Title
Functional genomics study of neuropsychiatric disorders through integration of chromatin regulation, transcriptomics, and metabolomics

Author
Boltz, Toni

Publication Date
2023

UNIVERSITY OF CALIFORNIA

Los Angeles

Functional genomics study of neuropsychiatric disorders through integration of chromatin regulation, transcriptomics, and metabolomics

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

by

Toni Ann Boltz

2023

ABSTRACT OF THE DISSERTATION

Functional genomics study of
neuropsychiatric disorders through integration of
chromatin regulation, transcriptomics, and metabolomics

by

Toni Ann Boltz
Doctor of Philosophy in Human Genetics
University of California, Los Angeles, 2023
Professor Roel A. Ophoff, Chair

This dissertation examines significant risk loci identified through GWAS for neuropsychiatric illness to prioritize causal variants and genes for future validation in functional assays. This investigation focuses on the impact of risk loci on molecular traits, and whether these functionally relevant variants are associated with GWAS-SNPs for neuropsychiatric phenotypes. Chapter 1 explores the integration of three complementary next-generation sequencing approaches, including genotyping, gene expression, and chromatin accessibility in hundreds of fibroblast cell lines of a multi-ancestry cohort of bipolar disorder patients and controls. Chapter 2 investigates the blood transcriptomes of a larger cohort, including a broader spectrum of bipolar disorder and schizophrenia diagnoses, and how application of computational cell type

deconvolution methods can further highlight potentially relevant genes. Chapter 3 delves into metabolomics of cerebrospinal fluid in Alzheimer's patients and healthy controls, allowing for the investigation of neurobiological mechanisms *in vivo*. We find that the integration of multiple omics levels significantly enhances our understanding of GWAS risk loci by uncovering the functional consequences and molecular mechanisms of disease-associated loci.

The dissertation of Toni Ann Boltz is approved.

Bogdan Pasaniuc

Valerie A. Arboleda

Noah A. Zaitlen

Roel A. Ophoff, Committee Chair

University of California, Los Angeles

2023

DEDICATION

Pentru familia mea.

For my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work could not have been done without the invaluable support of many people. Firstly, I am so grateful to Dr. Roel Ophoff for his mentorship throughout my graduate journey and for providing me with every opportunity to grow as a researcher. His guidance, encouragement, and unwavering support have been instrumental in shaping my research and academic growth - I am forever glad that he suggested I skip my third rotation and join his lab early. I would also like to thank Dr. Bogdan Pasaniuc for his expertise on the computational methods used throughout this dissertation, our meetings with him played a pivotal role in the success of these projects. Thank you as well to Dr. Noah Zaitlen and Dr. Valerie Arboleda for their encouraging feedback and insightful advice on my doctoral committee over the years.

Next, I would like to thank the members of the Ophoff lab for not only their extraordinary contributions to the science presented here, but also for their friendship and camaraderie over the years. I am so appreciative of Merel Bot, who not only did the entirety of the wet lab work for the datasets analyzed here (on top of being the lab manager and piloting the COVID wastewater program at UCLA), but also provided unwavering emotional support and encouragement throughout my time in LA. I truly could not imagine life in graduate school without Merel, and I'm so grateful for our close friendship that has blossomed over the years. I would also like to thank Dr. Lianne Reus, who's unparalleled passion for and expertise in neurodegeneration, as well as her ability to infuse fun into our work, made our collaborative efforts not only productive but also enjoyable. Thank you as well to Marcelo Francia, not only for his outstanding work across multiple projects, but also for the multitude of laughs we've shared over the years. Thank you to Dr. Lingyu Zhan, Juan de la Hoz, Naren Ramesh, Carolinne

Alvarado, Kevin Wojta, Dr. Artemis Zavaliangos-Petropulu, I am so grateful to all of you for making my time in the Ophoff lab so special.

I'd like to thank the members of the Bogdan lab for their collaborations and incredible support. Specifically, I'd like to thank Dr. Tommer Schwarz who so kindly welcomed me on my first day rotating in the Bogdan lab; the RNA-seq analyses done in this dissertation were a product of his remarkable efforts (and many, many Slack messages!) Thank you as well to Sandra Lapinska for her enthusiasm and willingness to help with analyses across both RNA-seq projects. Thank you to Dr. Arjun Bhattacharya and Kangcheng Hou for their technical expertise on various methods used throughout the dissertation. Thank you to Dr. Malika Kumar-Freund who helped me navigate my first year of graduate school - she was the one who explained the basics of genotyping and QTLs in a way that made it click. Thank you also to Gayane Hovhannisyan who helped so many times with (the sometimes impossible task of) scheduling meetings.

I'm so grateful to many UCLA alumni who have offered me help and kindness, including Dr. Kristina Garske, who spent many hours explaining the basics of ATAC-seq data analysis to me; Dr. Loes Olde Loohuis, who provided vital technical guidance on every project in this dissertation; and Dr. Jen Forsyth, who gave me amazing opportunities to learn polygenic risk scoring.

Thank you to BIG Summer, the undergraduate research program that I had the opportunity to participate in back in 2017. This program gave me the chance to see what life as a graduate student at UCLA would be like, and opened the door for me to pursue the PhD program here just a year later. I've also had the chance to mentor several BIG Summer students during my

time as a graduate student, and it’s been incredible getting to see such great talent come through this program.

I am so grateful to my life-long best friend Tanner Waters. Having such solid emotional support from the day we packed the car to drive out from Florida to California, all the way through to now, I don’t know if I could have done it without you. Thank you for always giving me the tough love I need, even if I don’t always seem super grateful in the moment, you really have had such a profound and wonderful impact on my life. I’m going to miss our life at Darlington so much, but I’m also so excited to see what the future holds for us. I must also give a shoutout to our cat Vinny who has brought so much joy to our home since we adopted him in 2020. Rest in peace, sweet angel.

Finally, I would like to thank my family. I am so lucky to have such amazing parents and grandparents who have raised me to be the person I am today. Thank you to my mom, who has always pushed me to do my very best, and who dropped everything to come with me to WCPG in Italy and supported me when I needed it most. Thank you to my dad who constantly reminds me that the most important thing in life is to be happy. Thank you to my grandparents for the love and support they’ve showered me with throughout my entire life. Mulțumesc din suflet, Papu și Maia, că m-ai crescut să devin femeia care sunt astăzi. I want to express my heartfelt gratitude to my sister Summer for always being there for me. Your wisdom, which seems beyond your years, makes me sometimes forget that you're my younger sibling. Thank you for being an amazing sister and a source of inspiration, I'm so excited to witness your journey in becoming the next Dr. Boltz. To my partner Evan, I am so incredibly grateful to have you by my side. I want to share my deepest gratitude for the love and comfort you have brought into my life

every single day since we met. I'm so excited to be moving over to the east coast with you and embark on the next chapter of our journey together.

CURRICULUM VITAE

EDUCATION

2018 – Current	PhD Candidate, Human Genetics University of California, Los Angeles | Los Angeles, CA
2014 – 2017	BS, Biochemistry and Molecular Biology BS, Computer Science University of Maimi | Coral Gables, FL

PUBLICATIONS

Boltz, Toni; Schwarz, Tommer; Bot, Merel; Hou, Kangcheng; Caggiano, Christa; Lapinska, Sandra; Duan, Chenda; Boks, Marco; Kahn, Rene; Zaitlen, Noah; Pasaniuc, Bogdan; Ophoff, Roel. (2023). Cell type deconvolution of bulk blood RNA-Seq to reveal biological insights of neuropsychiatric disorders. bioRxiv : the preprint server for biology. 10.1101/2023.05.24.542156. *Under revision for publication in the American Journal of Human Genetics.*

Boltz, Toni; Bot, Merel; Schwarz, Tommer; Lapinska, Sandra; Hou, Kangcheng; Garske, Kristina; Freund, Malika; Friemer, Nelson; Boks, Marco; Khan, Rene; Pasaniuc, Bogdan; Ophoff, Roel. (2023). Multi-omics study of primary fibroblast cell lines reveals shared allelic effects between ancestries. *In preparation for submission.*

Reus, Lianne; **Boltz, Toni;** Francia, Marcelo; Bot, Merel; Ramesh, Naren; Koromina, Maria; van der Flier, Wiesje F.; Jelle Visser, Pieter; van der Lee, Sven; Tijms, Betty M; Teunissen, Charlotte E.; Olde Loohuis, Loes; Ophoff, Roel. (2023). Quantitative trait loci mapping of

circulating metabolites in cerebrospinal fluid to uncover biological mechanisms involved in brain-related phenotypes. *In preparation for submission.*

Francia, Marcelo; Bot, Merel; **Boltz, Toni**; De la Hoz, Juan; Boks, Marco; Kahn, Rene; Ophoff, Roel. (2023). Fibroblasts as an in vitro model of circadian genetic and genomic studies. bioRxiv : the preprint server for biology. 10.1101/2023.05.19.541494. *Submitted and under review for publication.*

Schwarz, Tommer; **Boltz, Toni;** Hou, Kangcheng; Bot, Merel; Duan, Chenda; Olde Loohuis, Loes; Boks, Marco; Kahn, René; Ophoff, Roel; Pasaniuc, Bogdan. (2022). Powerful eQTL mapping through low coverage RNA sequencing. Human Genetics and Genomics Advances. 3. 100103. 10.1016/j.xhgg.2022.100103.

Reus, Lianne; Pasaniuc, Bogdan; Posthuma, Danielle; **Boltz, Toni;** Pijnenburg, Yolande; Ophoff, Roel. (2021). Gene Expression Imputation Across Multiple Tissue Types Provides Insight Into the Genetic Architecture of Frontotemporal Dementia and Its Clinical Subtypes. Biological Psychiatry. 89. 10.1016/j.biopsych.2020.12.023.

Introduction

Genome-wide association studies (GWAS) have successfully discovered genetic loci associated with the risk of developing complex brain-related disorders and diseases. Hundreds of independent loci with small effects have been reported for psychiatric conditions like bipolar disorder (BD)[1,2] and schizophrenia (SCZ).[2] Fewer loci have been reported for neurodegenerative disease such as Alzheimer's disease (AD),[3] including the particularly large effect of the APOE risk locus; the remaining loci tend to have similar magnitude of effects as their psychiatric disorder-associated counterparts. The vast majority of GWAS-significant loci are in non-coding regions of the genome and as such, the causal mechanism between the genetic variation and risk for these phenotypes remains unknown. These phenotypes are highly heritable, meaning that within a given population, there is a high proportion of phenotypic variance that is due to genetic factors.[4] Heritability estimates based on family studies range from 60-80% for BD and SCZ, yet, the SNP-based heritability is currently estimated to be about 25% for both disorders.[12] Similarly, AD is also estimated to be highly heritable at 60-80%,[3] though SNP-based heritability is estimated at 25-50% with about a quarter of the heritability attributable to the *APOE* locus.[5] The gap in heritability estimates suggests that while genetic contributions are playing a role, there is "missing heritability" that is unaccounted for in each of these phenotypes.

Neuropsychiatric conditions can be severely debilitating to those afflicted, and current treatment options are often variable at best in their efficacy, since relevant molecular mechanisms remain elusive. The symptoms of patients with BD include periods of mania and depression and irrational thoughts or decision-making; though given the overlap in symptomatology it is not uncommon for misdiagnosis to occur across severe mental illnesses like SCZ or major depressive disorder. Such misdiagnoses contribute to lower power to detect associations in

case-control studies. Globally, about 1-3% of the population suffers from bipolar disorder (about 2.8% in the United States),[6] and while schizophrenia is slightly less prevalent at around 1%,[7] other mental illnesses like depression and anxiety are becoming much more common, especially after the COVID19 pandemic.[8] Within the United States alone, there is a huge economic and humanistic burden attributed to bipolar disorder.[9] Direct costs including hospitalizations, clinic visits, and trying out various (often ineffective) pharmaceuticals are estimated around $46 billion annually, almost double the direct costs for the general population.[10] Indirect costs, including loss of productivity, unemployment, and caregiver burden are the main driver of economic burden, estimated to be about $156 billion annually.[9] Studies have also shown that there is a substantial decrease in the health-related quality of life for bipolar disorder patients, especially during depressive episodes.[9,11] These burdens emphasize the pressing need for better treatment options and outcomes for individuals suffering from bipolar disorder and related mental health conditions.

Similarly, while neurological diseases like Alzheimer's or frontotemporal dementia (FTD) have useful biomarkers, the manifestation of symptoms and specific underlying pathways remain unclear. Early symptoms of Alzheimer's and other forms of dementia include changes to personality, cognitive decline, and inability to create new memories or recognize people or objects, and eventually the buildup of amyloid-beta plaques and tau tangles in the brain lead to impaired communication, confusion, poor judgment, as well as difficulty with basic motor skills like speaking, swallowing, and walking.[12] AD is the most common form of dementia, accounting for 60-80% of cases, with the prevalence estimated at about 10% for individuals 65 and older, and this number is expected to rise assuming no major breakthroughs in treatment options.[12] Caring for AD patients, either at home through family members (usually unpaid) or in assisted living facilities / hospitals, incurs a massive economic burden, with both settings each estimated at around $340 billion in costs annually, again underscoring the urgent need for therapeutics.[12]

The application of genome-wide association studies to molecular quantitative traits has provided mechanistic insights into complex disease architectures.[13,14] However, these insights lag behind particularly for brain-related traits due to the inaccessibility of living brain tissue. Post-mortem tissue gene expression has been shown to be very different from living brain tissue gene expression,[15] thus there is a need for accessible tissue or biofluid samples from living donors. While gene expression is not highly correlated between different tissue types,[16] the cis-genetic effects - defined here as effects of variants on genes located within one megabase - are highly correlated,[16,17] suggesting the potential to still gain useful information from more accessible procedures such as skin biopsies or blood draw. Relatedly, cerebrospinal fluid (CSF) is not nearly as simple to obtain as skin or blood samples, though the relative safety of lumbar puncture makes it more readily accessible than brain tissue sampling from living donors. Given that CSF circulates around the brain and central nervous system, samples of this fluid can serve as a proxy for studying neurobiological mechanisms *in vivo.*[18]

My dissertation investigates significant risk loci identified through GWAS for neuropsychiatric illness to prioritize causal variants and genes for future validation in functional assays. This study focuses on the impact of risk loci on molecular traits, and whether these functionally relevant variants are associated with GWAS-SNPs for neuropsychiatric phenotypes. Though the causal effects of risk alleles on gene expression are most likely to be neuronal, we hypothesize that the genetic regulation is shared across tissues and will have the power to reveal relevant pathways. Numerous previous studies[16,17] have demonstrated the substantial concordance between blood and brain gene expression quantitative trait loci (QTL), and emphasize the significant gains in power to detect effects when large sample sizes of blood are used, especially when considering the difficulties with obtaining large sample sizes of brain tissue or limitations of post-mortem samples. Furthermore, while fibroblasts are not the preferred cell type

for studying psychiatric disease susceptibility, cis-genetic effects are generally shared across cell types.[17,19] The goal of this project is to integrate data from genetic variation, gene expression, chromatin accessibility, and metabolomics with known GWAS loci to provide molecular insights of the mechanisms that contribute to genetic risk of neuropsychiatric phenotypes.

Chapter 1 explores the integration of three complementary next-generation sequencing approaches, including SNP-genotyping, ATAC-seq, and RNA-seq in hundreds of fibroblast cell lines of a multi-ancestry cohort of BD patients and controls. QTL analysis of molecular data has identified genetic variants associated with traits such as gene expression, and colocalization of these functional QTL with GWAS risk loci has offered insights into the genetic basis of complex diseases. In this study, we employed gene expression (RNA-seq) and chromatin accessibility (ATAC-seq) datasets obtained from human primary fibroblasts to investigate QTLs in cohorts ascertained for bipolar disorder of European (n=150) and admixed American (n=96) ancestry. Our findings revealed a concordance of QTL effect sizes between European and American ancestry populations, indicating shared genetic architecture underlying gene expression and chromatin accessibility in primary fibroblasts. The integration of chromatin data with expression and genotypes allowed for finemapping of the eQTL pathways and importantly, we found that most SNPs and open regions of chromatin do not regulate its most proximal gene, highlighting the importance of including multi-omics levels. We then used this data to perform transcriptome-wide association (TWAS) and chromatin-wide association studies (CWAS) with brain-related and skin-related GWAS, identifying potentially causal gene-trait and chromatin-trait associations. The observed concordance of QTL effect sizes supports the notion of shared genetic regulatory mechanisms across ancestries in these cells. However, our results also emphasize the importance of having ancestry-specific reference panels for TWAS and CWAS, enhancing the reliability of genotype-phenotype associations. This study demonstrates the utility

of integrating RNA-seq and ATAC-seq data from human primary fibroblasts to uncover and fine-map QTLs in populations of European and American ancestry and contributes to a better understanding of the genetic basis of complex traits and diseases in diverse populations.

Chapter 2 investigates the blood transcriptomes of a larger cohort, including a broader spectrum of BD and SCZ diagnoses, and how application of computational cell type deconvolution methods can further highlight potentially relevant genes. While expression QTL analysis of bulk tissue is a common approach to decipher underlying mechanisms, this can obscure cell-type specific signals and mask trait-relevant mechanisms. While single-cell sequencing can be prohibitively expensive in large cohorts, computationally inferred cell type proportions and cell type gene expression estimates have the potential to overcome these problems and advance mechanistic studies. Using bulk RNA-Seq from 1,730 samples derived from whole blood in a cohort ascertained for individuals with BP and SCZ, this study estimates cell type proportions and their relation with disease status and medication. We found between 2,875 and 4,629 eQTL-associated genes (eGenes) for each cell type, including 1,211 eGenes that are not found using bulk expression alone. We performed a colocalization test between cell type eQTLs and various traits and identified hundreds of associations between cell type eQTLs and GWAS loci that are not detected in bulk eQTLs. Finally, we investigated the effects of lithium use, one of the main medications prescribed to treat the symptoms of bipolar disorder, on cell type expression regulation and found examples of genes that are differentially regulated dependent on lithium use. This study suggests that computational methods can be applied to large bulk RNA-Seq datasets of non-brain tissue to identify disease-relevant, cell type specific biology of psychiatric disorders and psychiatric medication. A version of this study is under revision at the American Journal of Human Genetics.

Chapter 3 delves into metabolomics of cerebrospinal fluid in Alzheimer's patients and healthy controls, which allows for the investigation of neurobiological mechanisms *in vivo*. In this study, we use metabolomics to measure the levels of 5,543 CSF metabolite levels, the largest panel in CSF to date, in nearly a thousand European individuals with genetic data. Individuals originated from two separate cohorts including a cognitively healthy cohort (n=490 subjects) and a well-characterized memory clinic cohort (n=487 subjects). We performed genome-wide metabolite quantitative trait loci (mQTL) mapping on CSF metabolomics and found 126 significant mQTLs, representing 65 unique CSF metabolite levels across 51 independent loci. We performed a metabolome-wide association study and colocalization analysis and identified 40 significant associations between CSF and brain traits, and similarly, we found colocalized gene-metabolite associations for over 90% of our genome-wide significant mQTL. These findings highlight metabolic pathways that may be involved in the dysregulation of neurodegenerative and psychiatric disorders.

Finally, in the Conclusion, I reflect on the progress of the field of neuropsychiatric genomics, summarize the findings and future directions from the three chapters, and discuss how the inclusion of samples from diverse ancestral populations is a necessary next step in human genomics research studies.

Chapter 1: Multi-omics study of primary fibroblast cell lines reveals shared allelic effects between ancestries

Authors: Toni Boltz[1], Merel Bot[2], Tommer Schwarz[3], Sandra Lapinska, Kangcheng Hou, Kristina Garske, Malika K. Freund, Nelson Friemer[2], Marco P. Boks[5], Rene S. Kahn[5,6], Bogdan Pasaniuc[1,3,4], Roel A. Ophoff[1,2]

1. Department of Human Genetics, David Geffen School of Medicine, UCLA

2. Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, UCLA

3. Department of Bioinformatics, David Geffen School of Medicine, UCLA

4. Department of Pathology and Laboratory Medicine, David Geffen **Authors:** Toni Boltz[1], Merel Bot[2], Tommer Schwarz[3], Sandra Lapinska, Kangcheng Hou, Kristina Garske, Malika K. Freund, Nelson Friemer[2], Marco P. Boks[5], Rene S. Kahn[5,6], Bogdan Pasaniuc[1,3,4], Roel A. Ophoff[1,2]School of Medicine, UCLA

5. Department of Psychiatry, Brain Center University Medical Center Utrecht, University Utrecht, Utrecht, the Netherlands

6. Department of Psychiatry, Icahn School of Medicine, Mount Sinai, NY, USA

INTRODUCTION

In recent years, genome-wide association studies (GWAS) have reached sample sizes in the millions of individuals, yet nearly 80% of individuals included in these studies are of European descent.[20] Similarly, databases which provide omics-level data, such as GTEx, generally consist mostly of European participants.[21] Previous studies[22,23,24] have shown that differences in linkage disequillibrium (LD) and allele frequencies can distinguish between populations of different ancestral backgrounds. Such differences have been shown to lead to false positives when European-based reference panels are used for studying non-European cohorts in genetic association studies,[25,26] thus there is a need for genetic and genomic datasets from diverse

populations. Incorporating data from diverse ancestries allows for the identification of both shared and population-specific genetic contributors to various traits, leading to a more comprehensive understanding of the underlying molecular mechanisms.

Furthermore, GWAS have successfully discovered loci associated with the risk of developing various complex diseases, including 64 independent loci currently reported for bipolar disorder.[1] However, the majority of these loci are in non-coding regions of the genome and as such, the causal mechanism between the genetic variation and disease risk often is unclear. Quantitative trait loci (QTL) studies[27] have identified non-coding SNPs that impact both gene expression and complex phenotypes, revealing mechanistic insights into disease architecture.[14,28] While fibroblasts are not the preferred cell type for studying psychiatric disease susceptibility, previous QTL studies have shown that cis-genetic effects are generally shared across cell types, with the caveat that brain tissue types are more highly correlated with each other than non-brain tissue types.[16,17] Relatedly, previous studies have shown that cis-effects on chromatin accessibility tend to be less context-dependent than on gene expression,[29] thus we expect better power to detect associations despite the tissue type. Given the ease of accessibility of a skin sample relative to a brain tissue sample, fibroblasts provide a unique opportunity to study biological samples from cohorts of hundreds of individuals.

In this study, we present a multi-ancestry cohort of bipolar disorder patients and controls with multi-omics data. Specifically, we investigate human primary fibroblasts derived from skin biopsies, originating from participants from the Netherlands (n=150), Colombia (n=50), and Costa Rica (n=46), with SNP-genotypes, gene expression via RNA-seq, and chromatin accessibility via ATAC-seq[30] (Assay for Transposon-Accessible Chromatin sequencing) data measured for all individuals, resulting in an extensive and unique dataset.

