R., Colodro-Conde, L., Couvy-Duchesne, B., Craddock, N., Crawford, G. E., Crowley, C. A., Dashti, H. S., Davies, G., Deary, I. J., Degenhardt, F., Derks, E. M., Direk, N., Dolan, C. V., Dunn, E. C., Eley, T. C., Eriksson, N., Escott-Price, V., Kiadeh, F. H. F., Finucane, H. K., Forstner, A. J., Frank, J., Gaspar, H. A., Gill, M., Giusti-Rodríguez, P., Goes, F. S., Gordon, S. D., Grove, J., Hall, L. S., Hannon, E., Hansen, C. S., Hansen, T. F., Herms, S., Hickie, I. B., Hoffmann, P., Homuth, G., Horn, C., Hottenga, J.-J., Hougaard, D. M., Hu, M., Hyde, C. L., Ising, M., Jansen, R., Jin, F., Jorgenson, E., Knowles, J. A., Kohane, I. S., Kraft, J., Kretzschmar, W. W., Krogh, J., Kutalik, Z., Lane, J. M., Li, Y., Li, Y., Lind, P. A., Liu, X., Lu, L., MacIntyre, D. J., MacKinnon, D. F., Maier, R. M., Maier, W., Marchini, J., Mbarek, H., McGrath, P., McGuffin, P., Medland, S. E., Mehta, D., Middeldorp, C. M., Mihailov, E., Milaneschi, Y., Milani, L., Mill, J., Mondimore, F. M., Montgomery, G. W., Mostafavi, S., Mullins, N., Nauck, M., Ng, B., Nivard, M. G., Nyholt, D. R., O'Reilly, P. F., Oskarsson, H., Owen, M. J., Painter, J. N., Pedersen, C. B., Pedersen, M. G., Peterson, R. E., Pettersson, E., Peyrot, W. J., Pistis, G., Posthuma, D., Purcell, S. M., Quiroz, J. A., Qvist, P., Rice, J. P., Riley, B. P., Rivera, M., Saeed Mirza, S., Saxena, R., Schoevers, R., Schulte, E. C., Shen, L., Shi, J., Shyn, S. I., Sigurdsson, E., Sinnamon, G. B. C., Smit, J. H., Smith, D. J., Stefansson, H., Steinberg, S., Stockmeier, C. A., Streit, F., Strohmaier, J., Tansey, K. E., Teismann, H., Teumer, A., Thompson, W., Thomson, P. A., Thorgeirsson, T. E., Tian, C., Traylor, M., Treutlein, J., Trubetskoy, V., Uitterlinden, A. G., Umbricht, D., Van der Auwera, S., van Hemert, A. M., Viktorin, A., Visscher, P. M., Wang, Y., Webb, B. T., Weinsheimer, S. M., Wellmann, J., Willemsen, G., Witt, S. H., Wu, Y., Xi, H. S., Yang, J., Zhang, F., eQTLGen, 23andMe, Arolt, V., Baune, B. T., Berger, K., Boomsma, D. I., Cichon, S., Dannlowski, U., de Geus, E. C. J., DePaulo, J. R., Domenici, E., Domschke, K., Esko, T., Grabe, H. J., Hamilton, S. P., Hayward, C., Heath, A. C., Hinds, D. A., Kendler, K. S., Kloiber, S., Lewis, G., Li, Q. S., Lucae, S., Madden, P. F. A., Magnusson, P. K., Martin, N. G., McIntosh, A. M., Metspalu, A., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Nordentoft, M., Nöthen, M. M., O'Donovan, M. C., Paciga, S. A., Pedersen, N. L., Penninx, B. W. J. H., Perlis, R. H., Porteous, D. J., Potash, J. B., Preisig, M., Rietschel, M., Schaefer, C., Schulze, T. G., Smoller, J. W., Stefansson, K., Tiemeier, H., Uher, R., Völzke, H., Weissman, M. M., Werge, T., Winslow, A. R., Lewis, C. M., Levinson, D. F., Breen, G., Børglum, A. D., Sullivan, P. F., and Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2018), "Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression," *Nature genetics*, 50, 668–681.