RESULTS

Population stratification apparent in SNP-genotypes but not in gene expression or chromatin accessibility

Given the three countries of origin within our cohort, we initially characterized the differences amongst these individuals at the SNP, gene, and accessible-chromatin levels, which were used to compute ancestry-specific expression (e)QTLs and chromatin-acccessibility (ca)QTLs. Principal component analysis (PCA) of the imputed genotypes depicted clear and significantly different ancestry-specific clusters (PC1 P_{ANOVA} = 5.6e-16, PC2 P_{ANOVA} = 5.7e-14 after correcting for batch year) (Figure 1.1A) (Table S1.1), as expected.[31] However, principal component analysis of the gene expression and peak matrices revealed stronger correlations with batch year than with ancestry (Table S1.1) (Figures 1.1B and 1.1C). This suggests that while population stratification is clearly detectable in the SNP-genotype data, it is not as impacted at the transcriptomic or accessible chromatin levels, which are more strongly impacted by batch effects. We present the QTL mapping analyses in both ancestry-specific and pooled-ancestry contexts. See Figure S1.1 for the PCA on genotypes overlayed with 1000 Genomes reference populations.

eQTL are concordant between ancestries

We first identified eQTLs within each ancestry group. Within the Dutch cohort (N=150) we identified 3,133 eGenes at an FDR of 5% (Table S1,2), versus 1,394 eGenes in the Costa Rican cohort (N=46) (Table S1.3) and 1,492 eGenes in the Colombian cohort (N=50) (Table S1.4). Subsetting to matching SNP-gene pairs, we found an R2 of 0.93 between the Dutch and Costa Rican eQTLs effects, an R2 of 0.96 between the Dutch and Colombian eQTLs effects, and an R2 of 0.95 between the Colombian and Costa Rican eQTLs effects (Figure 2A). Combining genotype and expression data from the ancestry groups into one cohort (N=246) allowed us to perform a meta-analysis with greater power to detect associations, and minimal

population stratification (lambdaGC = 1.08). This resulted in the identification of 5,258 eGenes at 5% FDR (Table S1.5).

caQTL are concordant between ancestries

For chromatin-accessibility QTLs (caQTLs), we identified 2,001 peaks with FDR-significant association to genetics in the Dutch cohort (Table S1.6), 1,292 peaks in the Costa Rican cohort (Table S1.7), and 1,625 in the Colombian cohort (Table S1.8). Subsetting to matching SNP-peak pairs, we find an R2 of 0.95 between the Dutch and Costa Rican caQTLs effects, an R2 of 0.96 between the Dutch and Colombian caQTLs effects, and an R2 of 0.96 between the Colombian and Costa Rican caQTLs effects (Figure 1.2B).

Similar to the meta-analysis eQTL mapping, combining the genotype and chromatin data from the ancestry groups into one cohort (N=246) gave better power to detect associations (lambdaGC = 1.00), resulting in 3,557 ePeaks at an FDR threshold of 5% (Table S1.9).

Very few opposite-effect eQTLs or caQTLs were found amongst these pairwise comparisons, though these few instances are likely due to differences in LD blocks and allele frequencies between these populations.

Causal effect correlations using local ancestry

To determine the degree of concordance between causal variant effect sizes across haplotype blocks, we used admix-kit[32] to compute a genetic correlation estimate, r_{admix}, for both caQTLs and eQTLs. Prior to estimation of r_{admix}, we utilized ADMIXTURE[33] to determine ancestries of interest for our admixed individuals, which showed that the average admixture proportions of the Costa Rican and Colombian individuals were approximately 62% European, 33% American, 5% West African, and 1% East Asian (Figure 1.3). Given that European and American haplotypes made up the vast majority (>95%) of the ancestry admixture for these individuals, a two-way

admixture was performed on 97 admixed individuals selected based on joint PCA with populations from the 1000 Genomes reference panel.

Then, given our assumption that cis-SNPs are more likely to affect expression than trans-SNPs, we obtained r_{admix} estimates for 39,850 caQTLs and 11,523 eQTLs. After exclusion of genes or peaks with low standardized heritability (<2.0) and low confidence interval (CI) widths (<0.5), results for 1,000 eQTLs (Table S1.10) and 987 caQTLs (Table S1.11) were meta-analyzed to obtain r_{admix} = 0.983 (C.I: [0.97, 0.996], p-value = 0.019) for caQTLs and r_{admix} = 0.956 (C.I: [0.943, 0.968], p-value = 1e-12) for eQTLs. This suggests that the causal allelic effect sizes are the same across ancestries for these QTL. The results are robust when restricting to genes with significant heritability and reasonable CI widths.

Fine-mapping of causal pathways

In order to identify potential causal paths starting from SNPs impacting the accessibility of a chromatin region which in turn leads to changes in expression of a proximal gene, we used the *pathfinder*[34] framework. Briefly, *pathfinder* uses a hierarchical statistical framework to fine-map SNPs with chromatin marks and chromatin marks with gene expression in order to predict causal paths from SNP to mark to gene expression.

Given the paired-sample design of this analysis, we decided to continue only with the pooled analysis, rather than limit power by subsetting each ancestry group. We defined 100kb regions centered around the transcriptional start sites (TSS) of the 5,258 eGenes identified in the meta-analysis eQTL analysis, though we filtered these down to regions with prior evidence of gene-chromatin and chromatin-SNP associations (see Methods), resulting in a total of 934 regions. This low retention of regions is consistent with the empirical data analysis performed in the original *pathfinder* study (17.7% here and 8.9% in Roytman et. al[34]). Of these 934 tested gene

regions, we identified almost half (n=428) with a posterior probability (PP) over 50% contained within the top ten paths in the region, and of these, 28 genes with PP>=90% in a single SNP-peak pair, suggesting high confidence that the expression of these genes is causally impacted by the fine-mapped SNP and chromatin mark (Supplementary File S1.1).

Notably, we found that most of the chromatin peaks and SNPs selected in top paths were not regulating the gene in closest proximity. Of the 428 genes with PP >= 50% contained in the top paths, we found only 94 to be regulated by its closest peak (21.9%), and 21 (4.9%) by its closest SNP (Table S1.12). Similarly, of the 28 genes with PP>=90% from a single SNP-peak pair, only 5 genes were regulated by its closest peak, and only one gene was regulated by its closest SNP (Table 1.1). This highlights the importance of including chromatin information in fine-mapping eQTL, given that a SNP or an open region of chromatin does not necessarily regulate the gene most proximal.

Transcriptome-wide associations to brain and skin-related traits

We performed a transcriptome-wide association study (TWAS) via the FUSION[35] framework. Given the differences in ancestral background within the cohort, we split the analysis into European (Dutch) and American (Colombian and Costa Rican together) subsets. The ancestry-specific analysis resulted in 412 heritable eGenes from the European analysis, and 508 heritable eGenes from the American analysis, with 67 genes overlapping between the two, a significant overlap with Fisher's exact test P = 6.2e-13.

For association to GWAS risk loci, we tested summary statistics for bipolar disorder[1] (European ancestry) and SCZ (both European[2] and Latino[36] ancestry) as well as fibroblast-related traits including UK BioBank[37] dermatological disease traits eczema and psoriasis. The association testing was performed separately per ancestry, then resulting Z scores were meta-analyzed via

inverse variance weighting (IVW), as pooling data across analyzing has been shown to lead to false positives.[38]

For bipolar disorder, we identified four significant genes with P_{IVW} <= 1e-4 (Bonferroni-corrected for 500 heritable genes), including LDL-receptor related protein *LRP11* (P_{IVW} = 2e-5) and TNF-receptor associated factor 7 *TRAF7,* P_{IVW} = 1e-81). We found the expression of the genes *TCHP* (trichoplein keratin filament binding) and *GABARAP* (GABA type A receptor-associated protein) to be significantly associated not only with BD (*TCHP* P_{IVW} = 1e-6; *GABARAP* P_{IVW} = 1e-5), but the European-ancestry SCZ GWAS (*TCHP* P_{IVW} = 8e-10; *GABARAP* P_{IVW} < 1e-300) as well. In addition, we identified the expression of the *PIGH* gene (phosphatidylinositol N-acetylglucosaminyltransferase subunit H, P_{IVW} = 4e-16) as significantly associated with the European-ancestry SCZ GWAS, and the gene *NCAM1* (neural cell adhesion molecule 1) as significantly associated with both the European-ancestry (P_{IVW} = 1e-4) and Latino-ancestry SCZ GWAS (P_{IVW} = 7e-4). The Latino-ancestry SCZ GWAS also revealed the genes *HS6ST1* (Heparan-Sulfate 6-O-Sulfotransferase 1, P_{IVW} = 9e-71) and *MICA* (MHC class I polypeptide-related sequence A, P_{IVW} = 5e-11) as significantly associated.

For the skin-related traits, we found that expression of the genes *PLD2* (phospholipase D2, P_{IVW} = 5e-9) and *RMND5B* (Required for Meiotic Nuclear Division 5 Homolog B, P_{IVW} = 1e-5) were significantly associated with psoriasis; *GINS1* (encodes for subunit 1 of GINS DNA-replication complex, P_{IVW} = 3e-94), *URGCP* (UpReGulator of Cell Proliferation, P_{IVW} = 1e-118), *JUND* (JunD transcription factor, P_{IVW} = 1e-4), and *ILF2* (InterLeukin enhancer binding Factor 2, P_{IVW} = 1e-300) significantly associated with eczema. See Figure 41.A for TWAS results with all tested traits after IVW meta analysis.

Interestingly, in comparing ancestry-specific eQTL genes (eGenes), we found overlapping significant eGenes between EUR and AMR-specific analyses for each of the tested traits, suggesting high confidence in the reliability of these eGenes. However, when comparing the ancestry-specific results or the IVW-meta analysis results with the pooled-analysis TWAS, we found almost no significant eGenes in common for any traits (only one gene-trait association, *NCAM1* with European-based SCZ GWAS, was significant in both the pooled and IVW-meta analyses), providing further evidence that pooled-ancestry analysis may be unreliable in the context of TWAS. See Supplementary File S1.2 for all versions of TWAS results, including SNP-based heritability for each gene.

Chromatin-wide associations to brain and skin-related traits

We performed a chromatin-wide association study (CWAS,[28] also called cis-trome wide) by again leveraging the FUSION-TWAS framework to associate regulatory elements with GWAS summary statistics. For association to GWAS risk loci, we used summary statistics for the same brain-related traits and dermatological traits as the TWAS analysis. Also paralleling the TWAS analysis, the association testing was performed separately per ancestry then meta-analyzed via inverse variance weighting (IVW).

We identified 11 P_{IVW}-significant open chromatin regions with association to bipolar disorder; 18 associated open chromatin regions for European-ancestry schizophrenia; and three associated open chromatin regions for Latino-ancestry schizophrenia. For the skin-related traits, we found four open chromatin regions significantly associated with psoriasis, and eight chromatin regions significantly associated with eczema. See Figure 1.4B for CWAS results after IVW meta analysis. See Supplementary File S1.3 for all CWAS results, including peak heritability.

DISCUSSION

In this study, we performed QTL analysis on gene expression and chromatin accessibility data from fibroblast cells of a multi-ancestry bipolar cohort. QTL effect sizes were found to be concordant between the European and admixed American populations, suggesting that the genetic variants that affect either gene expression or chromatin accessibility tend to have similar effects across these populations. However, while the effect sizes may be consistent, the underlying genetic architecture and linkage disequilibrium (LD) patterns can vary significantly between ancestries,[39] contributing to the clustering of ancestral groups in principal component analysis of genetic data. For more reliable integration of molecular data into GWAS, it is necessary to match the genetic ancestral background to functional omics datasets and LD reference panels. This approach not only improves the accuracy of genomic studies[39,20] but also fosters inclusivity and furthers genomic research efforts on a global scale.

Multi-omics datasets have advanced the field of molecular biology by providing a comprehensive and integrated view of biological systems at various levels. By collecting genomics, transcriptomics, and chromatin data across hundreds of individuals, we and others have shown that such datasets can unravel molecular mechanisms underlying biological processes. Such comprehensive analyses facilitate in deciphering complex biological patterns and predicting interactions at the molecular level. Regarding our TWAS and CWAS analyses, we find over double the number of chromatin peaks (45) than genes (20) associated with the tested traits. This is consistent with the prostate cancer findings in the original CWAS method paper,[28] providing further evidence to the idea that gene expression tends to be more context-dependent than chromatin accessibility.[29] However, none of the CWAS-significant peaks overlapped with peaks found in the top pathways (PP>=50%), thus we were unable to further map these to potential genes aside from using the gene most proximal to the peak locus.

Relatedly, the closest genes to the CWAS-significant peaks have not been previously associated with brain-related or skin-related GWAS loci.

As an illustrative example, the *MICA* gene (major histocompatibility complex (MHC) class I polypeptide-related sequence A) was significantly associated with the Latino GWAS for schizophrenia in our TWAS analysis. *MICA* has been previously found to be associated with schizophrenia,[40,41] and integrating the expression measures for this gene region with chromatin data suggested a causal pathway with posterior probability >0.98 from rs2442724 to a locus of open chromatin in the MHC region (chr6:31,367,110-31,369,941), upstream of the *MICA* gene (chr6:31,368,488-31,383,092). While the MHC region is known to be difficult to parse given extensive LD,[42] the large cohort of paired transcriptome/cistrome data allowed for fine-mapping of the regulation of this particular gene. Several MHC genes have been previously reported as associated with psychiatric disorders,[43,44–46] implicating this region as a probable risk locus.

Another gene, Long Intergenic Non-Protein Coding RNA 933 (ID: ENSG00000259728, genomic location chr15:85,114,155-85,121,355), for which we identified a highly probable (PP=0.99) pathway has been previously associated with BD[1] and SCZ[2] via the PsychENCODE TWAS.[47] The causal path included SNP rs12900391 with a chromatin peak at chr15:84,542,517-84,544,020. While the exact function of this gene is not well understood, the genetic associations with BD found by the latest wave of the Psychiatric Genomics Consortium have suggested the involvement of pathways regulating insulin secretion, calcium channel activitiy, and signaling of endocannabinoids and glutamate receptors.[1] We also found a probable pathway for the gene Gamma-Aminobutyric Acid (GABA) Receptor Subunit Rho-2 (*GABRR2)*. GABA is the major inhibitory neurotransmitter in all mammalian brains, and while direct association between BD or SCZ and this gene has not yet been detected, the locus of this gene on chromosome 6q has had associations to psychiatric illness.[48–50] Similarly, we identified a

probable pathway for the *GATAD2A* gene, previously found to be associated with SCZ in the latest GWAS.[2] We identified probable pathways for other interesting brain-relevant genes including *CRCP*, which encodes a membrane protein that functions as part of a receptor complex for a small neuropeptide that increases intracellular cAMP levels in the brain.[51]

While we identify novel molecular mechanisms for genes potentially relevant to disease biology, there are several limitations to our study. Although fibroblasts can offer valuable insights into certain biological processes due to shared cis-genetic effects, their relevance to complex brain-specific pathways remains limited. However, fibroblasts have previously been used in studying circadian rhythms, a phenotype which is known to be dysregulated in bipolar disorder, suggesting that relevant molecular mechanisms are at least partially preserved.[52,53]

Secondly, this investigation focused solely on cis-QTLs, while trans-QTLS, though more difficult to ascertain, could potentially unveil additional regulatory elements involved in the brain and skin-related traits studied here. Furthermore, while our study is an important step forward in the inclusion of Latin American individuals in genomic studies, the lack of samples from individuals of African and Asian countries and other diverse ancestries represents a limitation, as genetic variants that are common to these populations may be missing or very rare in the European and American samples included here. Lastly, the scarcity of Latino-based GWAS datasets on which to perform ancestry-matched TWAS underscores the urgent need for more extensive efforts to incorporate diverse ancestral backgrounds in genetic studies. Utilizing GWAS and functional data from diverse ancestries is crucial for understanding the genetic basis of complex traits in a globally inclusive manner.

METHODS

Sequencing data collection

Skin biopsies were collected in 2010-2012 from individuals of Colombian ancestry and Costa Rican ancestry, and collected in 2013-2014 from individuals of Dutch ancestry to generate human primary fibroblasts. Fibroblasts were isolated by taking skin biopsies. Primary fibroblast cultures were established following standard procedures[54] and stored as frozen aliquots in liquid nitrogen. Fibroblasts were thawed out in batches of 8 lines at the time and grown to confluence in T75 culture flasks in standard culture media (DMEM containing 10% fetal bovine serum (FBS) and 1x Penicillin-Streptomycin). Upon reaching confluence, cells were passaged to 12 well plates at a density of $1x10^5$ for ATAC and $2x10^5$ for RNA. The next day cells were collected for further processing.

This cohort includes BD patients and unrelated healthy controls. DNA and RNA extractions from these cells were performed simultaneously from the same batch of cells in order to minimize technical artifacts that may confound the mediation analysis. Relatedly, nine of the Dutch samples were sequenced for RNA-seq and ATAC-seq in both batches and correlation of the gene expression and peak intensities was high, particularly for gene expression (Table S13). For duplicated sample pairs with high correlation (R2>0.85) for both gene expression and chromatin peak intensity, we randomly selected one in the pair of IDs to include in downstream analyses. For the two pairs with low correlation in chromatin peak intensities, quality control revealed that the second batch had significantly lower read depth for these individuals, thus we included the samples originating from the first batch in these instances.

RNA-seq data generation and processing

RNA was extracted from fibroblasts in order to assess the levels of gene expression across the genome. Cells were lysed using 350uL RLT lysis buffer from the Qiagen RNeasy mini kit. Lysed cells were then scraped off the plate, transferred to a Qiaschredder (Qiagen 79656) and centrifuged for 2 min at max speed to further homogenize. Cell lysates were kept in -80 until extraction. RNA from cell lysates was extracted using the Qiagen RNeasy mini kit (Qiagen

74106). Cell lysates were extracted in a randomized order to prevent batch effects in downstream analysis. In order to collect total RNA including small RNAs, the standard extraction protocol (Purification of Total RNA from Animal Cells using Spin Technology) was adjusted by making the following changes:

- adding 1.5 volumes of 100% ethanol, instead of 70%, after the lysis step (step 4 in handbook protocol)

- adding 700 mL of buffer RWT (Qiagen 1067933) instead of the provided RW1 (step 6 in handbook protocol)

TruSeq Stranded polyA selected library preps were generated and samples were sequenced on the Illumina HiSeq 4000 sequencer with 75-base paired end reads, at an average of 50 million mapped reads per sample. The resulting FASTQ files were pseudo-aligned to hg19 using kallisto,[55] resulting in a matrix of transcripts per million (TPMs) which were aggregated to the gene level. The expression matrix was filtered for protein-coding genes and outliers based on technical variation (n genes = 18,886 remain after filters). Principal component analysis was performed on the log-transformed counts matrix to identify and remove outlier samples. Here we include covariates for sex, first two genotype PCs, and the year of the sequencing batch. Previous studies have shown that there are likely unmeasured or “hidden” factors that reduce the power to detect associations in next-generation sequencing data, therefore we also performed PEER[56] (Probabilistic Estimation of Expression Residuals) factor analysis to find such hidden determinants of variation. PEER factors were computed separately per ancestry group, including 5 factors for the Colombian and Costa Rican groups, and 15 for the Dutch group, given the difference in sample sizes. The resulting residuals matrix that remained after accounting for all factors was then used as input for downstream analyses.

ATAC-seq data generation and processing

ATAC-seq libraries were generated as previously described (https://www.nature.com/articles/nmeth.2688). Samples were sequenced on the Illumina HiSeq 4000 sequencer with 75-base paired end reads, at an average of 39 million mapped reads per sample. Trimmed reads were aligned to the hg19 reference genome using bowtie2[57], and filtering steps were taken to remove unmapped reads, non-primary alignment, and low-quality reads, as recommended by the ENCODE standards for ATAC-seq data analysis.[58] Reads were then input into MACS2[59] in order to call peaks that are significantly enriched against the local background using a false-discovery rate (FDR) correction threshold of 0.05 and modeled by the Poisson distribution. From this initial set of peaks, blacklisted regions[60] as identified by ENCODE were removed in order to exclude regions that have anomalous or unusually high signals due to repetitive or unstructured sequences. Then, the remaining peaks across all samples were combined to form the consensus regions, using the bedtools intersect function to stitch together any regions within 147bp into one larger peak region. This resulted in over 418,000 consensus peaks, which were limited to only those peaks that are called in at least 30% of individuals, reducing the number of peaks to 77,957. These consensus regions, in conjunction with the filtered reads (bam files) per individual, were used in the R package featureCounts to determine the number of reads each individual had within each peak, with higher counts of reads per peak indicating greater accessibility of that region of the genome. This resulted in an N x P peak counts matrix where N = number of individuals and P = number of peak regions. This matrix was log-transformed to account for skewness and ensure normalization. Principal component analysis (PCA) was then performed on the log-transformed counts matrix to identify and remove outlier samples. Resulting PCs were correlated via Spearman's rank correlation against various technical factors in order to determine drivers of variation within the data that are unrelated to underlying biology. Using a Bonferroni significance level of $P < 2e\text{-}05$, we found that sex, read depth, fraction of mitochondrial DNA, TSS enrichment score, median fragment size, and fraction

of reads in peaks were correlated with the first two genotyping PCs. The genotyping and sequencing batch year was also included in order to account for batch effects.
Using the log-transformed counts matrix as the input measures and the independent technical factors identified from the PC correlations plus age as covariates, we used PEER to find hidden confounders. PEER factors were computed separately per ancestry group, including 5 factors for the Colombian and Costa Rican groups, and 15 for the Dutch group, given the difference in sample sizes. The resulting residuals matrix that remained after accounting for all factors was then used as input for the QTL analysis.

Genotyping and imputation

Samples were genotyped and imputed separately in two batches, the first consisting of 129 individuals of European ancestry genotyped via the OmniExpressExome platform, with the second batch consisting of 21 individuals of Dutch ancestry, 50 individuals of Colombian ancestry, and 46 individuals of Costa Rican ancestry genotyped via the Global Screening Array. Genotypes were first filtered for Hardy-Weinberg equilibrium p value < 1.0e-6 for controls and p value < 1.0e-10 for cases, with minor allele frequency (MAF) > 0.01. Genotypes were then imputed into the 1000 Genomes Project phase 3[61] reference panel by chromosome using RICOPILI v.1[62] separately per genotyping platform, then subsequently merged, applying an individual-missingness threshold of 10%, SNP-missingness of 5%, and MAF > 0.05 for post-merge quality control. Imputation quality was assessed by filtering variants where genotype probability > 0.8 and INFO score > 0.1 resulting in 2,747,786 autosomal SNPs that were common across both datasets.

QTL mapping

QTL analysis was performed with MatrixEQTL[63] using a cis-locus distance defined as +/-1Mb around the peak midpoint or gene TSS, initially done separately per ancestry group. We

included the identity by state (IBS) similarity matrix of the genotypes as an error covariance term within the model. Associations that remain after an FDR threshold of 5% were retained for downstream analysis. Correlations above R2 of 90% of the resulting SNP-gene or SNP-peak effect sizes per ancestry group suggested that the groups could be combined for a gain in power, and thus the analysis was repeated with all individuals together and with an added covariate term for ancestry.

Assessing ancestry-specific causal effects

Ancestry specific causal effects were calculated using admix-kit[32] on 97 admixed individuals, including 45 Costa Ricans, 50 Colombians, and 2 Dutch. These admixed individuals were determined by performing a joint PCA with the imputed genotype and 1000 Genomes Project reference panel.[61] The first 4 PCs, maximum sample distance of 1.5, sample t-range of (0.05, 0.95), and super populations EUR and AMR were used via the select-admix-indiv function in admix to select individuals to include in our analysis. To determine the super populations as well as the ancestries to use for the correlation estimates, we utilized ADMIXTURE[33] to compare our cohort with four reference populations, including Europeans (CEU), West Africans (YRI), Americans (PEL), and East Asians (CHB) from 1000 Genomes Project to determine admixture proportions with supervised analysis. With these admixed individuals, we inferred local ancestry using RFmix.[64]

The admix-kit package was used to determine the similarity among the genome-wide causal allelic effects across specified local ancestries in admixed individuals.[32] To calculate genetic correlations for each gene or peak, we built a window-based genetic relationship matrix, GRM, using a cis-locus distance defined as +/- 1Mb around each peak or gene rather than a genome-wide GRM to focus on the contribution of cis-SNPs on expression. The window-based GRM is used to estimate log-likelihood at different r_{admix} values to obtain the point estimate, credible interval, and p-value for each gene or peak. To obtain the genetic correlation across all genes or

peaks, we meta-analyze across the genes/peaks via the meta-analyze-genet-cor() function in admix.

Prior to meta-analysis, due to our low sample size, we exclude genes or peaks whose standardized heritability (hsq_est/hsq_stderr) is less than 2, confidence interval widths are below 0.5, and genes/peaks who had more than one credible interval.

Fine-mapping of SNP-chromatin-gene pathways

We used the *pathfinder* method which accounts for both SNP LD and the correlation structure between chromatin marks by using a multivariate normal distribution. This method iterates through each possible path to determine its corresponding posterior probability, thus enabling us to prioritize SNPs and chromatin peaks that mediate gene expression, which can then be prioritized for functional validation. We restricted the regions tested by taking the transcriptional start site (TSS) for each eGene and pulling out all SNPs and chromatin peaks within 50kb upstream or downstream of the TSS. These regions were then filtered via a two-stage regression analysis, wherein the gene expression values were regressed on the proximal chromatin marks, and for models with a resulting p-value less than 0.05, we regressed SNP genotypes in that region onto the residuals from the initial peak-gene regression. Any regions with at least one p-value less than 0.05/(number of SNPs in region) were retained for pathfinder analysis. All chromatin peaks within the region were correlated pairwise via the cor() function in R, and LD for all SNPs within the region was calculated through plink –r2 square. We then inputted these regions into the pathfinder.R script (https://github.com/meganroytman/pathfinder/blob/master/pathfinder.R).

TWAS

To identify eQTL associated with GWAS traits, we performed a TWAS using the FUSION[35] software (http://gusevlab.org/projects/fusion). First, we generated weights for all 5,258 FDR-

significant eGenes using the FUSION.compute_weights.R script, restricted to loci +/- 1Mb around the lead SNPs per each gene. We used the PEER-corrected gene expression thus no additional covariates were included in the model. In generating the weights, eGenes were first filtered for those with significant SNP-heritability (p-value <= 0.05).

We used the FUSION.assoc_test.R script to test for association between the gene weights and GWAS for bipolar disorder, schizophrenia (EUR and Latino ancestry), and UKBB skin-related GWAS for eczema and psoriasis. Gene-trait pairs were selected based on the best performing model after five-fold cross validation, including for Best Unbiased Linear Predictor (BLUP), elastic net (ENET), Least Absolute Shrinkage and Selection Operator (LASSO), and just using the top SNP. To account for LD structure we used an in-sample LD panel. The --coloc flag was included to perform colocalization[65] on any genes that had an association with the trait of interest with TWAS.P < 0.05.

CWAS

Similarly, to identify caQTL associated with GWAS traits, we performed a CWAS[28] using the FUSION software. First, we generated weights for all FDR-significant ePeaks using the FUSION.compute_weights.R script, restricted to loci +/- 1Mb around the lead SNPs per each gene. We used the PEER-corrected peak intensities thus no additional covariates were included in the model.

We used the FUSION.assoc_test.R script to test for association between the peak weights and GWAS for bipolar disorder, schizophrenia (EUR and Latino ancestry), and UKBB skin-related GWAS for eczema and psoriasis. Peak-trait pairs were selected based on the best performing model after five-fold cross validation, including for Best Unbiased Linear Predictor (BLUP), elastic net (ENET), Least Absolute Shrinkage and Selection Operator (LASSO), and just using the top SNP. To account for LD structure we used an in-sample LD panel. The --coloc flag was included

to perform colocalization on any peaks that had an association with the trait of interest with TWAS.P < 0.05.

Chapter 1 Figures

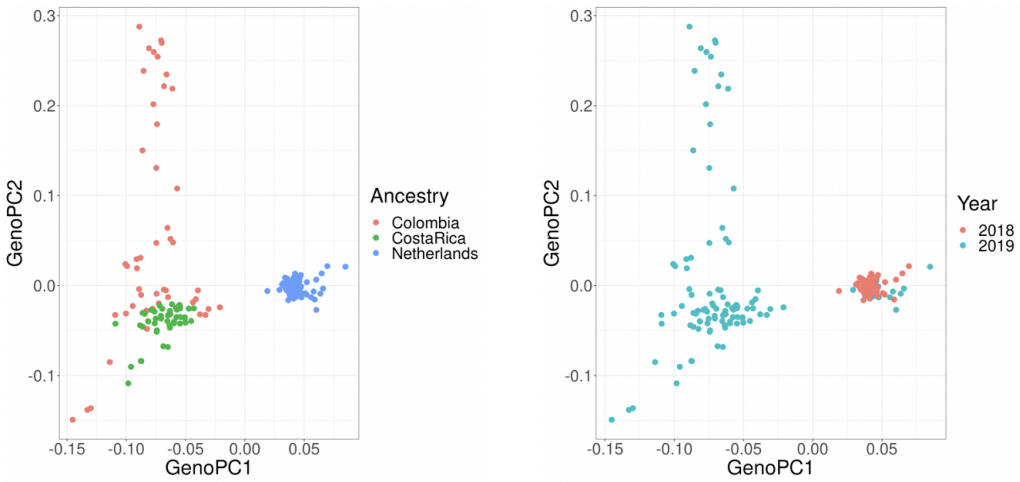
Figure 1.1. Principal component analysis of genotype, expression, and chromatin datasets.

(A) PCA on genotypes, each dot represents an individual. The left plot is colored by country of origin and the right plot is colored by sequencing batch year. PC1 explains 4.9% of the variance while PC2 explains 2.8%.

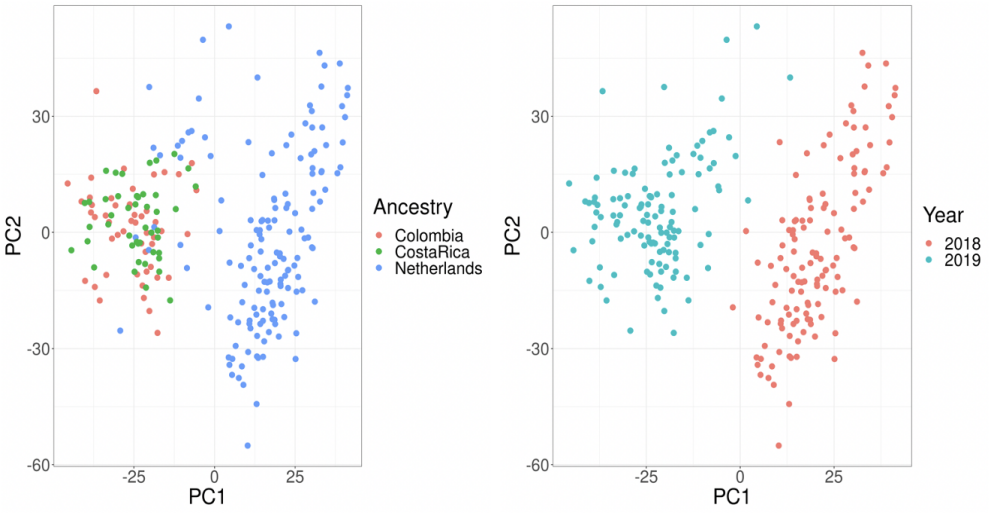
(B) PCA on gene expression. PC1 explains 10.8% of the variance while PC2 explains 6.8%. Regarding association to ancestry group, we found that PC1 P_{ANOVA} = 0.0027 and PC2 P_{ANOVA} = 0.113 after correcting for batch year.

(C) PCA on chromatin peaks. PC1 explains 72.7% of the variance while PC2 explains 3.2%. Regarding association to ancestry group, we found that PC1 P_{ANOVA} = 0.45 and PC2 P_{ANOVA} = 0.6 after correcting for batch year. Note that ATAC-seq read depth was the strongest contributing factor to the large variance in PC1.

1.1A Genotype PCA



1.1B Expression PCA



1.1C Chromatin PCA

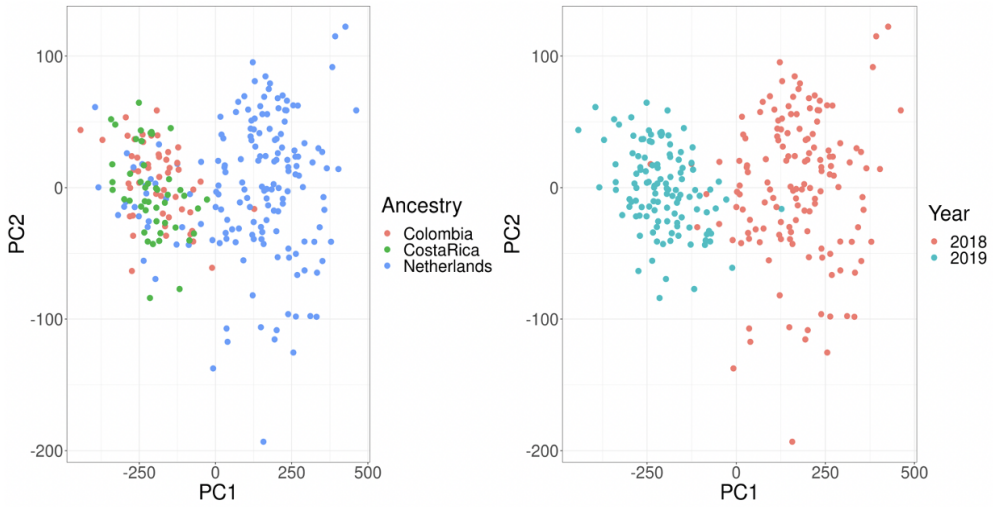


Figure 1.2. Expression and chromatin-accessibility QTLs are largely concordant between European and American populations.

(A) Each point represents a significant SNP-gene pair (note that many genes / peaks are associated with multiple SNPs, thus may be represented in multiple points).

(B) Each point represents a significant SNP-peak pair.

Figure 1.2A

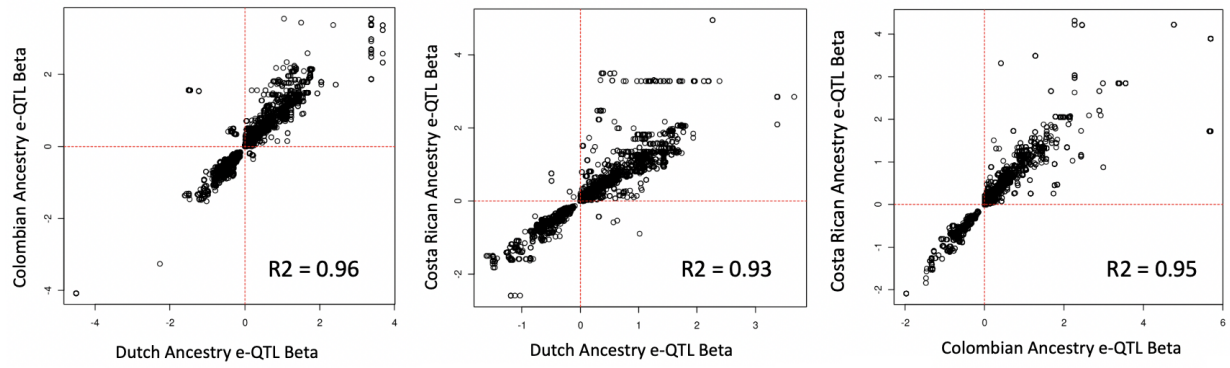


Figure 1.2B

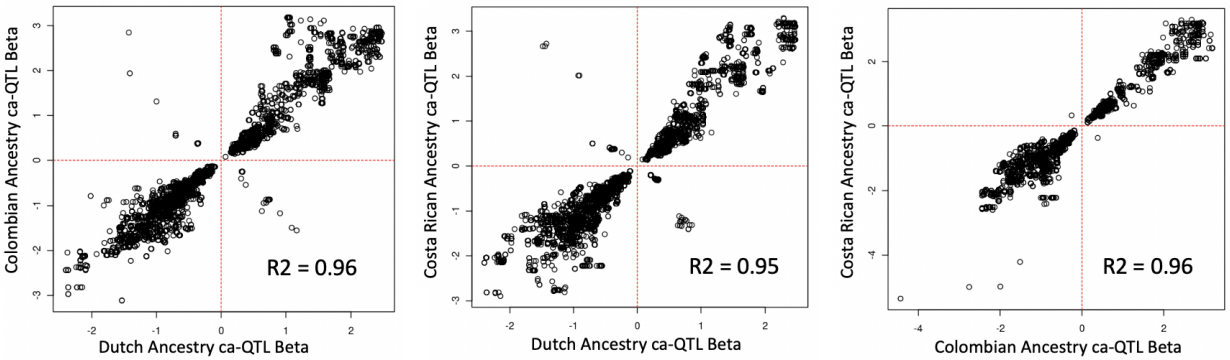


Figure 1.3. Proportions of reference populations via ADMIXTURE.

Population reference panels from 1000 Genomes include PEL (Peruvian from Lima, Peru; AMR superpopulation), CEU (Utah residents from Northern and Western Europe; EUR superpopulation), YRI (Yoruba in Ibadan, Nigera; AFR superpopulation), and CHB (Han Chinese in Bejing, China; EAS superpopulation).

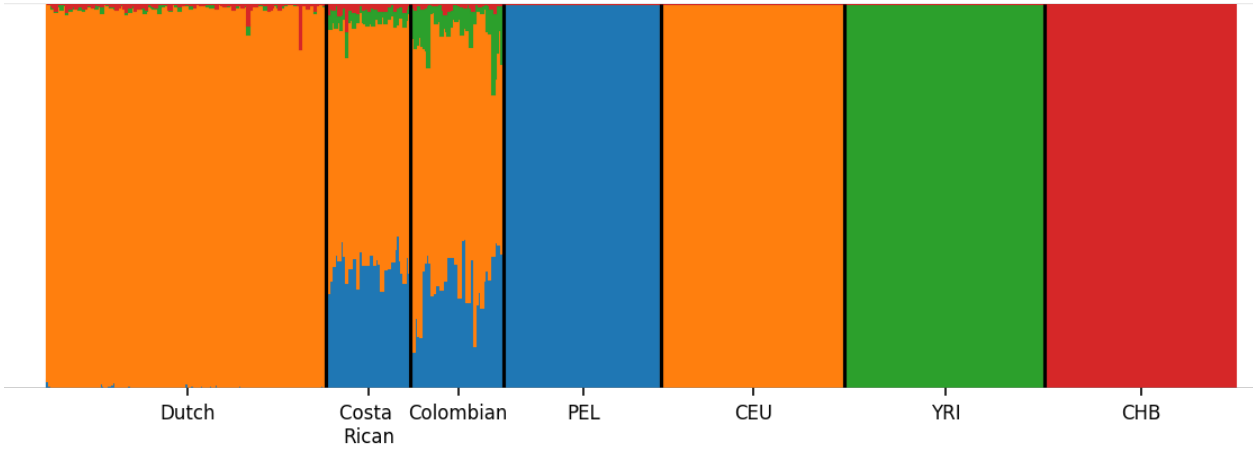
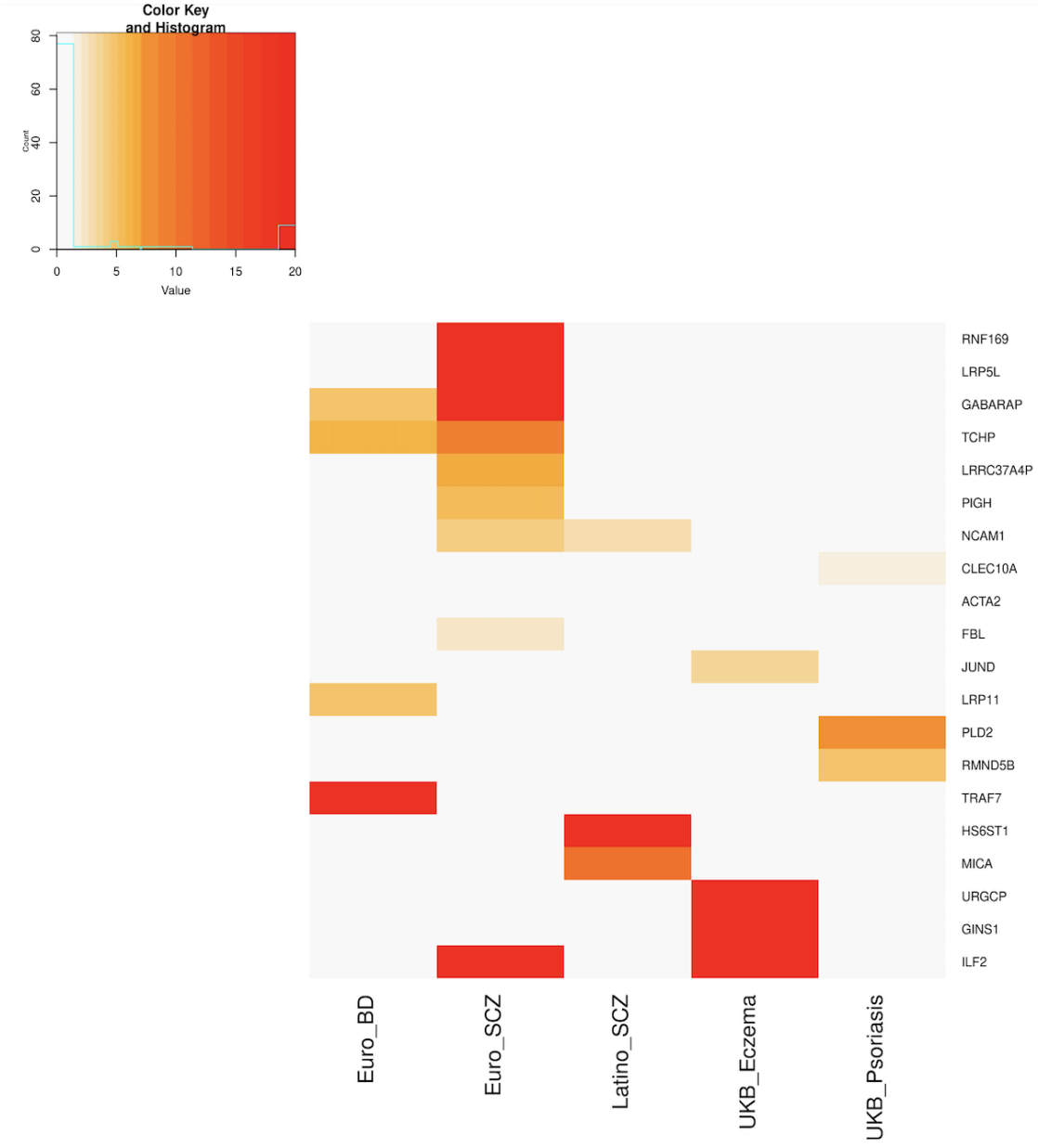
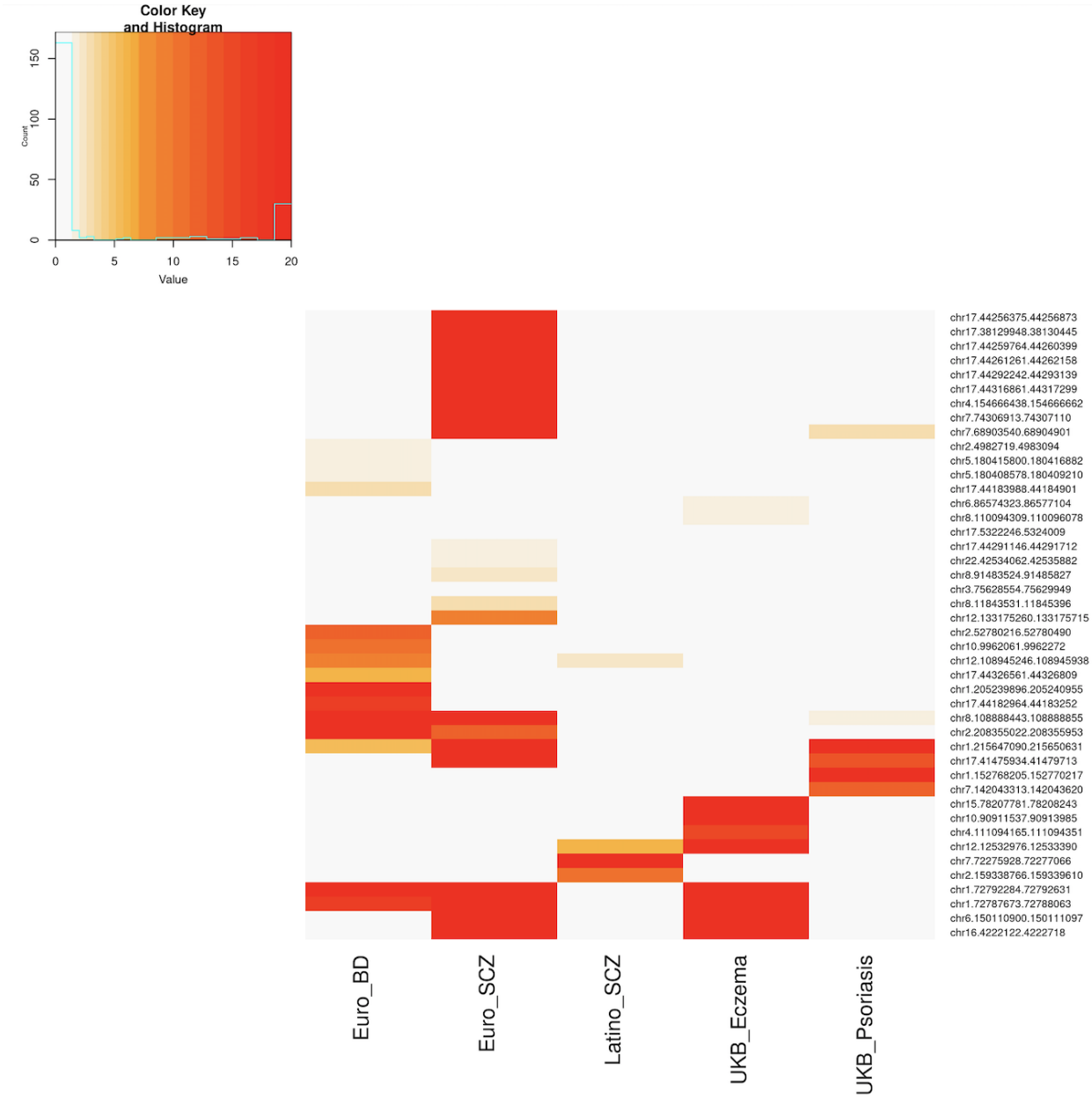


Figure 1.4. Significant TWAS and CWAS associations with brain-related and skin-related phenotypes. Each cell is colored by the degree of significance (-log10(p)) with white for non-significant associations.

(A) TWAS.



(B) CWAS.


Color Key
and Histogram
Count
0
50
100
150
0
5
10
15
20
Value
chr17.44256375.44256873
chr17.38129948.38130445
chr17.44259764.44260399
chr17.44261261.44262158
chr17.44292242.44293139
chr17.44316861.44317299
chr4.154666438.154666662
chr7.74306913.74307110
chr7.68903540.68904901
chr2.4982719.4983094
chr5.180415800.180416882
chr5.180408578.180409210
chr17.44183988.44184901
chr6.86574323.86577104
chr8.110094309.110096078
chr17.5322246.5324009
chr17.44291146.44291712
chr22.42534062.42535882
chr8.91483524.91485827
chr3.75628554.75629949
chr8.11843531.11845396
chr12.133175260.133175715
chr2.52780216.52780490
chr10.9962061.9962272
chr12.108945246.108945938
chr17.44326561.44326809
chr1.205239896.205240955
chr17.44182964.44183252
chr8.108888443.108888855
chr2.208355022.208355953
chr1.215647090.215650631
chr17.41475934.41479713
chr1.152768205.152770217
chr7.142043313.142043620
chr15.78207781.78208243
chr10.90911537.90913985
chr4.111094165.111094351
chr12.12532976.12533390
chr7.72275928.72277066
chr2.159338766.159339610
chr1.72792284.72792631
chr1.72787673.72788063
chr6.150110900.150111097
chr16.4222122.4222718
Euro_BD
Euro_SCZ
Latino_SCZ
UKB_Eczema
UKB_Psoriasis

Chapter 2: Cell type deconvolution of bulk blood RNA-Seq to reveal biological insights of neuropsychiatric disorders

Authors: Toni Boltz*#[1] and Tommer Schwarz#[2], Merel Bot[3], Kangcheng Hou[2], Christa Caggiano[2], Sandra Lapinska[2], Chenda Duan[4], Marco P. Boks[5], Rene S. Kahn[5,6], Noah Zaitlen[2,7], Bogdan Pasaniuc[1,2,8,9], Roel Ophoff*[1,2,3,10]

1. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
2. Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA, USA
3. Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA
4. Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA
5. Department of Psychiatry, Brain Center University Medical Center Utrecht, University Utrecht, Utrecht, the Netherlands
6. Department of Psychiatry, Icahn School of Medicine, Mount Sinai, NY, USA
7. Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA
8. Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
9. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
10. Department of Psychiatry, Erasmus University Medical Center, Rotterdam, the Netherlands

INTRODUCTION

One limitation of standard eQTL studies is that they generally use expression estimates from bulk tissue.[13,66] While this is informative, it has been shown that there are many cell type specific mechanisms driving biology,[19,67] which can be missed when looking at a collection of many cell

types. In recent years, single cell RNA-Seq has allowed for the profiling of the gene expression of an individual cell, giving us a clearer picture of cell type gene expression. However, single cell RNA-Seq experiments are considerably more expensive than bulk RNA-Seq[68]. To leverage the advantages of each of these approaches, we can use methods to estimate cell type gene expression from bulk RNA-Seq expression.