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T.-H., D'Ambrosio, D., Gallins, P., Ha, M. J., Hottenga, J. J., Huang, S., Kattenberg, M., Kochar, J., Middeldorp, C. M., Qu, A., Shabalin, A., Tischfield, J., Todd, L., Tzeng, J.-Y., van Grootheest, G., Vink, J. M., Wang, Q., Wang, W., Wang, W., Willemsen, G., Smit, J. H., de Geus, E. J., Yin, Z., Penninx, B. W. J. H., and Boomsma, D. I. (2014), "Heritability and genomics of gene expression in peripheral blood," *Nature genetics*, 46, 430–437.

Wu, Y., Burch, K. S., Ganna, A., Pajukanta, P., Pasaniuc, B., and Sankararaman, S. (2022), "Fast estimation of genetic correlation for biobank-scale data," *American journal of human genetics*, 109, 24–32.

Wu, Y., and Sankararaman, S. (2018), "A scalable estimator of SNP heritability for biobank-scale data," *Bioinformatics*, 34, i187–i194.

Xiong, X., Hou, L., Park, Y. P., Molinie, B., GTEx Consortium, Gregory, R. I., and Kellis, M. (2021), "Genetic drivers of m^{6}A methylation in human brain, lung, heart and muscle," *Nature genetics*, 53, 1156–1165.

Yang, C., Farias, F. H. G., Ibanez, L., Suhy, A., Sadler, B., Fernandez, M. V., Wang, F., Bradley, J. L., Eiffert, B., Bahena, J. A., Budde, J. P., Li, Z., Dube, U., Sung, Y. J., Mihindukulasuriya, K. A., Morris, J. C., Fagan, A. M., Perrin, R. J., Benitez, B. A., Rhinn, H., Harari, O., and Cruchaga, C. (2021), "Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders," *Nature neuroscience*, 24, 1302–1312.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., and Visscher, P. M. (2012), "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits," *Nature genetics*, 44, 369–375.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011), "GCTA: a tool for genome-wide complex trait analysis," *American journal of human genetics*, 88, 76–82.

Zhai, J., Kim, J., Knox, K. S., Twigg III, H. L., Zhou, H., and Zhou, J. J. (2018), "Variance Component Selection With Applications to Microbiome Taxonomic Data," *Frontiers in microbiology*, 9, 509.

Zhang, Y., Yang, H. T., Kadash-Edmondson, K., Pan, Y., Pan, Z., Davidson, B. L., and Xing, Y. (2020), "Regional Variation of Splicing QTLs in Human Brain," *American journal of human genetics*, 107, 196–210.

Zhang, Z., Luo, K., Zou, Z., Qiu, M., Tian, J., Sieh, L., Shi, H., Zou, Y., Wang, G., Morrison, J., Zhu, A. C., Qiao, M., Li, Z., Stephens, M., He, X., and He, C. (2020), "Genetic analyses support the contribution of mRNA N^6-methyladenosine (m^{6}A) modification to human disease heritability," *Nature genetics*, 52, 939–949.

Zhou, H., Sinsheimer, J. S., Bates, D. M., Chu, B. B., German, C. A., Ji, S. S., Keys, K. L., Kim, J., Ko, S., Mosher, G. D., Papp, J. C., Sobel, E. M., Zhai, J., Zhou, J. J., and Lange, K. (2020), "OPENMENDEL: a cooperative programming project for statistical genetics," *Human genetics*, 139, 61–71.

Zhou, H., Hu, L., Zhou, J., and Lange, K. (2019), "MM Algorithms For Variance Components Models," *Journal of computational and graphical statistics*, 28, 350–361.

Zhu, Z., Bakshi, A., Vinkhuyzen, A. A. E., Hemani, G., Lee, S. H., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., LifeLines Cohort Study, Esko, T., Milani, L., Mägi, R., Metspalu, A., Hill, W. G., Weir, B. S., Goddard, M. E., Visscher, P. M., and Yang, J. (2015), "Dominance genetic variation contributes little to the missing heritability for human complex traits," *American journal of human genetics*, 96, 377–385.

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., and Yang, J. (2016), "Integration of summary data from GWAS

and eQTL studies predicts complex trait gene targets," *Nature genetics*, 48, 481–487.