There exist many methods[69,70] to estimate cell type expression from bulk RNA-Seq. Here, we elected to use CIBERSORTx[71] and bMIND[72] to estimate cell type proportions and cell type expression, respectively. Computational methods for analyzing bulk gene expression data have the potential for being advantageous in some applications as it is possible to obtain much larger sample sizes using bulk RNA-Seq instead of single cell RNA-Seq. While most single cell RNA-Seq studies have sample sizes in the range of several hundreds of cells from a small number of individuals, leveraging low-coverage bulk RNA-Seq allows us to obtain samples from hundreds to thousands of subjects.[73] We used the low-coverage RNA-seq dataset described in Schwarz, et. al.[10] as the primary dataset for analysis of cell type deconvolution in this study.

Associations between immune-related traits and neuropsychiatric disorders have been previously reported[74], and we hypothesized that using blood-based expression can provide relevant information regarding the biology of such disorders.[16,75,76] In this work we used cell type deconvolution methods to derive cell type specific estimates for gene expression from bulk blood RNA-seq, specifically within a cohort including psychiatric patients and controls of European ancestry. We used these results to conduct cell type cis-eQTL analyses, and compared the shared and unique cell type associations. We show that these cell type eQTL results derived from deconvoluted bulk RNA-Seq are consistent with eQTLs from scRNA-Seq. We performed colocalization analysis to find loci driving GWAS associations in either neuropsychiatric or blood-based traits and cell type gene expression. We go on to identify

several examples of “opposite-effect” eQTLs, where a cell type eQTL signal demonstrates gene expression regulation in the opposite direction from that observed in a bulk eQTL study. Finally, we explored the effects of lithium use[77] on cell type expression, and identified several cases of lithium-SNP interaction dictating presence of an eQTL.

RESULTS

Computationally-derived cell type estimates are reliable

Figure 2.1 provides a graphical abstract of the pipeline used in this study to generate putative cell-type specific eQTLs. To estimate cell type gene expression in whole blood, we analyzed bulk blood RNA-seq of bulk RNA-Seq (N = 1,730) using computational deconvolution tools. First, we estimated cell type proportions using the LM22 signature matrix and CIBERSORTx (**Figure 2.2A**). We found that these proportion estimates are consistent with standard white blood cell reference ranges,[78] for which generally neutrophils have the highest abundance, lymphocytes (including T cells, B cells, natural killer (NK) cells combined) the second highest abundance, and monocytes the lowest abundance. However we note that blood cell type proportions vary across individuals depending on numerous factors such as medication use, current illness, and age.[79] We confirmed that the proportions estimated via CIBERSORTx are consistent with the complete blood count measures taken in the clinic for a subset (N=143) of individuals in our dataset (**Supplementary Figure 2.1**). We observed a pearson correlation (R^2) of 0.76 for cell type proportions estimated in neutrophils using CIBERSORTx and proportions measured in clinic, 0.85 for lymphocytes, and 0.48 for monocytes. These results suggest that the computationally estimated proportions are reliable.

Next, we used these proportion estimates and bMIND expression deconvolution (**Methods**) to estimate cell type expression. Consistent with biological expectations, we found that correlation of estimated expression between different cell types is high, as all cell types are derived from

the same tissue (**Figure 2.2B**). Next, we investigated whether computationally estimated cell type expression could successfully detect differences in expression between different cell types, despite there being a high correlation structure between different cell types. Principal component analysis confirmed that the major sources of variation in the dataset are attributable to differences in cell type expression (**Supplementary Figure 2.2**). These results suggest that using large cohorts of bulk RNA-Seq in blood, paired with computational deconvolution tools, can successfully detect differences in expression dependent on cell type composition.

Finally, we contrasted computationally-derived cell-type estimates with single cell RNA-Seq (scRNA-Seq) data.[80,81] We compared median TPM (transcripts per million) estimates across six cell types and find moderate correlation between the reference single-cell expression and computationally derived expression, ranging from R^2 of 0.11 in naive B cells to R^2 of 0.27 in CD8 T cells (**Supplementary Table 2.1** and **Supplementary Figure 2.3**). To further check how well computationally estimated expression compares to expression derived from scRNA-Seq, we correlated expression estimates between the two reference scRNA-Seq datasets in monocytes, the one cell type with data available in both reference datasets. We found that the median TPM of the 2,836 eGenes (genes with an associated eQTL) in both datasets have an R^2 of 0.22, comparable to the R^2 observed when comparing computationally estimated expression with scRNA-Seq.

Cell type eQTL analysis reveals more refined biological signal compared to bulk eQTL

Next, we performed eQTL analyses on the resulting cell type expression estimates to find evidence of genetic regulation of cell type expression. We restricted to the eight cell types with average proportion > 2% including: naive B Cells, memory B Cells, CD4 naive T Cells, CD4 memory T cells, natural killer cells, monocytes, and neutrophils. We conducted local-eQTL mapping with a 1 Mb window using QTLtools (**Methods**), to identify between 2,875 and 4,629

eQTL-genes (eGenes) with a significant association at FDR correction level of 5%, across the eight different cell types (**Figure 2.3A**). In total, we identified 5,752 eGenes with a significant association in at least one of the eight main cell types. We show that there exists a range of concordance of effect sizes for eGenes found in both the individual cell type analyses and the bulk eQTL analysis (**Figure 2.3B** and **2.3C).** This confirms findings from previous studies showing a strong shared genetic effect on gene expression across cell types. We observed that most eGenes are detected as significant in either just one, or all eight cell types (**Supplementary Figure 2.4**).

Additionally, we found evidence of cell type "opposite-effect" eQTLs, where a SNP in a given cell type shows an association with the same eGene as detected using bulk RNA-Seq, but in the opposite direction. One such example is the eQTL for *FCGR3B* (Fc fragment of IgG receptor IIIb); while the bulk eQTL had an effect size of -1.3, the effect size in neutrophils and T cell types ranged between 0.49 and 0.86. Similarly, the eQTL for *MACF1* (Microtubule actin crosslinking factor 1) had effect sizes between -1.1 and -0.15 for the T cell types, versus effect sizes ranging between 0.21 and 0.28 for the bulk and remaining immune cell types. *MACF1* is known to be involved in neurite growth during brain development and has previously been linked to schizophrenia.[82] These examples are especially interesting as it supports the idea that gene expression at the cell type level can uncover nuances of biological mechanisms that go undetected when only using bulk-level analyses. Similar effects have been observed in other studies using both single cell RNA-Seq[83] and deconvoluted bulk RNA-Seq.[84]

To further validate these cell type eQTLs, we compared the results of this analysis with results from eQTL analysis using single cell RNA-Seq from the eQTLCatalogue and BLUEPRINT consortiums.[81,85,86] We restricted to the protein coding genes identified as eGenes using the computational deconvolution approach. Generally, we found that the two approaches to cell type

eQTL mapping show strong concordance. For example, in neutrophils, we found that 2,921 out of the 4,629 genes (63%) with a significant association using the computational deconvolution approach also had a significant association in using single-cell RNA-Seq, correcting at an FDR level of 5%. Among these eGenes, comparing the association with the same leading SNP in both of these datasets (**Figure 2.3D**), we observed a correlation (R^2) of 0.66 between their effect sizes. Similar effect size correlations, for T cells CD4, B cells, and monocytes are shown in **Supplementary Figure 2.5**. This suggests that the computational deconvolution approach to large scale bulk RNA-Seq projects can be used to obtain accurate cell type eQTL estimates.

Integration of cell type specific eQTL with brain and blood trait GWAS

For every gene with a significant eQTL, we used FUSION[87] to estimate the gene expression heritability across each of the contexts, or the proportion of variance in gene expression explained by variance in genetics. Only those genes with significant heritability after five-fold cross validation per each context were retained for further analysis. **Table 2.2** provides the summarized statistics of the significantly heritable genes and the gene with highest estimated SNP-heritability per cell type. An advantage of investigating eQTLs at the cell type level is that it provides a more precise view of biological mechanisms driving the association between gene expression and phenotype. In order to investigate whether there exists variants that drive both the expression of genes in a specific cell type and a GWAS trait, we conducted Transcriptome Wide Association Study (TWAS)[87] and colocalization[65] analyses using the significant ct-eQTLs from the eight main cell types previously mentioned, and GWAS of several neuropsychiatric and blood-based phenotypes. **Figure 2.4A** provides an overview of the overlap across the contexts, both for brain-related and blood-based traits.

GWAS for neuropsychiatric traits tested include: BP,[1] SCZ,[88] major depressive disorder (MDD),[89] alcohol dependence,[90] cannabis use disorder,[91] migraines,[92] insomnia,[93] attention-

deficit/hyperactivity disorder (ADHD),[94] and Alzheimer's disease.[95] In total there were 710 eGenes found to be associated only in bulk and no other cell type, and 168 eGenes found to be associated in one or more cell types and not in the bulk (**Table 2.3**). Regarding colocalization, in total there were 68 eGenes found to have colocalized SNPs between expression and trait only in the bulk and no other cell type, and 50 eGenes found only in one or more cell types and not in the bulk (**Table 2.3**).

Of the 50 eGenes found to have a colocalization posterior probability with the same variant impacting both gene expression and the GWAS trait (PP4>0.8) in a cell type but not in the bulk, half have a higher median TPM across the GTEx v8 brain tissue types than in GTEx whole blood. This suggests that these genes are relevant for brain functions despite being detected in immune cell type specific expression estimates. An example of one such gene is *HTR6*, a serotonin receptor targeted by certain antidepressant and antipsychotic medication, found to be strongly associated and colocalized with BP in the most recent Psychiatric Genomics Consortium (PGC) study on bipolar disorder[1] which used brain-derived gene expression weights from the PsychENCODE project.[96] Conditioning on *HTR6* memory B cell-specific expression using FUSION completely removed the significant GWAS signal at this locus, suggesting that the genetic factor driving gene expression also encompasses the BP association signal (**Figure 2.4B**). The same held true for other immune cell types in which *HTR6* was colocalized with BP, including naive B cells and CD4 T cells. This demonstrates the utility of using cell type deconvolution methods in large cohorts of an easily-accessible tissue like blood, since it is able to capture gene expression regulation relevant in brain cell types that otherwise are not detectable in bulk blood eQTLs.

GWAS for blood-based traits tested include: systemic lupus erythematosus[97] (an autoimmune disorder), mean corpuscular volume, mean corpuscular hemoglobin,[98] red blood cell width

distribution, monocyte count, eosinophil count, lymphocyte count, platelet count, white blood cell count, and red blood cell count.[99] In total there were 1,765 eGenes found to have associations only in bulk and no other cell type, and 493 eGenes found only in one or more cell types and not in the bulk (**Table 2.4**). Regarding colocalization, in total there were 488 eGenes found only in the bulk and no other cell type, and 229 eGenes found only in one or more cell types and not in the bulk (**Table 2.4**).

Within the blood-based traits we again found examples of opposite-sign effects in certain cell types when compared to the bulk. For example, when considering systemic lupus erythematosus (SLE) as a trait, we found for the *IRF5* gene, natural killer cells have a TWAS Z-score of -10.7 whereas the bulk has a score of +3.91, suggesting distinct mechanisms that are dependent on the cell type context. IRF5 (interferon regulatory factor 5) is known to be implicated in SLE,[100,101] though the exact mechanism by which it is dysregulated in the context of disease remains unknown. See the TWAS Supplementary Tables to view all FUSION TWAS and colocalization results.

Lithium-dependent genetic regulation of gene expression

Given the large number of BP probands in our study sample, we were interested to see whether there were BP-specific effects that could be observed using cell type deconvoluted expression. Since lithium is the most commonly used drug to treat these patients and it has also been established that lithium use has an effect on the blood transcriptome,[102,103] we hypothesized that lithium-dependent genetic regulation of the blood transcriptome may exist. Among the 1,045 bipolar disorder patients in this cohort, 709 were taking lithium at the time of blood draw ("Lithium-User") and 336 were not ("Lithium Non-User").

When stratifying by cases versus controls (with all BP and SCZ individuals included as cases), we found significant differences in cell type proportion for CD4 T cells (p=1.8e-7, higher in controls), natural killer resting cells (p=1.2e-7, higher in controls), and neutrophils (p=2.3e-8, higher in cases). Next, considering only the cases of BP, we stratified those who use lithium versus those who do not, and found significant differences in cell type proportion for CD4 naive T cells (p=8e-4, higher in non-users), CD4 memory T cells (p=4e-4, higher in non-users), natural killer resting cells (p=3e-4, higher in non-users), and neutrophils (p=1.5e-9, higher in users). However, when we only include lithium non-users within the BP cases, and compare those against the controls, we found no significant differences in proportion for any of the cell types. See **Supplementary Figure 2.6** for example plots of all three tests using neutrophils. This suggests that the use of lithium within the BP cases drives these differences in cell type proportion, rather than disease status itself, consistent with previous findings.[103]

We validated the effect of lithium use on blood cell types in a separate cohort of individuals who had electronic health data from the University of California, Los Angeles ATLAS Community Health Initiative.[104,105] Specifically, we included self-reported European patients with a PheCode for bipolar disorder who also had laboratory test orders for complete blood counts and noted whether they had a prescription order for lithium (n=1302 with lithium, n=6208 without). In comparing the neutrophil count between BP patients who had never been prescribed lithium (or before they were prescribed lithium) and those who had a prescription order for lithium, we found that there was a significant (logistic regression p=2.09e-07) elevation of neutrophils in patients with a prescription for lithium (**Supplementary Figure 2.7**). Furthermore, for a subset of BP patients within the ATLAS dataset, we also have records for neutrophil counts both before and after the patient was prescribed lithium. Using a Wilcoxon-signed rank test with continuity correction, we found a significant difference between the neutrophil counts between the two groups (p=0.0228) when including individuals of any ancestry (n=376), though when restricting

to only European individuals (n=229), the significant difference is lost (p=0.2) (**Supplementary Figure 2.7**). The replication of this finding in this large external dataset provides further evidence to suggest that cell type proportion is impacted by lithium usage, though the implications of this are yet to be understood.

Next, we investigated whether estimated cell type expression is a significant predictor for case/control status or lithium use. Restricting to the genes with the highest variance in each cell type, we built logistic regression models to separately predict case/control status and lithium use, including the same covariates as the previous proportion-based models. However, we find that gene expression does not provide additional predictive value over the cell type proportions for either case/control status or lithium use.

To investigate lithium-dependent genetic regulation, we performed an interaction model eQTL scan between lithium users and nonusers, testing whether there exist SNPs whose cell type or cell type specific expression regulation is dependent on the presence of lithium. To do this, we included an interaction term for the genotypes and lithium status in the regression model (**Methods**). Using bulk expression, we only identified one gene with such an association (FDR p-value < 0.10). With cell type expression derived from bMIND, we identified as many as 34 such eGenes (in monocytes), and a total of 110 examples of genes (Li-eGenes) that show differential regulation of cell type expression, compared to just one gene that shows differential regulation of bulk expression (**Supplementary Table 2.3**). We found that 97 of the eGenes that have significant differential lithium regulation exhibit opposite effect sizes between the lithium user and nonuser groups, at the cell type level. The remaining 13 Li-eGenes show same direction effect sizes between the lithium user and nonuser groups, with significantly different magnitudes. For example, in naïve B cells, *KITLG* (ENSG00000049130) shows opposite effect eQTLs based on rs11104703 (**Figure 2.5A**). While in monocytes we see that *TNFRSF11A*

(ENSG00000105641) shows differential effect size, in the same direction, based on rs79143095 (**Figure 2.5B**). Due to the large number of samples used in this analysis, we are powered to detect small differences, like these.

In order to directly measure expression differences between lithium users and nonusers, we conducted a differential expression analysis test using limma[106] initially in the bulk dataset (**Methods**). Comparing the two groups, we tested 17,194 genes from bulk expression measures. We found 100 genes with evidence of differential expression in bulk (FDR < 0.05), with log fold changes of the significant genes ranging from -0.191 to 0.177, suggesting low impact of lithium on differential expression (**Figure 2.5C**). Out of the 100 differentially expressed genes found here, 33 were previously reported in Krebs, et. al,[103] a significant overlap according to Fisher's exact test (OR = 6.43, p = 4.74e-14). Overlapping genes include *FBXL2* - a gene highly expressed in the brain and involved in neuronal signaling, and *CNTNAP3* - which mediates interactions between neurons and glial cells. See Differential Expression Supplementary Tables for full lithium differential expression results.

Though previous studies have not found substantial evidence of differential expression in the blood transcriptome between cases of BP or SCZ and controls,[103,107] we were interested in investigating this within our own cohort given the uniquely large sample size. Using the bulk RNA-seq and the same 17,194 genes selected in the lithium-user differential expression analysis, we found 64 genes with FDR < 0.05, of which nine genes overlapped with the significant genes found in the lithium analysis. Log fold changes of the significant genes ranged only from -0.126 to 0.104, suggesting that if these genes are truly a result of disease status, the differences are minimal (**Supplementary Figure 2.8**). See Differential Expression Supplementary Tables file for full case/control differential expression results.

For the cell-type specific differential expression analyses, we leveraged the differential expression function available through the bMIND software. In the case-control analysis, we found four differentially expressed genes in Neutrophils (FDR<0.05), including *TSPAN2* and *CFAP45* both of which were reported in the Krebs et. al. lithium differential expression study.[103] We found 24 differentially expressed genes in memory B cells, and 21 in naive B cells (with 18 differentially expressed genes in common between the two B cell types). Interestingly, when conducting the lithium user versus non-user analysis, we did not find any differentially expressed genes in any cell type. While this may be a result of the smaller sample set used in the lithium analysis as compared to the case-control analysis, it also may reflect that the effects of lithium are only found at the bulk level due to its impact on cell type composition, rather than changes in gene expression within individual cell types. See Differential Expression Supplementary Tables file for q-values of all cell type specific differential expression results.

DISCUSSION

We show that cell type deconvolution of bulk blood RNA-seq provides novel insights not only for immune-relevant biology, but also neuropsychiatric disease biology. While bulk eQTLs tend to provide a greater number of associations overall, we find that cell type specific eQTLs provide unique associations not otherwise detectable in bulk. Many of these unique cell type associations have high expression in brain tissue types, and harbor several example genes that have been previously implicated in BP TWAS[1] studies using brain tissue. This demonstrates that large cohorts of an easily accessible tissue like blood is useful for deciphering biology for brain-related phenotypes when cell type deconvolution is applied. An important caveat, however, is that the associations with brain-related traits found in this study are most likely to be shared genetic mechanisms between blood cell types and brain cell types, rather than blood cell type-specific biology.

Considering the BP TWAS results alone, there were 82 total eGenes with an opposite direction of effect in a cell type than in the bulk eQTL analysis (defined as having an opposite-sign TWAS Z-score for the same gene and the same trait). For example, we found 63 eGenes, significantly associated with BP, that have an opposite direction of effect in CD8-T cells when compared to bulk expression. *ARID5A*, a gene implicated in the most recent PGC bipolar disorder TWAS[1] is one example of these genes. In the bulk expression the TWAS Z-score of *ARID5A* and bipolar disorder is -4.99 (TWAS Z-score -5.32 in PGC BP study), whereas in CD8-T cells it is +6.02. This gene was also found to be colocalized with PP4>0.8 in the CD8 T Cell test, though it does not pass the colocalization threshold in the bulk test or PGC3 BP test. The same is true for *ARID5A* in CD4 memory resting T cells (TWAS Z-score +6.56). Similarly, the methyltransferase gene *WDR82* in CD4 Naive T cells has a positive (TWAS Z-score +3.72) association with BP, whereas the bulk expression has a negative (TWAS Z-score -3.98) association at the same locus (TWAS Z-score -6.75 in PGC BP study). There are many such examples of these genes across each of the cell types and the various traits that we examined.

Examples of novel BP-associated genes were also discovered, including *RILPL2*, found to be colocalized in the context of memory B cells, monocytes, natural killer resting cells, and CD8 T cells, but not in the bulk. This gene is highly expressed in whole blood in adults (median TPM 27.42 in GTEx), but is also crucial for dendritic-spine morphogenesis in developing neurons[108] Similarly, *CAMKK2* (calcium/calmodulin dependent protein kinase kinase 2), a gene found to be colocalized in the context of monocytes, neutrophils, and CD4 T cells is highly expressed both in whole blood and in brain tissues (particularly cerebellar hemisphere and cerebellum according to GTEx). While *CAMKK2* has not been implicated in a BP TWAS, the large PGC GWAS points toward calcium channel signaling as a potential therapeutic target for BP,[1] and indeed a loss-of-function mutation in this gene has been previously linked to BP status.[109] We

consider these to be potential BP-relevant genes that are interesting candidates for experimental validation.

We replicated previous findings that immune cell type composition is impacted by lithium use rather than BP status. We also replicated several previously reported genes that are differentially expressed in whole blood in response to lithium, in addition to reporting novel lithium-response genes. Although lithium has been prescribed as a mood stabilizer for decades, its precise mechanism of action is still unclear.[110] Lithium has been shown to increase the activity of the transcription factor CREB (cAMP response element-binding protein),[111] a protein involved in neuronal plasticity.[112] Here, we found that *ATF4*, an eGene in all cell types and the bulk, which encodes for CREB-2, has opposite directions of effect in T cell types than in the other immune cell types or bulk. We found a similar pattern for the *AKT1* (Rho-family-alpha serine/threonine-protein kinase) eGene. AKT1 protein levels in brain tissue have been previously associated with both schizophrenia and bipolar disorder, and although genetic associations exist,[113] they do not pass genome-wide multiple testing correction.

While we find promising lines of evidence that immune cell type specific expression is useful for discovering candidate brain-relevant genes, there are several limitations to our study. Firstly, while our cohort had an ample number of BP patients, the number of SCZ samples was much lower, and thus underpowered for a diagnosis-specific analysis. Furthermore, we only test SNP-gene pairs in *cis*, whereas *trans* eQTLs are known to be more context-specific,[17] so we miss distal associations that are potentially biologically relevant to the phenotypes of interest. By using computationally-derived expression estimates, there is a greater possibility for spurious associations that are not related to biology, dependent on the specific method of decomposition/deconvolution chosen. Also by using low-coverage RNA-seq, we may be missing important eGenes that are not as highly expressed in blood. Finally, our study consists of all

European-ancestry individuals, but to gain a more comprehensive and inclusive understanding of the biology between immune cell types and psychiatric conditions, in addition to better fine-mapping these eQTL, many more samples of diverse ancestries need to be analyzed in future work.

Collectively, this suggests that while the bulk whole blood gene expression provides a greater number of significant findings overall, cell type specific expression allows us to observe additional biological mechanisms that are not possible to capture when only using gene expression measures from bulk alone.

METHODS

Cohort description

The samples included are from a study with individuals ascertained for bipolar disorder (BP) or schizophrenia (SCZ). The cohort consists of 1,045 individuals with BP, 84 individuals with SCZ, and 601 controls with whole blood RNA-seq and corresponding genotypes (N=1,730 after excluding first degree relatives) included for all individuals.

Bulk RNA-Sequencing

Bulk RNA-sequencing was performed at the UCLA Neurogenomics Core, using the TruSeq Stranded plus rRNA and GlobinZero library preparation method, as described previously.[73] We used FASTQC to visually inspect the read quality from the lower-coverage whole blood RNA-Seq (5.9M reads/sample). We then used kallisto[55] to pseudoalign reads to the GRCh37 gencode transcriptome (v33) and quantify estimates for transcript expression. We aggregated transcript counts to obtain gene level read counts using scripts from the GTEx consortium (https://github.com/broadinstitute/gtex-pipeline).

Genotyping pipeline

Genotypes for the individuals included in the cohort were obtained from the following platforms: OmniExpressExome (N = 816), Psych Chip (N = 522), COEX (N = 162), Illumina550 (N=19), and Global Screening Array (N=211). Given that the SNP-genotype data came from numerous studies, the number of overlapping SNPs across all platforms was < 80k, prompting us to perform imputation separately for each genotyping platform, as previously described in Schwarz, et. al. 2022. Briefly, genotypes were first filtered for Hardy-Weinberg equilibrium p value < 1.0e-6 for controls and p value < 1.0e-10 for cases, with minor allele frequency (MAF) > 0.01, then were imputed using the 1000 Genomes Project phase 3 reference panel[61] by chromosome using RICOPILI v.1[62] separately per genotyping platform, then subsequently merged. Imputation quality was assessed by filtering variants where genotype probability > 0.8 and INFO score > 0.1. We restricted it to only autosomal chromosomes due to sex chromosome dosage, as commonly done.[114]

Cell type proportion estimation

We estimated the proportion of cell types of the bulk whole blood RNA-seq datasets using CIBERSORTx, with batch correction applied and LM22 signature matrix as the reference gene expression profile. The LM22 signature matrix uses 547 genes to distinguish between 22 human hematopoietic cell phenotypes, though here we restrict to 8 cell types with proportions > 0.02.

Complete blood counts (CBC) lab tests from the clinic were provided for a subset of the cohort (N=143), providing us ground truth measures (in units of 10^9 cells per liter) for neutrophils, lymphocytes, monocytes, basophils, and eosinophils. To make the counts comparable to the proportions outputted by CIBERSORTx, we divided the counts of the cell type of interest by the sum of counts across all cell types in an individual, providing the count ratio shown in Supplementary Figure 2.1.

Cell type expression estimation

We log2-transformed the matrix of bulk TPM measures before inputting into bMIND since the largest expression measure was greater than 50 TPM. Using the cell type proportions derived from CIBERSORTx in conjunction with these log-transformed bulk expression measures, we used bMIND in order to derive cell type expression estimates, with flag np=TRUE.

bMIND derived estimates and cis-eQTL mapping

Using output from bMIND, we transformed expression estimates from log2(TPM) to counts using sequencing library sizes, restricting to sufficiently expressed genes (estimated count > 1.0 in 40% of individuals). Expression estimates were then standardized (mean = 0) then performed cis-eQTL analysis mapping using QTLTools, using a defined window of 1 Mb both up and downstream of every gene's TSS, for sufficiently expressed genes (TPM > 0.1 in 20% of individuals). We run the eQTL analysis in permutation pass mode (1000 permutations, and perform multiple testing corrections using the q value FDR procedure, correcting at 5% unless otherwise specified. We then restrict associations to the top (or leading) SNP per eGene.

TWAS and colocalization

We used the FUSION pipeline to perform TWAS on the normalized cell type specific expression estimates and normalized bulk expression measures, residualizing each expression matrix by its first 50 principal components to account for variation due to technical (non-biological) factors. Imputed genotypes were restricted to those that overlap with the 1000 Genomes LD reference panel, providing 272,652 SNPs on which to perform the analysis. A window of 500kb upstream and 500kb downstream of the lead SNP for each eQTL was used as the cis-region to be tested. Gene-trait pairs were selected based on the best performing model after five-fold cross validation, including for Best Unbiased Linear Predictor (BLUP), elastic net (ENET), Least Absolute Shrinkage and Selection Operator (LASSO), and just using the top SNP.

We tested for colocalization of GWAS and eQTLs using the –coloc flag within the FUSION/TWAS pipeline. Colocalization is only performed in those gene-trait associations with $p < 0.05$. In each cell type, we tested eGenes with a significant association between expression and SNP (**Tables 2.4 and 2.5**). We report SNPs with a colocalization probability (PP4) > 0.80 .

Cell type specific regressions using estimated cell type proportions and gene expression

We built logistic regression models to evaluate the effect of cell type proportion on case/control status, and lithium use status within only the BP cases. These models included the proportion of one cell type at a time, along with covariates including age, sex, RNA concentration, and RNA integrity number (RIN) as predictors. In testing the differences in cell type proportions between different binary outcomes, we used the glm() function in R with family=binomial.

We also used logistic regression to predict either case control status or lithium use (only in BP cases) from cell type expression estimates after residualizing for 50 expression PCs. Variable numbers of genes were included based on genes with most variance per cell type, using a range of 100 to 1000 genes with an interval of 100. Covariates include age, sex, RNA RIN, RNA concentration, and cell type proportion estimates. A random 70% of individuals were sampled to use for training, and 30% for testing the prediction.

Electronic Medical Record Validation Cohort

ATLAS is an opt-in biobank that enrolls patients when they visit UCLA for a blood draw. ATLAS is a diverse biobank that includes patients from a variety of genetic ancestries that live across the greater Los Angeles region.[115] Registered ATLAS researchers can access deidentified electronic health record data for patients, consisting of outpatient and inpatient encounters, including information on diagnoses, procedure orders, laboratory orders, and prescription orders. As of 2022, there were approximately 50,000 participants enrolled in ATLAS. A

complete description of the ATLAS project and data is available in [104].

Bipolar patients were identified in ATLAS using the diagnosis table. The bipolar phenotype was defined as any patient who had at least one diagnosis of any of the ICD 10 codes included in the bipolar Phecode Map 1.2.[116] Neutrophil counts (measured as 10^3 counts/μL) were determined using test results for complete blood count laboratory orders. We restricted this analysis to those individuals with self-reported European ancestry. To prevent severe outliers from biasing results, test results with a neutrophil count greater than 2 standard deviations from the median count value in all bipolar patients were removed. Lithium prescription orders were found by querying the prescription order table for medications of any dose or format that were classified as psychiatric medication and had the generic name lithium.

Neutrophil count data for patients with a bipolar Phecode were separated into three categories: tests administered before the patient was prescribed lithium, tests administered after the first lithium prescription order, and tests for patients without a lithium prescription order. Since many patients had multiple complete blood count orders, the median neutrophil count per patient per category was calculated. Median neutrophil counts were compared between bipolar patients after their first lithium prescription and bipolar patients without a lithium prescription using a logistic regression (implemented in R). Max age and sex were used as covariates. For the subset of patients who had complete blood count tests taken before and after a lithium prescription order, we used a paired Wilcoxon rank test to increase power, implemented in R using the wilcox.test(paired=TRUE) command.

Interaction model

To test whether there exists an interaction between SNP-lithium usage, we included an interaction component in the regression model, as such: $y = \beta * X + \beta * l + \beta * (X + l) + covariates$

where *X* refers to the genotype at a particular SNP, and *l* refers to lithium use.

Differential expression analysis

We used the limma eBayes function with trend=true to conduct differential expression tests in the bulk dataset. We include only those genes with at least 1 TPM in at least 436 individuals (about 25% of the total 1,730 individuals included in the analysis), leaving 17,194 genes to be tested. We then log2-transform this matrix and compute the first 50 expression principal components to be included as covariates. In the lithium user vs non-user analysis, only cases were included to avoid confounding effects caused by disease status, while in the case-control analysis, all individuals diagnosed with BP or SCZ were included as cases and non-affected individuals included as controls.

For the cell-type specific differential expression analysis, we use the bmind_de() function as included in the bMIND software package. To keep the methods comparable to the bulk analysis, we also use the log2-transformed expression measures as input along with the first 50 expression PCs as covariates.

Data Availability

The lower-coverage RNA-seq and the corresponding genotypes generated and analyzed during this study have been deposited in dbGAP (accession number phs002856.v1).

Chapter 2 Tables

Table 2.1: Cell type proportion estimates from CIBERSORTx and number of eQTLs per cell type.

Cell type	Mean cell type proportion estimate (s.d.)	Number of eGenes (FDR < 0.05)
Naive B cells	0.025 (0.020)	4,009
Memory B cells	0.020 (0.014)	3,571
CD8 T cells	0.025 (0.025)	2,875
Naive CD4 T cells	0.15 (0.042)	3,082
Memory Resting CD4 T cells	0.066 (0.034)	3,284
Resting NK Cells	0.066 (0.029)	3,858
Monocytes	0.050 (0.039)	3,483
Neutrophils	0.51 (0.094)	4,629
Bulk (directly from RNA-Seq)	1.0	7,302

Table 2.2: FUSION heritability results. Number of Sig. Genes refers to the number of genes that remain significantly (P<0.05) heritable after five-fold cross validation. Q1 = first interquartile, Q3 = third interquartile. Overall, the bulk data shows higher heritability estimates across each of the statistics. Of note is that every gene listed is distinct for each context, including genes that are relevant to neuronal function, such as *NSG1* (neuronal vesicle trafficking associated), *CAMKK2* (calcium dependent kinase, involved in neuronal differentiation and synapse

formation) and *BTG1* (B-cell translocation gene 1, found to be involved in neural stem cell renewal).[117]

Panel	Number of Sig. Genes	Min	Q1	Median	Mean	Q3	Max	Gene with Max h2
Bulk	5,113	0.0041	0.026	0.055	0.096	0.12	0.728	*TRBV28*
B Cells Memory	2,541	0.0035	0.024	0.044	0.075	0.093	0.68	*BTG1*
B Cells Naive	1,552	0.0056	0.024	0.045	0.078	0.095	0.579	*PI16*
Monocytes	2,431	0.0052	0.025	0.045	0.077	0.095	0.584	*NSG1*
NK Cells Resting	2,763	0.0042	0.024	0.045	0.078	0.098	0.61	*BCAT1*
Neutrophils	1,605	0.0056	0.025	0.048	0.083	0.10	0.69	*CAMKK2*
T Cells CD4 Memory Resting	1,989	0.0057	0.026	0.047	0.080	0.099	0.63	*SBF2*
T Cells CD8	2,033	0.0057	0.024	0.042	0.069	0.081	0.63	*FGFBP2*
T Cells CD4 Naive	2,147	0.0053	0.024	0.044	0.075	0.092	0.56	*CROT*

Table 2.3: TWAS & Colocalization neuropsychiatric trait results. Shared refers to the number of significant (FDR < 0.05) genes that are in common with the bulk TWAS-significant gene set, whereas unique refers to those that are not present in the bulk TWAS-significant gene set.

Cell Type	Significant eGenes	Num Significant TWAS Genes, Shared	Num Significant TWAS Genes, Unique	Num Genes with Coloc PP4>0.8, Shared	Num Genes with Coloc PP4>0.8, Unique
B Cells Naïve	4,009	90	43	43	13
B Cells Memory	3,571	142	58	62	25
T Cells CD8	2,875	108	50	50	15
T Cells CD4 Naïve	3,082	120	46	56	19
T Cells CD4 Memory Resting	3,082	115	43	55	22
NK Cells Resting	3,858	156	72	73	21
Monocytes	3,483	126	52	62	24
Neutrophils	4,629	76	35	35	9
Bulk	7,302	906	/	155	/

Table 2.4: TWAS & Colocalization blood-based trait results. Shared refers to the number of significant (FDR < 0.05) genes that are in common with the bulk TWAS-significant gene set, whereas unique refers to those that are not present in the bulk TWAS-significant gene set.

Cell Type	Significant eGenes	Num Significant TWAS Genes, Shared	Num Significant TWAS Genes, Unique	Num Genes with Coloc PP4>0.8, Shared	Num Genes with Coloc PP4>0.8, Unique
B Cells Naïve	4,009	922	164	289	78
B Cells Memory	3,571	1,582	207	511	106
T Cells CD8	2,875	1,276	168	414	93
T Cells CD4 Naïve	3,082	1,349	183	445	88
T Cells CD4 Memory Resting	3,082	1,257	150	419	80
NK Cells Resting	3,858	1,712	254	557	119
Monocytes	3,483	1,484	212	484	113
Neutrophils	4,629	969	159	331	60
Bulk	7,302	3,893	/	1,175	/

Chapter 2 Figures

Figure 2.1: Graphical overview of pipeline. Figure created in BioRender.

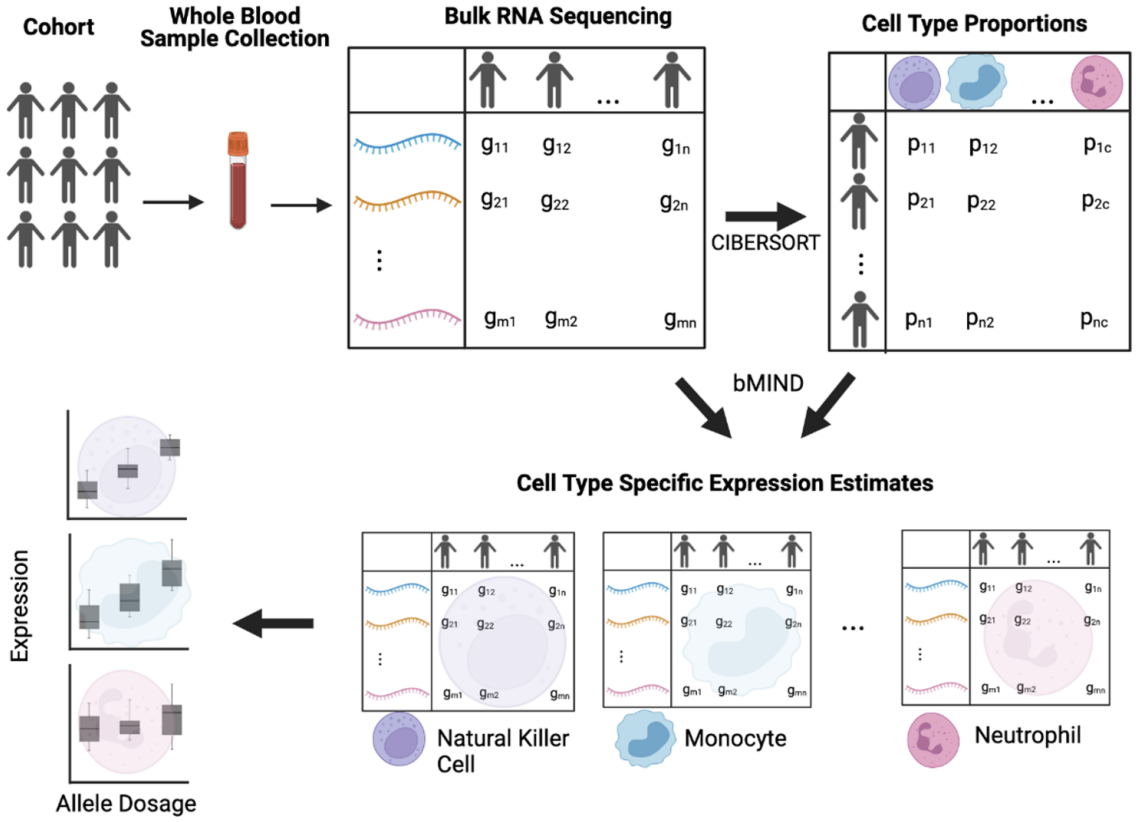


Figure 2.2: Cell type expression from computational deconvolution methods

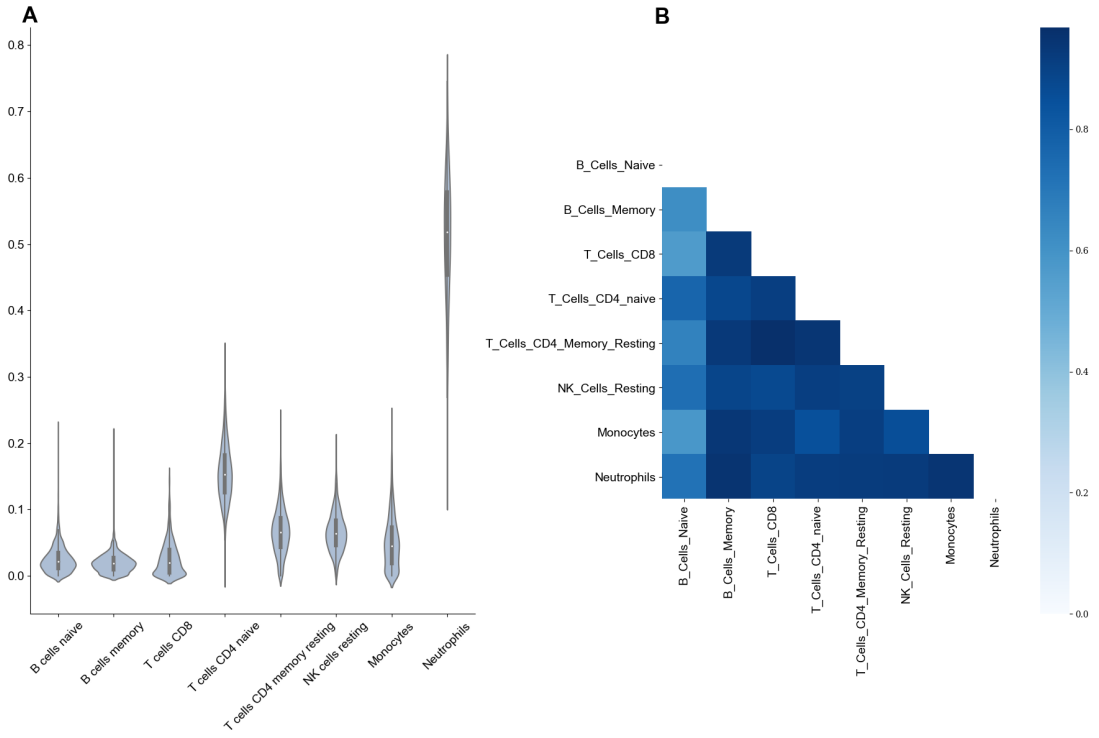


Figure 2.2A: Cell type proportion predictions from CIBERSORTx - A violin plot showing the range of estimated cell type proportions for all 1730 individuals in each of the eight major cell types.

Figure 2.2B: R^2 of expression between each cell type - A heatmap of correlations (measured by R^2 of mean expression across samples) between the eight main cell types captured by CIBERSORTx.

Figure 2.3: eQTLs per cell type, effect size correlation with reference dataset and bulk dataset.

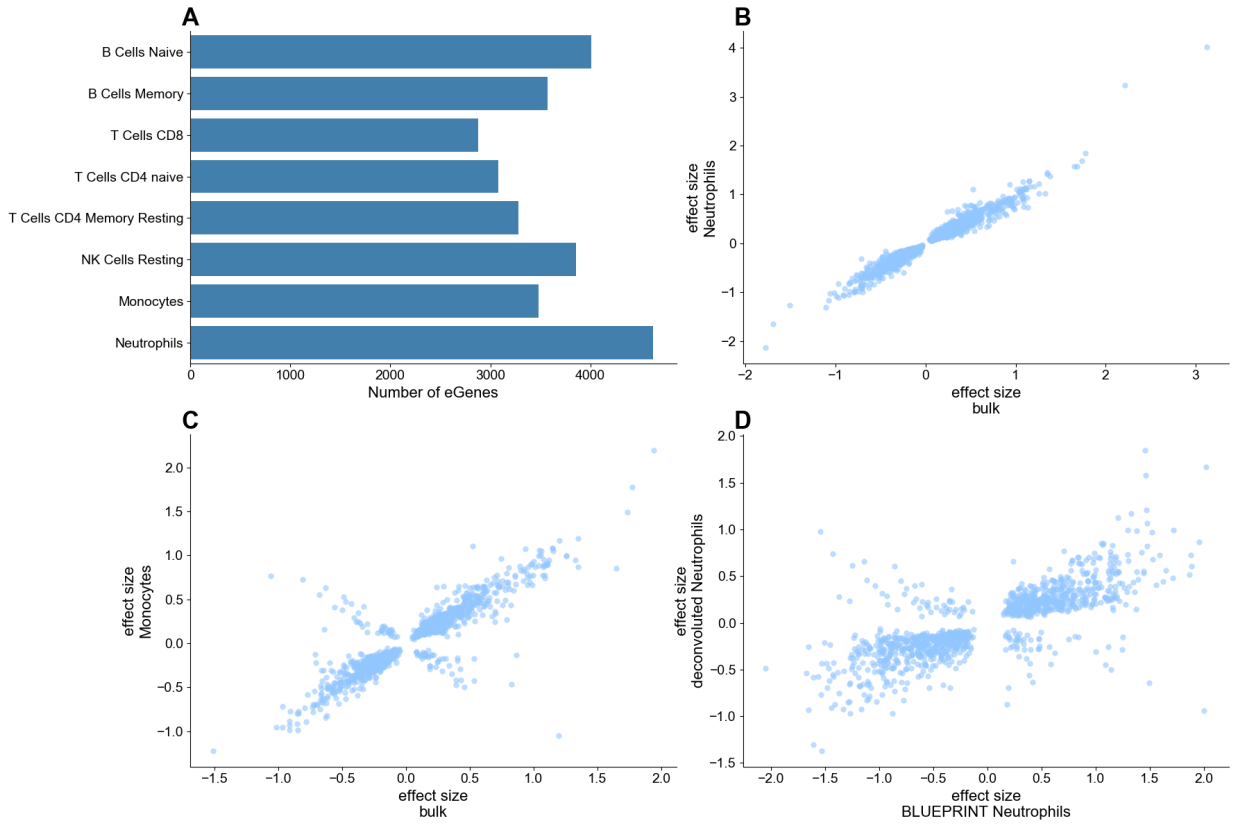


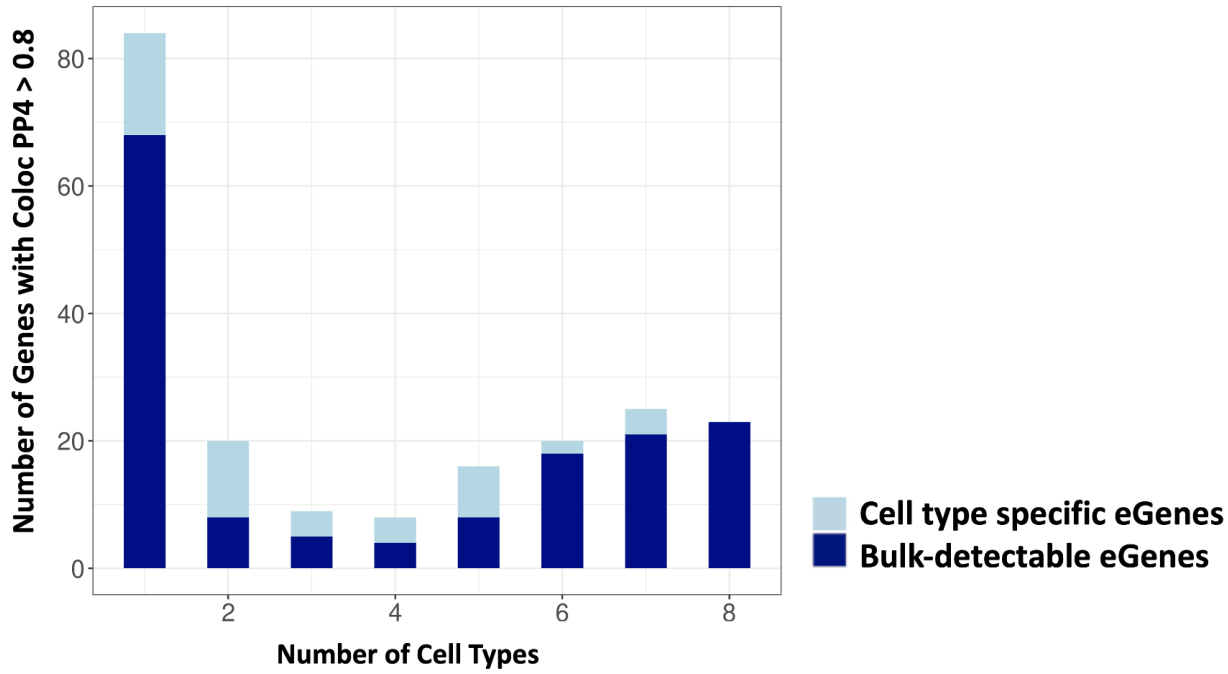
Figure 2.3A: Number of associations identified per cell type - Number of eGenes with a significant association identified for the eight major cell types detected by CIBERSORTx, using a FDR cutoff of 0.05 .

Figure 2.3B: Comparison of effect size between shared cis-associations with Neutrophils - Restricting to the eGenes with a significant association in both the bulk eQTL analysis and neutrophil eQTL analysis, we compare the estimated effect sizes of the most significant eQTL associations.

Figure 2.3C: Comparison of effect size between shared cis-associations with Monocytes - Restricting to the eGenes with a significant association in both the bulk eQTL analysis and monocyte eQTL analysis, we compare the estimated effect sizes of the most significant eQTL associations.

Figure 2.3D: Comparison of effect sizes between shared cis-associations using reference single cell RNA-seq. Restricting to the eGenes with a significant association in both the BLUEPRINT reference neutrophil eQTL analysis and our neutrophil eQTL analysis, we compare the estimated effect sizes of the most significant eQTL associations.

Figure 2.4: Colocalization and enrichment analyses of cell type specific eQTLs.



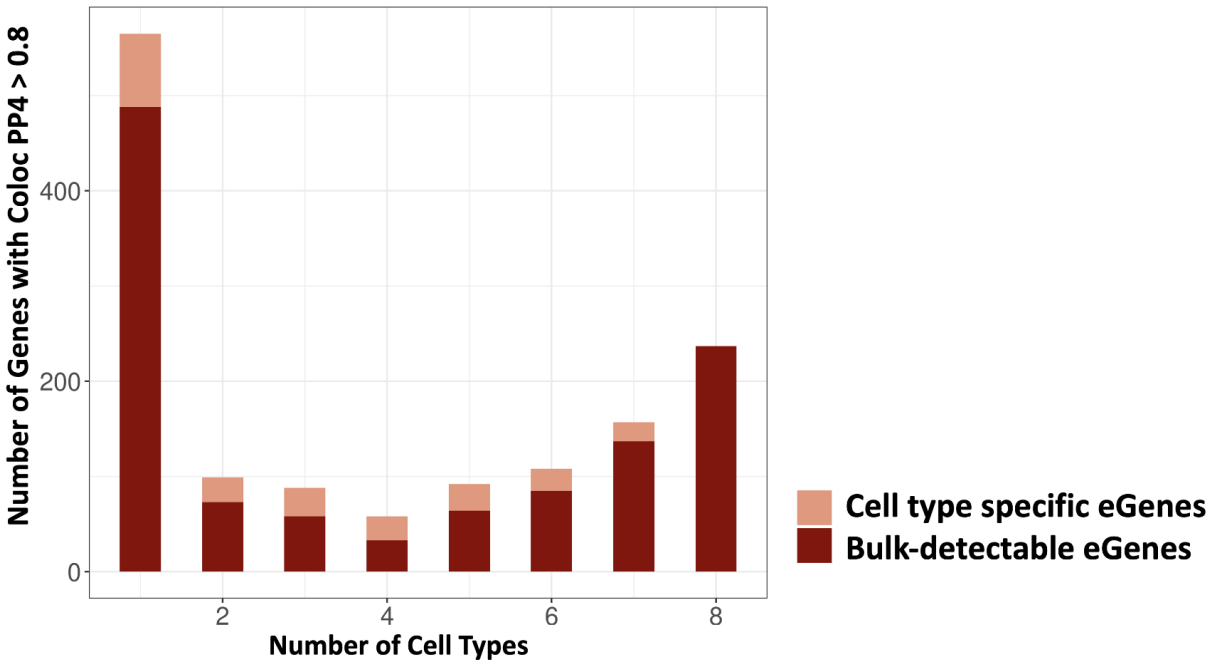


Figure 2.4A: (top) Number of genes with coloc PP4>0.8 across contexts in neuropsychiatric traits. (bottom) Number of genes with coloc PP4>0.8 across contexts in blood-based traits.

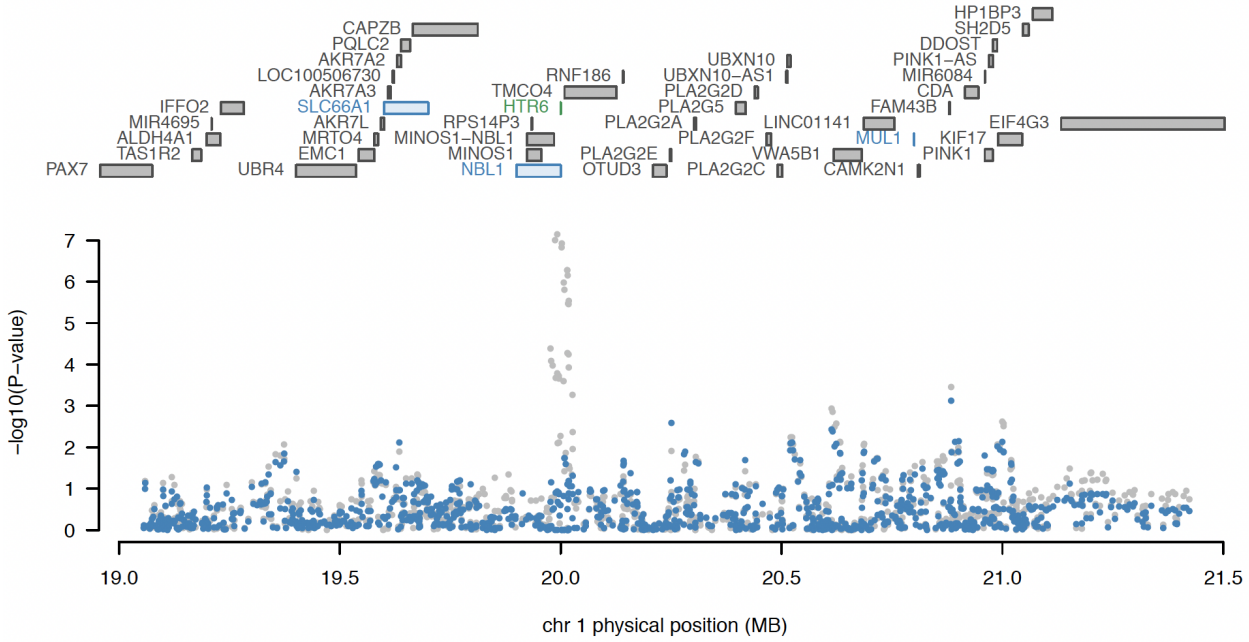


Figure 2.4B: Conditional analysis of *HTR6* expression in memory B cells.

All genes in the locus are included in the top panel, with marginally TWAS associated genes highlighted in blue, and those jointly significant (*HTR6*) in green. The bottom panel includes a Manhattan plot of the GWAS data before (gray) and after (blue) conditioning on the imputed expression of *HTR6* in memory B cells. Figure generated by FUSION.post_process.R script.

Figure 2.5: Lithium user vs non-user analyses.

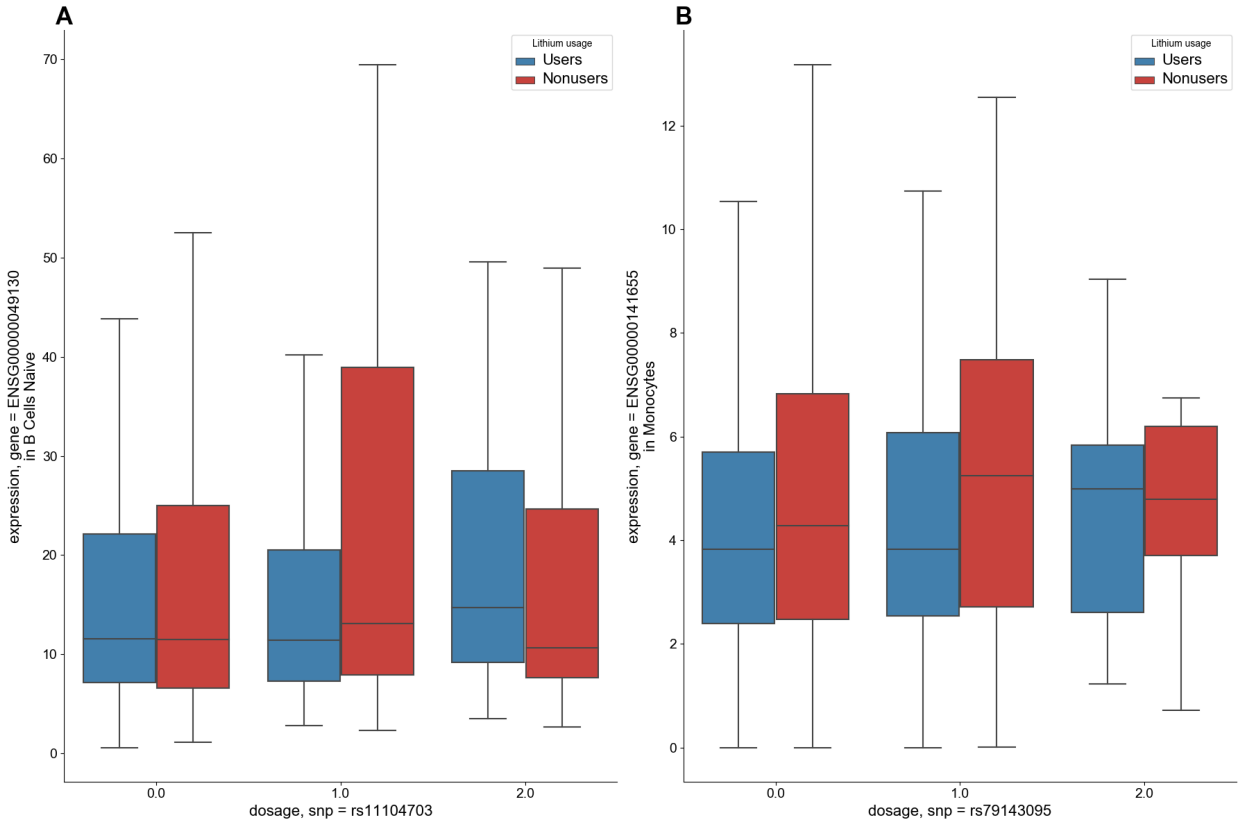


Figure 2.5A: Boxplots showing the expression of *KITLG* (ENSG00000049130) in naïve B cells, stratified by dosage of SNP rs11104703 in lithium users versus nonusers.

2.5B: Boxplots showing the expression of *TNFRSF11A* (ENSG00000105641) in monocytes, stratified by dosage of SNP rs79143095 in lithium users versus nonusers.

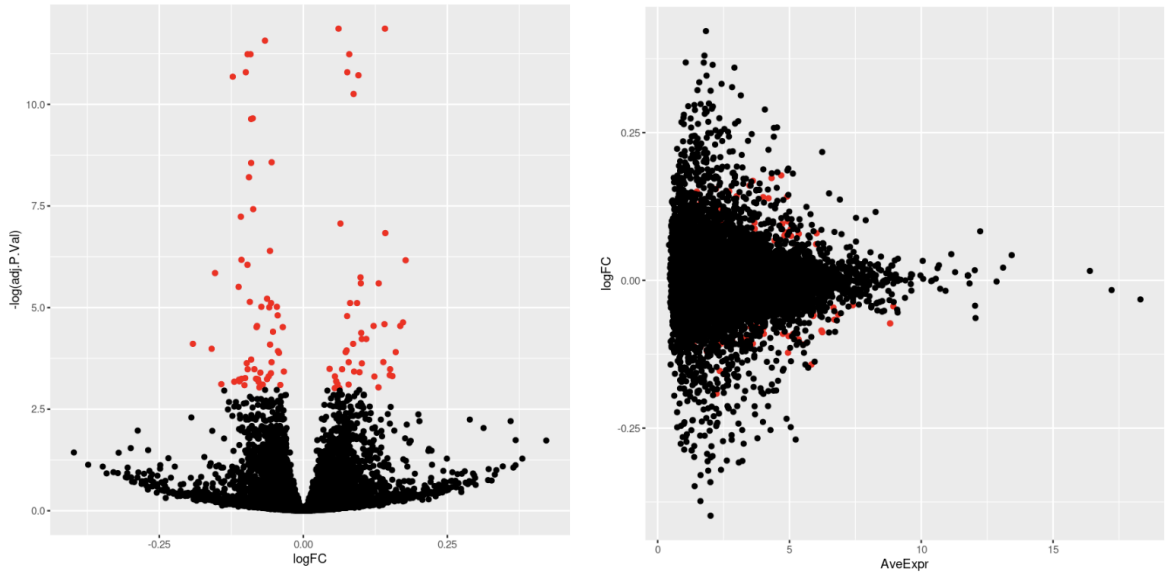
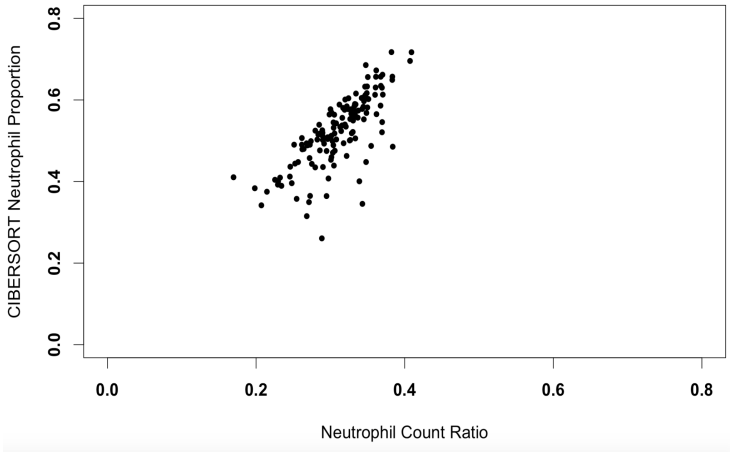
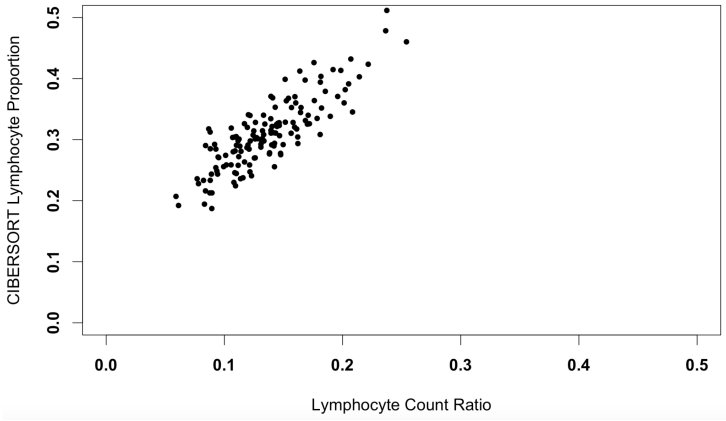


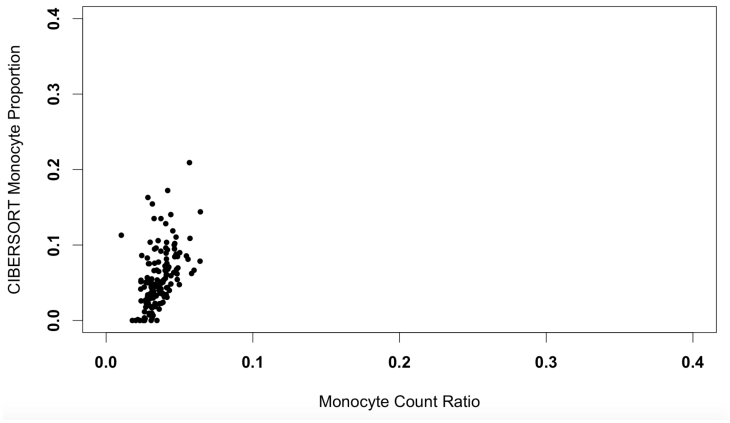
Figure 2.5C: Differential gene expression results for lithium users vs lithium non-users: (left) Volcano plot which highlights differentially expressed genes (FDR < 0.05) in red (N=100 total differentially expressed genes). (right) Average expression of each gene vs the log fold change (logFC) of each gene, with differentially expressed genes highlighted in red.

Supplementary Figures

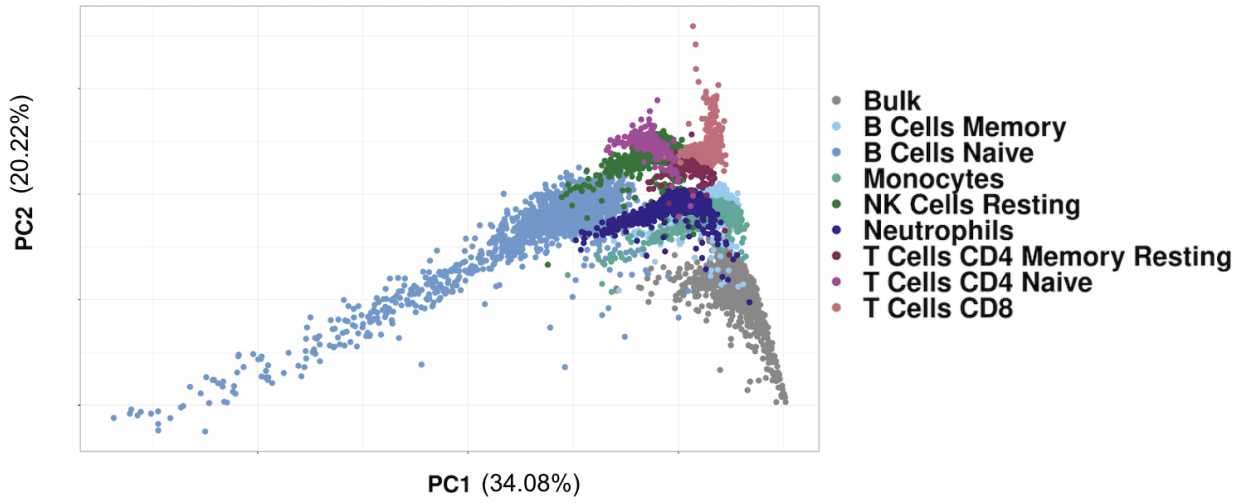
Supplementary Figure 2.1: Scatterplots of CIBERSORTx-estimated cell type proportions vs complete blood count proportions. We find generally high concordance between computationally estimated and measured ground truth cell type proportions using a subset of our cohort. Pearson's correlation R2 for neutrophils = 0.76, for lymphocytes = 0.85, for monocytes = 0.48.


CIBERSORT Neutrophil Proportion
0.0
0.2
0.4
0.6
0.8
Neutrophil Count Ratio


CIBERSORT Lymphocyte Proportion
0.0
0.1
0.2
0.3
0.4
0.5
Lymphocyte Count Ratio

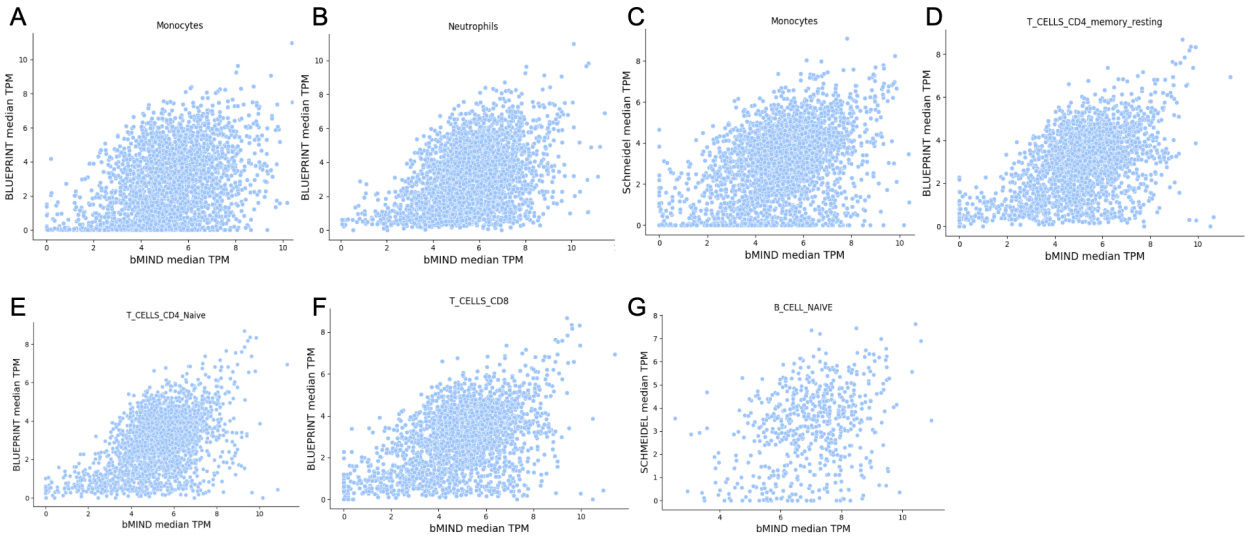

CIBERSORT Monocyte Proportion
0.0
0.1
0.2
0.3
0.4
Monocyte Count Ratio

Supplementary Figure 2.2: PCA of cell type expression.

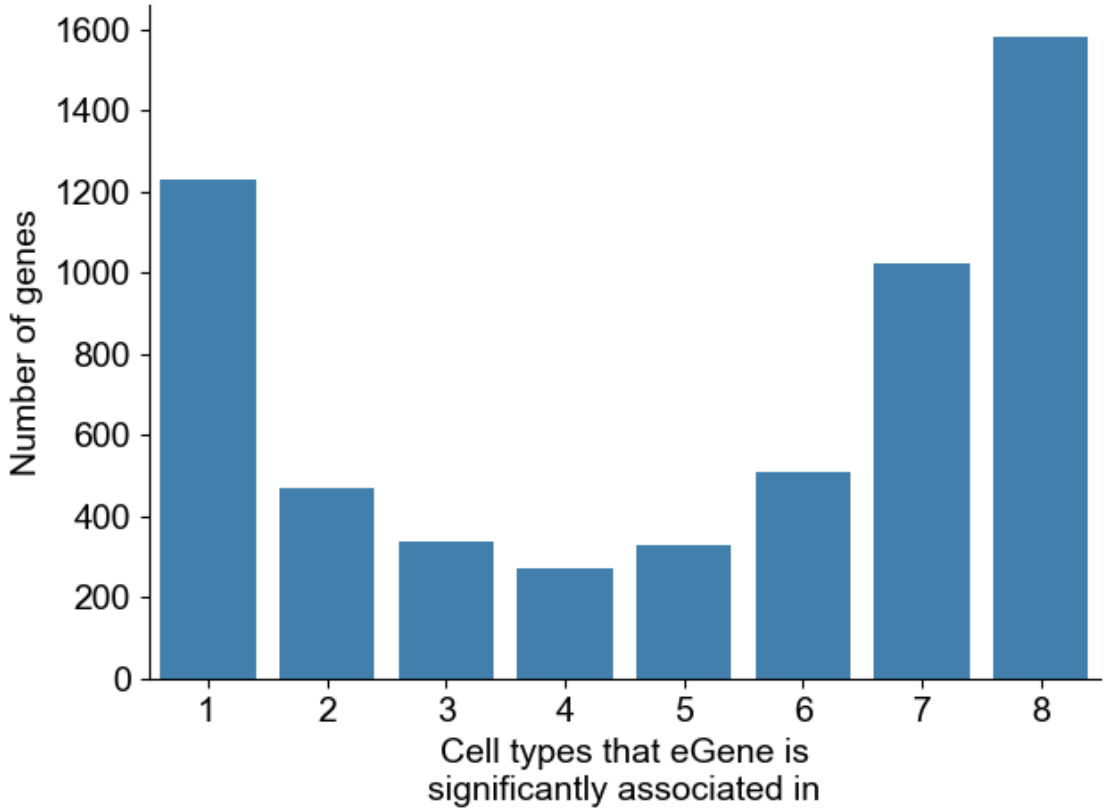


Supplementary Figure 2.3: Scatterplots of expression estimated from bulk vs single-cell

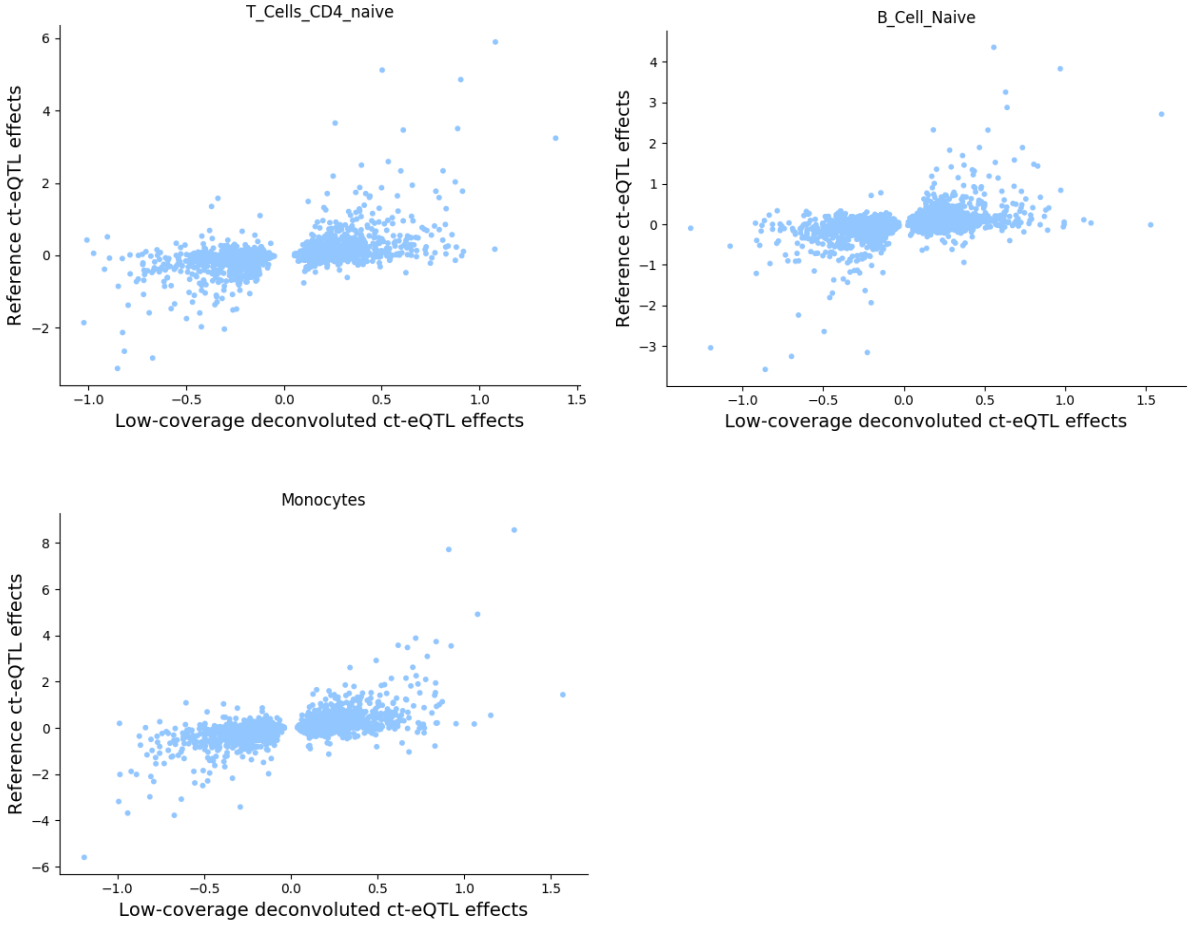
Using two scRNA-Seq datasets as references, we compare the median TPM values for genes detected as eQTLs using both scRNA-Seq and computationally deconvoluted bulk RNA-Seq.



Supplementary Figure 2.4: Distribution of shared eGenes across cell type contexts.



Supplementary Figure 2.5: Effect size correlations between reference single cell eQTL and the deconvoluted eQTL. T cells CD4 naive R2 = 0.27; B cells naive R2 = 0.22; Monocytes R2 = 0.36.



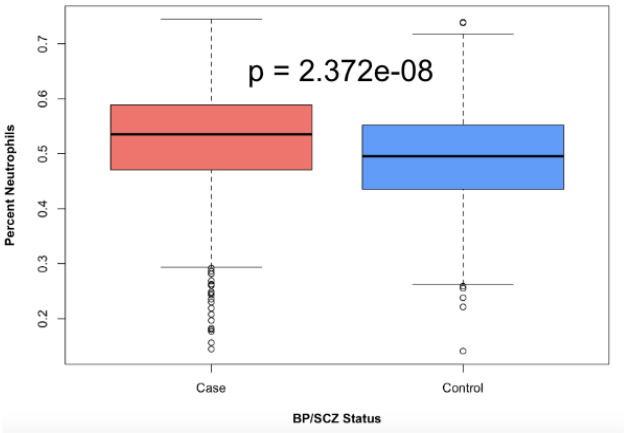
Supplementary Figure 2.6: Neutrophil count elevated for lithium users.

A. Difference in neutrophil proportion (after accounting for covariates including age, sex, RIN, and RNA concentration) between BP/SCZ cases and controls.

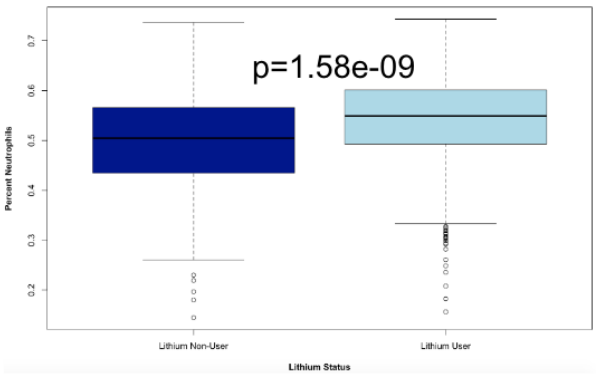
B. Difference in neutrophil proportion between lithium users and non-users (after accounting for covariates), only within BP cases.

C. Difference in neutrophil proportion between lithium non-users and controls (after accounting for covariates).

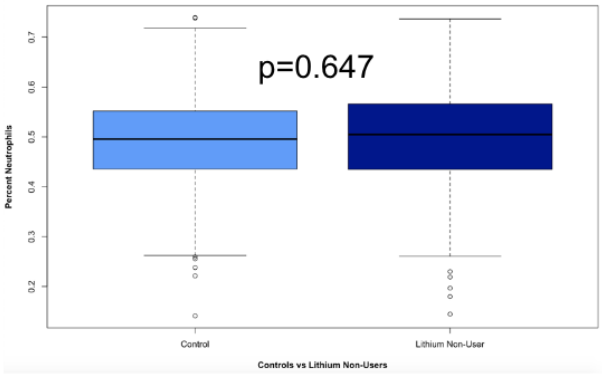
A.



B.



C.



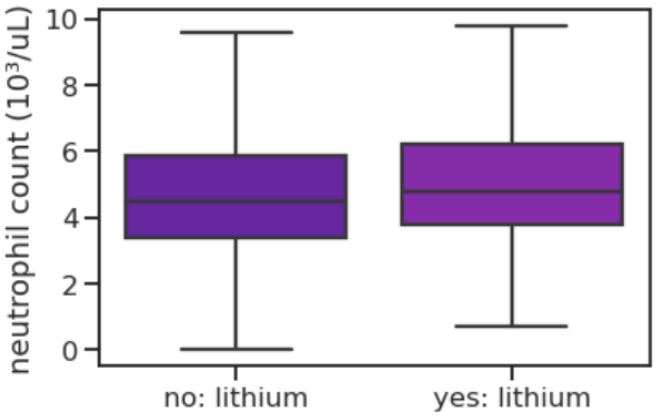
Supplementary Figure 2.7: Neutrophil count elevated for lithium users in UCLA ATLAS.

A. Median neutrophil count across self-reported European patients, with covariate correction for age and sex (p=2.09e-7).

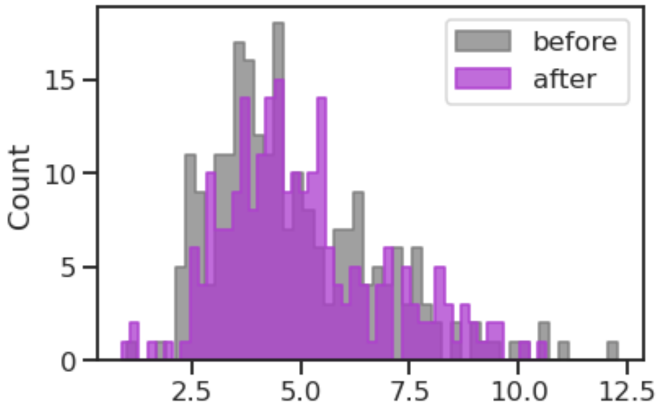
B. Neutrophil count distribution across self-reported European patients before and after lithium prescription.

C. Neutrophil count distribution across patients (n=382, all ancestries) before and after lithium prescription.

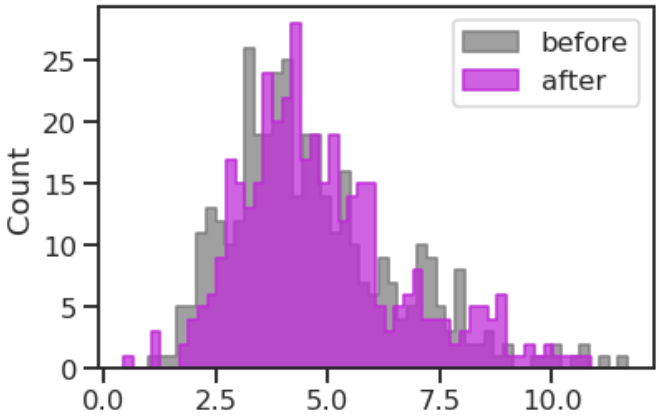
A.



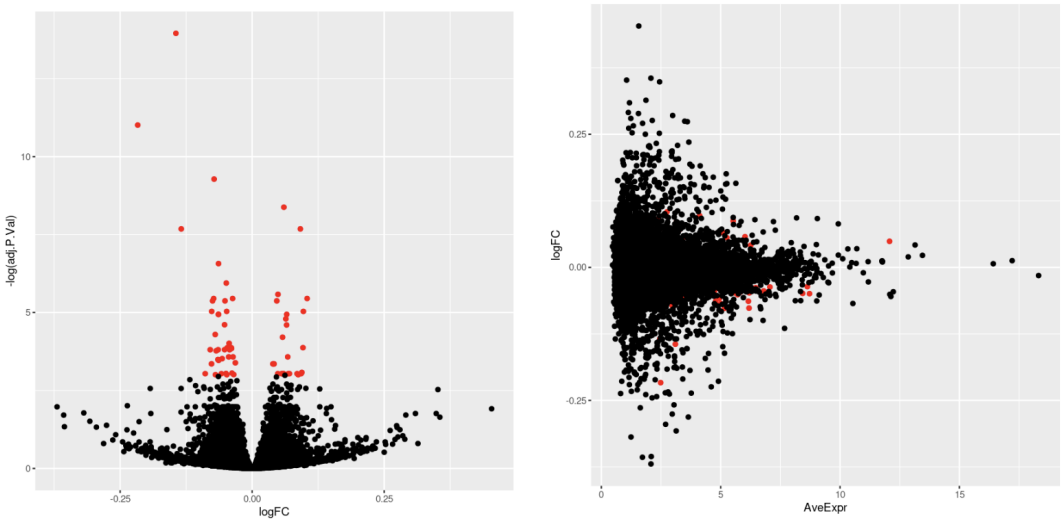
B.



C.



Supplementary Figure 2.8: **Differential gene expression results for BP or SCZ cases vs controls:** (left) Volcano plot which highlights differentially expressed genes (FDR < 0.05) in red (N=64 total differentially expressed genes). (right) Average expression of each gene vs the log fold change (logFC) of each gene, with differentially expressed genes highlighted in red.



Supplementary Table 2.1: Correlations of median expression between reference single cell RNA-Seq datasets and computationally derived expression estimates

Restricting to the genes identified as eGenes using both the single cell RNA-Seq reference dataset and the computationally derived cell type expression, we report R^2 values for the median TPM for genes across samples.

Cell type	Reference	R^2	Number of genes
Monocytes	BLUEPRINT	0.14	2896
Neutrophils	BLUEPRINT	0.16	3239
CD4 Memory T Cells	BLUEPRINT	0.26	2504
CD4 Naive T Cells	BLUEPRINT	0.27	2504
CD8 T Cells	BLUEPRINT	0.24	2504
B Cell Naive	Schmeidel	0.11	624
Monocytes	Schmeidel	0.15	2896
Monocytes*	Schmeidel/BLUEPRINT	0.22	2836

Supplementary Table 2.2: Linear models for lithium usage prediction

Cell type	Model R^2	Effect size	SE	p-value
Monocytes	0.06	-1.28	0.36	<0.001
Neutrophils	0.09	1.08	0.15	<0.001

CD4 Memory T Cells	0.06	-1.68	0.42	<0.001
CD4 Naive T Cells	0.06	-1.52	0.36	<0.001
CD8 T Cells	0.06	-2.09	0.55	<0.001
B Cell Naive	0.06	-3.06	0.79	<0.001
B Cell Memory	0.05	0.59	1.09	0.59
Resting NK Cells	0.06	-1.92	0.52	<0.001

Supplementary Table 2.3: Number of eGenes differentially regulated by Lithium. Using an FDR cut-off of $p < 0.10$, we look at the number of eGenes with a significant SNP-lithium interaction. “Same-direction” Li-eGenes have the same direction of effect sizes between lithium users and nonusers, and “opposite-direction” Li-eGenes have the opposite direction of effect sizes between lithium users and nonusers.

Cell type	Number of Li-eGenes	Number of “same-direction” Li-eGenes	Number of “opposite-direction” Li-eGenes
Naive B cells	24	4	20
Memory B cells	15	3	12
CD8 T cells	2	1	1

Naive CD4 T cells	25	1	24
Memory CD4 T cells	2	0	2
Resting NK cells	5	0	5
Monocytes	34	3	31
Neutrophils	3	1	2

Chapter 3: Quantitative trait loci mapping of circulating metabolites in cerebrospinal fluid to uncover biological mechanisms involved in brain-related phenotypes

Authors: Lianne M. Reus*[1,2,3], Toni Boltz*[1], Marcelo Francia[4], Merel Bot[4], Naren Ramesh[4], Maria Koromina[5], Wiesje F. van der Flier[2,3], Pieter Jelle Visser[2,3,6,7], Sven van der Lee[2,3,8], Betty M. Tijms[2,3], Charlotte E. Teunissen[9], Loes Olde Loohuis[4], Roel A. Ophoff[1,4]

* shared first-author

1. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

2. Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, The Netherlands

3. Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands

4. Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA

5. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York; Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York

6. Department of Psychiatry, Maastricht University, Maastricht, The Netherlands

7. Department of Neurobiology, Care Sciences and Society, Division of Neurogeriatrics, Karolinska Institutet, Stockholm, Sweden

8. Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, the Netherlands

9. Neurochemistry Lab and Biobank, Department of Clinical Chemistry, Amsterdam Neuroscience, Amsterdam UMC, Amsterdam, The Netherlands

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified many genetic risk loci contributing to human diseases and traits. The functional interpretation of these risk loci is crucial for treatment development and biomarker identification, but has been challenging in practice. Quantitative trait loci (QTL) studies provide valuable mechanistic insights for various disease etiologies.[118] The study of metabolomics in particular allows the detection of small changes in endogenous and exogenous compounds, thereby closely reflecting the current physiological state of cells, tissue or organisms.[119–123]

Due to the relative inaccessibility of *in vivo* brain tissue and its surroundings, insights into biological processes underlying psychiatric and neurodegenerative disorders have been limited.[14,124] Cerebrospinal fluid (CSF) participates in waste removal of neural metabolism through interaction with interstitial fluid surrounding brain cells.[18] CSF can be collected *in vivo,* and thus serves as a relevant source for products of ongoing biological mechanisms in the brain. CSF studies on neurodegenerative disease have been successful in identifying biological processes involved, such as abnormal regulation of lipid metabolism and increased inflammation.[125–127] So far, CSF QTL mapping studies have shown that the proteome[124,128,129] and metabolome[130–132] are (at least partly) under genetic control, which implies that insight in its genetic architecture can aid in the biological interpretation of genetic risk loci.

In this study, we performed a genome-wide metabolite-QTL (mQTL) study on the largest mass-spectrometry (MS)-based CSF metabolomic panel studied so far (5,543 metabolites) using data from 977 individuals of European ancestry. The assayed metabolites include those of primary

metabolism, biogenic amines, and complex lipids. We identified 82 significant mQTLs for 65 CSF metabolites and 51 independent loci, of which 58/65 (89.2%) have not been detected before in CSF and eight not in blood, saliva or urine. When integrating our CSF mQTLs with pre-existing summary statistics data on ten psychiatric and neurodegenerative disorders, we identified 23 CSF metabolites associated with brain-related traits. Many of the CSF metabolites with a mQTL (56.9%, 37/65) colocalized with brain-specific eQTLs. This study indicates that the multi-omic integration of genetics with CSF metabolomic data helps to identify which molecular mechanisms underlie neurobiological mechanisms.

RESULTS

QTL mapping of CSF metabolites

CSF samples were assessed for metabolic compounds via three platforms that each measure different components of the metabolome: GC-TOF MS/MS (primary metabolism, 393 metabolites, 273 of which are unannotated), CSH-QTOF MS/MS (complex lipids, 3,532 metabolites, 3,262 unannotated), and HILIC-QTOF MS/MS (biogenic amines, 1,618 metabolites, 1,194 unannotated) (Table S3.1). Given a high degree of concordance between observed mQTL effects in both the cognitively healthy subjects cohort as the memory clinic cohort (Table S3.2), we performed a meta-analysis in the combined dataset of 977 subjects to boost power to detect associations (Figure 3.1). We did not observe genomic inflation for any of the CSF metabolites (range λ=0.97-1.03) (Table S3.1). Heritability estimates for CSF metabolite levels ranged from h^2_{SNP}=3%-49% (mean = 15%, SD = 12%), with the highest reported heritability for N-Acetylhistidine (49%), N-epsilon-Dimethyl-L-lysine (42%) and ethylmalonic acid (23%).

Detection of independent and novel CSF mQTL signals

Our genome-wide mQTL study on 5,543 metabolites and 6,189,630 SNPs identified 4,999 significant SNP-metabolite pairs after Bonferroni-correction (P<6e10^{-11}). After applying a filter on independent LD blocks (based on clumping SNPs within 1Mb and LD R^2<0.2) we observed 126 significant SNP-metabolite pairs, representing 65 unique CSF metabolite levels across 73 loci (Figure 3.2A). Following stepwise conditional analyses on the subset of CSF metabolites (n=19) with multiple independent loci (n=80), we identified in total 82 mQTLs (for 51 independent risk loci and 65 metabolites): 53 metabolites had one independent locus, eight metabolites had two loci, and four metabolites had three (or more) independent loci (Table 3.1, Table S3.3, Figure 3.2C). This is in line with previous studies, demonstrating that most of the heritability of metabolite levels is confined to a single locus,[130] often uniquely regulated by the gene that encodes a protein involved in the specific metabolic pathway. CSF metabolites under a more complex regulatory architecture included ribonic acid (*ENOSF1/TYMS*), ethylmalonic acid (*UNC119B*), and two unannotated metabolites (i.e., *PYROXD2* for 9.31_161.13, *TYMS* for 9.40_165.04). Similarly, many independent mQTLs (n=33) harbored associations for more than one metabolite, involving a total of 39 different metabolites with at most five metabolites at one locus (Figure 3.2D) (i.e., *FADS2*). This suggests pleiotropy at these loci, potentially involving a key enzyme in multiple metabolic pathways, or a group of metabolites that are all part of a single pathway.

Most mQTL associations were detected using the HILIC-QTOF MS/MS platform (84.1%, 69/82 SNP-metabolite associations), followed by GC-TOF (9.8%, 8/82 SNP-metabolite associations) and CSH-QTOF, (6.1%, 5/82 SNP-metabolite associations).

Most CSF mQTLs located in intronic regions

To examine the functional annotation of our identified CSF mQTLs, we queried the Gencode V41[133] basic gene annotation database for the 51 unique, independent SNPs significantly associated with these metabolites (Figure 3.2B). We found that variants within intronic regions

of protein coding genes account for 60.7% of loci, variants within promoters (defined as 10kb upstream of transcription start site) of genes account for 1.9% of independent loci, variants within untranslated regions (UTRs) for 3.9%, and variants within exons for 3.9%. The paucity of exonic variants is a known trend among QTL and GWAS studies,[134,131,135] whereas the abundance of intronic variants suggests potential role of enhancer activity[136,137] in the regulation of these metabolites. The remaining variants (29.4%) were considered intergenic and thus have no currently known functional annotation.

QTL mapping of CSF metabolites identifies novel and validates known biological pathways

Of the mQTL-associated CSF metabolites with annotation that we identified, thirteen (13/20, 65%) have been previously detected in QTL studies on CSF[130] (n=5), blood[119–121] (n=10), saliva[122] (n=3) or urine[123] (n=3) metabolite levels (Figure S3.1, matched by metabolite then for SNPs in LD R2>0.8).

Eight annotated CSF mQTLs were novel (not previously reported in CSF, blood, saliva or urine), including those for hypaphorine, bicalutamide, 3'-O-Methylcytidine, PEP-38:3 or PEO-38:4, inosine, ribose-5-phosphate, 3-Aminotyrosine and proline (Table S3.4). One novel mQTL consisted of SNPs within the *SLC22A5* locus and CSF levels of exogenous metabolite hypaphorine (chr5:132389258:T:C, *P*=1.52e-12, no causal SNPs observed with fine-mapping) (Figure S10-S11). Another novel mQTL locus included the *SACM1L* locus on chromosome 3 and associated with CSF proline levels (rs73058498, $P=1.88e10^{-12}$). Statistical and functional fine-mapping with SuSiE selected the lead SNP rs73058498 (posterior inclusion probabilities (PIPs) range 0.25 - 0.40) as part of the only 95% credible set for proline (see Table S3.9 for fine-mapping SNPs within 95% credible sets for which their summed PIPs exceed 50%).

The complex lipids platform harbored one mQTL locus, representing both previously identified and novel *pleiotropic* effects of *FADS2* (on chromosome 11) with CSF levels of four different phosphatidylcholines, PC 36:3 Isomer B (i.e., two isomers with different m/z ratios), PC 38:3 (two isomers) and and glycerophospholipid PE P-38:3/PE O-38:4 (novel association; rs174556, range P=8.52e^{-15}-2.11e^{-19}). Statistical and functional finemapping with FINEMAP highlighted the lead SNP rs174556 as part of a 50% credible set (along with eight other SNPs) for three out of four of the lipids, and at least two methods selected rs174556 as part of 95% credible sets (along with six other SNPs) for two of those three lipids (Table S3.9).

The strongest association was reported for *PYROXD2* on chromosome 10 with CSF levels of five inter-correlated metabolites (Pearson's R2 > 0.83): N(epsilon)-Dimethyl-L-lysine and unannotated CSF metabolites 7.25_144.10, 9.29_260.20, 9.31_161.13, 9.30_130.09 (top SNP rs10883083 and chr10:98392298, range P=5.03e-181-3.58e-214 for all CSF metabolites) (Figure S3.2 - S3.6). Fine-mapping prioritized six potentially causal SNPs, including top SNP rs10883083 (as well as rs10786415, rs2147896, rs4539242, rs59667296, rs942814) in a 95% credible set for N(epsilon)-Dimethyl-L-lysine across all four methods. The top SNP rs10883083 was also part of the 95% credible sets for the other unannotated CSF metabolites at this locus (Table S3.9). The *PYROXD2* locus has been shown to associate with CSF levels of a similar (but not identical) metabolite called N6-methyl-L-lysine, suggesting that this locus plays a role in the regulation of lysine metabolism.[130]

Another association was observed for *NAT16* (N-acetyltransferase 16) on chromosome 7 with CSF levels of three inter-correlated CSF metabolites: N-Acetylhistidine, 8.10_220.07 and 7.40_268.13 (rs740104, P= 8.33e-74, P=4.47e-61, P=4.68e-25, respectively) (Figure S3.7-S3.9). Fine-mapping identified two 95% credible sets for N-Acetylhistidine and 8.10_220.07, one of which (size = 4) included the top SNP rs740104, the other included only two SNPs,

rs12540617 (PIPs range 0.72-0.85) and rs2227653 (PIPs range 0.14 - 0.28). For 7.40_268.13 we found only one 95% credible set (size = 4), including the top SNP (Table S3.9). N-Acetylhistidine is expressed in the human brain and other organs, but its specific function is unknown. This association has been reported before in blood plasma.[138]

Another previously detected mQTL locus (in human plasma[139], saliva[122] and urine[123]) included four independent SNPs on *ENOSF1* (on chromosome 18) that associated with CSF levels of ribonic acid (rs11081229, P=2.36e10^{-12}, rs2790, P=1.49e10^{-27}, rs2847334, P=1.09e^{-27}, rs6506537, P=1.27e10^{-20}). Fine-mapping revealed two 95% credible sets for the ribonic acid mQTL, though only the SuSiE method prioritized the top SNP rs2847334 as part of a 95% credible set. Both SuSiE and PolyFun-SuSiE highlighted rs11081266 (PIPs range 0.19 - 0.41) and rs2790 (PIPs range 0.25 - 0.37) as the SNPs with the highest PIPs in a 95% credible set (Table S3.9).

CSF metabolite levels associate with risk loci for brain-related disorders

To identify metabolites with CSF levels associated with brain-related disorders, we performed metabolome-wide association (MWAS) analyses using the framework provided by FUSION.[35] Out of the 220 CSF metabolites that had sufficient predictability (i.e., at least one SNP with P<5e-8), the R^2 between the predicted and actual metabolite levels ranged from 0.005 to 0.67 (mean = 0.05, SD = 0.09). The HILIC-QTOF platform had the best performance and included more CSF metabolites as compared to the other platforms (Figure S3.12-3.13, Table S3.5).

CSF levels of 220 metabolites were tested for association with ten brain-related phenotypes, including Alzheimer's disease[3], dementia with Lewy bodies[140], stroke[141], amyotrophic lateral sclerosis[142], bipolar disorder[1], schizophrenia[2], major depressive disorder[143], attention deficit hyperactivity disorder (ADHD)[94], insomnia[144], and alcohol abuse disorder.[145] We identified 40

significant (FDR<0.05) metabolite-phenotype associations, including ten traits and 31 CSF metabolite levels (Figure 3.3, Figure S3.14, Table S3.6). The strongest MWAS associations (with annotation) included: bipolar disorder with CSF metabolite levels of PC 36:3 Isomer B, PC 38:3, PE P-38:3/P EO-38:4 (*FADS2* on chromosome 11, range *P*-FDR=5.9e-08-4.6e-09), schizophrenia with CSF metabolite levels of 3-Aminotyrosine (*EMX1* on chromosome 2, *P*-FDR=4.25e-05). Three metabolite-trait associations had supporting evidence from colocalization analysis (posterior probability (PP)4>0.8), including bipolar disorder-PC36:3 Isomer B (*FADS2*), bipolar disorder-aspartic acid (intergenic region on chromosome 18 most nearby gene *CDH2*, rs2847408) and schizophrenia-2.06_142.09 (*SETD7*). Of these metabolome-wide significant genes with supporting colocalization evidence, the association of the intergenic region nearby *CDH2* for bipolar disorder was novel, showing no genome-wide evidence for association in the corresponding GWAS (minimal *p* within ±1 Mb of the gene's region=2.06e-05)[1].

Identified CSF mQTLs colocalize with known brain eQTLs

To identify whether any CSF mQTL colocalizes with eQTL for brain-specific gene expression, we used the PsychENCODE TWAS weights for 13,490 genes in a FUSION association test, followed by colocalization analysis. Of the 65 metabolites with a significant mQTL, we found that 37 (56.9%) have a high probability of colocalization (PP4>0.8) with brain-specific eQTL across 28 genes (Table S3.7).

We found that the mQTL on chromosome 14 regulating CSF levels of guanosine and inosine, both purine nucleosides, is highly likely to be colocalized (PP4>0.98) with the locus regulating expression levels of *PNP,* or purine nucleoside phosphorylase, in the brain. Within our dataset, guanosine and inosine levels are significantly correlated with an R2 of 0.65. Panyard, et. al previously reported the guanosine association at the *PNP* locus using CSF[130] metabolomics, as well as an association with schizophrenia.[130]

Furthermore, we found a group of complex lipids that have high colocalization probability with a few genes at a locus on chromosome 11. Three of these lipids (two annotated as PC 36:3 Isomer B and PC 38:3) had colocalization PP4>0.9 with the *FADS1* (fatty acid desaturase 1) gene locus (Figure 4). These three lipids and an additional one (another PC 38:3 isomer) were colocalized (PP4>0.8) with the *FEN1* (flap endonuclease 1) gene locus. Similarly, these four lipids plus an additional one (annotated as "PE P-38:3 or PE O-38:4") were colocalized (PP4>0.9) with the *TMEM258* (transmembrane protein 258) gene locus. Each of these lipids have a Pearson's R2 > 0.83 with each other, suggesting that these lipids and their isomers are metabolites of a particular pathway. Sall, et. al.[146] has previously grouped these genes together into a syntenic block of conserved MDD risk genes involved in myelination and lipid metabolism. The levels of these metabolites also showed highly significant associations with bipolar disorder in our MWAS.

Additionally, we applied the recently developed method isoTWAS[147] in order to identify specific metabolite-associated isoforms that may otherwise be diluted in the gene-level TWAS approach. After adjusting p-values for multiple testing and applying a PIP threshold of 0.8, we find that 59 metabolite QTL (90.7%) are significantly associated across 96 different gene isoforms (61 unique genes) (Table S3.8). Of the 139 gene-metabolite pairs from isoTWAS, seven of these are direct replications of the TWAS findings (a significant overlap, Fisher's exact test p-value = 3.6e-10), including the three of the lipids associated with the *FADS1* gene locus (as well as four unnamed HILIC metabolites associated with the genes *CTNNA1*, *PKD2L1*, *PYROXD2*, *SRSF12*). Many metabolites that were TWAS-significant were also isoTWAS-significant for different genes, including inosine and guanosine associated with isoforms of *EIF2B2* rather than *PNP*. EIF2B is known to have guanine nucleotide exchange factor activity,[148] suggesting that isoTWAS is useful in identifying additional potential mechanisms of genetic regulation between metabolites and genes.

DISCUSSION

We performed genome-wide QTL mapping on 5,543 circulating CSF metabolite levels in a cohort of 977 living individuals. We identified 82 significant mQTLs representing 51 independent loci and 65 unique CSF metabolite levels, of which 58/65 (89.2%) have not been detected before in CSF and eight not in blood, saliva or urine.

We found both novel and replicated mQTL associations, of which most metabolites seemed to be regulated by the gene locus that encodes the protein for which they are substrates or products. One novel example found here is hypaphorine, a plant metabolite that can be found in the human metabolome after ingestion of legumes (e.g., lentils, chickpeas).[149] Hypaphorine was found to be regulated by the *SLC22A5* locus, a gene for the OCTN2 protein, which transports carnitine into the cell. Hypaphorine has been shown to have inhibitory effects on a SLC22 family members due to its similarity in structural backbone to carnitine.[150] A strongly significant and replicated example is the association between ribonic acid and the *ENOSF1* locus, the protein for which catalyzes the dehydration of sugars, including ribonic acid.

Similarly, we observed that 56.9% (37/65) of the CSF metabolites with a mQTL colocalized with brain-specific eQTLs, and 90.7% (59/65) colocalized with brain specific isoform-QTLs (isoQTLs). Collectively, these eQTL-CSF mQTL colocalizations provide insights into the shared neurobiological mechanisms and possible interactions between these genes and CSF metabolite levels. One such eQTL-mQTL colocalization we found is at the *PNP* locus with CSF levels of guanosine and inosine. PNP is involved in the purine nucleotide salvage pathway by catalyzing the conversion of guanosine to guanine as well as inosine to hypoxanthine.

When integrating our identified CSF mQTLs with pre-existing genome-wide summary statistics on ten psychiatric and neurodegenerative disorders, we identified 31 CSF metabolites of which predicted levels associated with brain-related traits. The strongest MWAS associations were with bipolar disorder and schizophrenia, of which an intergenic region nearby *CDH2* for bipolar disorder was novel. We also observed metabolite associations for Alzheimer's disease, stroke, and ADHD.

Several studies have previously reported the lipid mQTL hotspot at the *FADS1/2* locus,[151,130,131] which we augment with additional lipids here. This locus has also been identified as a GWAS risk region for bipolar disorder,[1,152] and here we found this locus to be significantly associated with phosphatidylcholine lipid levels, glycerophospholipid levels, and bipolar disorder through our MWAS analysis. Colocalization with brain eQTLs and isoQTLs suggested a common genetic mechanism regulating the levels of these lipids and the levels of the *FADS1* gene. Heterozygous knockouts of the *FADS1/2* genes in mice have shown episodic phenotypes such as increases in hyperactivity (marked by increased wheel-running), depression-like episodes, and changes in circadian rhythms, suggestive of the symptoms commonly associated with bipolar disorder.[153] Similarly, lipid dysregulation is known to play a role in AD.[154] While we did not find significant differences in the levels of these lipids in our ADC cohort (data not shown), genetic variants in the *FADS1* locus have also been implicated in increased risk for Alzheimer's disease,[155] suggesting further pleiotropy at this locus for this neurodegenerative phenotype as well. This genetic region has been described previously as both highly polygenic and pleiotropic, mostly influencing complex lipids synthesized from arachidonic acid, a product of the rate-limiting enzymes encoded by *FADS2* (and *FADS1*).[151]

This study successfully identified many mQTL associations in CSF that were replicated in two separately ascertained cohorts; nonetheless, several limitations should be taken into account.

First, some metabolites had no annotations, or annotations with low reliability scores (based on the Metabolomics Standards Initiative (MSI) scale), which makes these associations difficult to interpret biologically. As an example, we observed a novel mQTL association for variants located on *POU4F2* and CSF levels of a metabolite annotated as a prostate cancer medicine, bicalutamide (MSI level=3, i.e., moderate reliability). However, we did not observe sex differences for this association despite bicalutamide being prescribed mainly in men for treating metastatic prostate cancer (bicalutamide inhibits testosterone by binding to androgen receptors). The POU4F2 protein (encoded by *POU4F2*) expression levels have been implicated in promoting tumor growth in various cancer types through the Hedgehog signaling pathway.[156] Altogether this suggests that our newly identified mQTL could represent a true biological pathway for prostate cancer, but this could not be verified with the current annotation information on this metabolite. Furthermore, our study consisted solely of European-ancestry individuals. To gain a better understanding of the biology between CSF metabolites and brain related disorders and gene expression, as well as to improve the fine-mapping accuracy of these QTL, many more samples of diverse ancestries must be included in future studies.

Since CSF collection can occur *in vivo* in healthy and diseased individuals, our results highlight that the specific integration of genetics with CSF metabolomic data could help understanding how genetic factors contribute to ongoing molecular mechanisms underlying neuropsychiatric disorders and neurobiological mechanisms. This study indicates that the multi-omic integration of genetics with CSF metabolomic data helps to identify which molecular mechanisms underlie diseases and disorders of the brain. Large-scale genome-wide studies on the CSF metabolome are limited, thus expanding on these studies is imperative for gaining biological insight in psychiatric and neurodegenerative disorders. As the summary statistics generated by this study can also be useful for studying other diseases and traits, data will be shared with the scientific community via the European Bioinformatics Institute GWAS catalog.

METHODS

Study sample(s)

In total, 977 study samples (age 52.7±16.6 years, 35.9% female) were included from a memory clinic cohort and a cohort of cognitively healthy subjects in the Netherlands.

Memory clinic cohort: Memory clinic samples originated from three cohorts, that are all related to Alzheimer center Amsterdam, including the Amsterdam Dementia Cohort (ADC),[157] the 90+ study[158] and the Twin Study.[159,160] The ADC started collecting samples in the year 2000 and is an ongoing, observational follow-up study of patients who visited the memory clinic of the Alzheimer Center at Amsterdam UMC, location VU University medical center (VUmc).[157] Dementia was diagnosed according to diagnostic guidelines for neurodegenerative disease.[161–164] The Innovative Medicine Initiative European Information Framework for AD (EMIF-AD) 90+ study includes cognitively healthy elderly above 90 years old, and is aimed to identify factors associated with resilience to cognitive impairment in the oldest-old.[158] Monozygotic twins (one subject per twin pair) were invited from the Netherlands Twin Register[159] to participate in the PreclinAD study as part of the EMIF-AD project (http://www.emif.eu/).[160] A detailed description of the cohorts is provided in the supplemental materials.

Cognitively healthy subject cohort: Cognitively healthy volunteers were recruited at outpatient pre-operative screening services in four hospitals in Utrecht, the Netherlands (August 2008 until March 2010).[132] We included patients undergoing spinal anesthesia for minor elective surgical procedures, who ranged between 18 and 60 years of age and had all four grandparents born in The Netherlands or other Northwestern European countries (Belgium, Germany, UK, France and Denmark). Each candidate participant received a telephone interview to exclude subjects with self-reported psychotic or major neurological disorders (e.g., stroke, brain tumors, neurodegenerative diseases) and to record any use of psychotropic medication.

An overview of characteristics from the cognitively healthy subjects cohort (N=490, all cognitive healthy controls) and memory clinic cohort (N=487) is presented in Table S10. The Amsterdam sample includes n=220 controls (75 normal cognition, 145 subjective cognitive decline), n=87 subjects with mild cognitive impairment (MCI) and n=180 patients with dementia. Patient groups in the Amsterdam sample differed from each other as expected, with the AD-type dementia group including more *APOE*-e4 carriers, less *APOE*-e2 carriers and having more abnormal AD CSF biomarkers compared to the MCI and healthy controls. Cognitively healthy subjects were less often female, were younger, had a lower *APOE*-e4 and a higher *APOE*-e2 frequency, as compared to the memory clinic subjects.

All participating studies were approved by their respective Medical Ethics Committee. Informed consent, either from the patient or from the legal representative, was obtained from all participants.

CSF data collection

Memory clinic cohort: CSF samples were collected via lumbar puncture, using a 25-gauge needle and syringe. CSF levels of amyloid-beta 42 (Aβ42), total tau (t-tau) and hyperphosphorylated 181 tau (p-tau) were determined as part of the diagnostic work-up, using enzyme-linked immunosorbent assays (ELISA) (Innotest: Fujirebio, Ghent, Belgium).[157] For the ADC cohort, CSF Aβ42 values were adjusted for drift over time as described previously.[165] Biomarker abnormality cut-offs for the ADC cohort are CSF Aβ42 < 813 pg/ml, CSF t-tau > 375 pg/ml, ratio t-tau/ Aβ42 > 0.52.[166]

Cognitively healthy subject cohort: A sample of 6 ml of CSF was obtained from each subject via lumbar puncture and immediately stored in fractions of 0.5 and 1 ml at −80 °C, as described previously. [132]

CSF metabolite processing

In total, 5,543 CSF metabolites were measured across three different metabolite assays at the West Coast Metabolomics Center at UC Davis, including GC-TOF MS (primary metabolism), CSH-QTOF MS/MS (complex lipids), and HILIC-QTOF MS/MS (biogenic amines) (https://metabolomics.ucdavis.edu/).

Metabolite levels were first screened for missingness across each cohort, removing any metabolites which had missing data for >20% of individuals. For remaining metabolites, we imputed missing values to half the median value for the corresponding metabolite across the cohorts, reasoning that these metabolites are likely present in quantities too low to detect within these individuals, and as done previously[127]. Inverse-rank normalization was performed on all metabolites to ensure normality for QTL mapping.

Genotyping and imputation

Memory clinic samples were genotyped with the Illumina Global Screening Array (GSA) and cognitively healthy control data was genotyped with the OmniExpress Exome array. Quality control prior to imputation has been described in depth elsewhere.[167] Autosomal genotypes were first filtered for SNPs with <2% SNP-missingness and >5% minor allele frequency (MAF) using plink[168] separately per cohort. Individuals with call rate <98% were excluded. Genotype vcf files were then uploaded into the TopMed server for imputation and liftover to hg38. Post-imputation quality was assessed by filtering variants with imputation R2<0.3, providing about 8 million SNPs per cohort for downstream analyses. Imputed genotypes were then merged between the two cohorts, and then once more filtered for variants with <2% SNP-missingness and >5% MAF.

QTL mapping pipeline

Plink v1.90b[168] –linear function was used to perform the QTL mapping across each of the metabolites separately. Covariates were adjusted via the –covar flag, including age, sex, and first 3 genotyping PCs. Initially cohorts were analyzed separately and the resulting effect sizes for any metabolite-SNP pairs found to be genome-wide significant in either cohort were then checked for replication. For the Amsterdam cohort we performed QTL mapping with additional covariate adjustment for diagnostic status, which did not significantly change results (data not shown). The effect sizes of each of these loci per metabolite were tested for concordance via Pearson's correlation. Meta analysis was performed after merging individual-level genotype data across both cohorts.

Replication across cohorts

In order to quantify the concordance of the associations between both cohorts, for each metabolite, we subsetted the resulting summary statistics to any metabolite-SNP pairs associated with $p<6e^{-11}$ in either cohort. In total, we identified 30 metabolites with significant genetic regulation in the cognitively healthy controls, and 29 metabolites in the memory clinic samples, of which 16 metabolites were significant after Bonferroni correction in both cohorts with the same direction of effect at the same SNPs. Of the remaining 27 mQTL that showed Bonferroni-significance in one cohort but not the other, 19 mQTL were nominally significant in the other cohort, with a p-value of at least 0.05 and same direction of effect at the same SNP, resulting in a direct replication rate of 81%. Out of the eight metabolite-SNP pairs that did not directly replicate, six were significant in the Memory clinic cohort but not replicated in the Cognitively healthy cohort, and the remaining two were not replicated in the Memory clinic cohort. No SNPs with LD R2 > 0.5 replicate with P < 0.05 in the opposing cohort, though when we consider the meta-analysis summary statistics, all SNP-metabolite associations are replicated with P <= 0.0088 (see Table S3.2). Furthermore, these eight specific instances each

have at most two SNPs passing the Bonferroni-corrected threshold, suggesting that they are borderline significant associations.

Variant annotation

We downloaded the GENCODE v.41 basic gene annotation GTF file to interpret potential functional consequences of the loci found to be associated with metabolites. We manually define promoters as the region 10Kb upstream of the TSS (using bedtools[169] –flank -s -r 10000 -l 0), introns as regions within "gene" annotations that are not already covered by "exons" (bedtools –subtract), and intergenic regions as any region not covered by any annotations (bedtools –complement). Independent loci were selected per metabolite by subsetting the genotype files to only the SNPs that reached at least $p<6e10^{-11}$ significance, then clumping (using plink –clump command) these SNPs into separate regions with LD R2 < 0.2. All remaining SNPs for any metabolites were then concatenated, deduplicated, and reformatted into a BED format file for use with the bedtools –intersect command.

CSF metabolome-wide association study pipeline

To identify metabolites whose local-regulated CSF level is associated with brain-related phenotypes, we performed MWAS analyses using FUSION software (http://gusevlab.org/projects/fusion). First, we generated weights for all CSF metabolites with at least one genome-wide significant signal in the cohorts combined (*P*<5e-8). We restricted to loci +/- 500Kb around the lead SNPs per each metabolite, with the note that some metabolites were polygenic and had more than one associated genetic loci. For polygenic metabolites, any significant SNPs outside of the +/-500Kb range of another genome-wide significant SNP were included as a separate model. Age, sex, study site and the first three genotype PCs were included as covariates in the –covar flag of the FUSION.compute_weights.R script.

We used the FUSION.assoc_test.R script (default settings) to test for association between the CSF metabolite weights and GWAS for Alzheimer's disease[3], dementia with Lewy bodies[140], stroke[141], amyotrophic lateral sclerosis[142], bipolar disorder[1], schizophrenia[2], major depressive disorder[143], ADHD[94], insomnia[144] and alcohol abuse disorder.[145] To account for LD structure we used 1000 Genomes data (all ancestries, build 38) data as LD reference panel. The --coloc[65] flag was included to perform colocalization on any metabolites that had an association with the trait of interest with *P*-TWAS < 0.05. Colocalization analysis further narrowed down the metabolite-trait associations to those with a single variant influencing both the CSF metabolite level and trait, and associations with a PP4>0.8 were classified as having supporting colocalization evidence.
Models include between 889 and 7431 SNPs. The top1 model was chosen as the best model type for 62.3% of the metabolites, followed by LASSO models (21.4%), elastic net models (10%), and blup (6.4%).

Results on gene–tissue associations per phenotype were corrected for multiple comparisons using a 5% FDR significance threshold. Significant MWAS loci were identified as novel if the strongest associated SNP was not nominally significant (*P* > 1e-5) in the corresponding GWAS within ±1 Mb of the transcriptional start site of the gene's region.

Fine-mapping of mQTLs

To predict which SNP(s) within mQTL associations were most likely to be causal, we used FINEMAP,[170] SuSiE,[171] PolyFun-FINEMAP, and PolyFun-SuSiE[172] as fine-mapping methods. The summary statistics for these metabolites were lifted over to hg19 to match the UKBB LD reference panel as well as the UKBB functional annotations[172], both of these are composed of British ancestry individuals. Top loci to be fine-mapped were defined by selecting the SNP with the lowest p-value for each metabolite and including any SNPs within the region 500kb upstream and 500kb downstream of the lead SNP. Results were processed for credible sets

with summed fine-mapping posterior inclusion probability (PIP) of at least 0.5 across fewer than 10 SNPs, and SNP-associations that overlapped across multiple methods were prioritized as most likely to be the causal.

Integration with brain eQTLs

To identify mQTL that colocalize with eQTL for brain-specific gene expression, we lifted over the PsychENOCDE TWAS weights from hg19 to hg38 and then performed a FUSION TWAS analysis. Each of the summary statistics for metabolites with Bonferroni-significant loci were included as separate input GWAS for the FUSION.assoc_test.R script with the PyschENCODE TWAS weights, using the 1000 Genomes European-specific (EUR) LD panel. The --coloc 0.05 flag was included to perform colocalization on any metabolites that were associated with gene expression with *P*-TWAS < 0.05. We consider any metabolite-gene pairs with coloc.PP3 > coloc.PP4 and coloc.PP4 > 0.8 to be significantly colocalized.

In a similar manner, we also applied the isoform-level TWAS method[147] using precomputed weights per isoform derived from adult PsychENCODE data, as provided by the isoTWAS developers. In total we tested 7,530 genes, each with varying numbers of isoforms, that had positive heritability with a p-value < 0.05 within the PsychENCODE data. Similar to the TWAS analyses, we used the 1000 Genomes EUR LD reference panel. We then performed probabilistic fine-mapping of the significant associations, and filtered the resulting statistics to those transcripts in the credible sets with an adjusted screen p-value less than 0.05, permutation p-value less than 0.05, and PIP >= 0.8.

Data availability

GWAS summary statistics on all CSF metabolite levels that were generated as part of this study have been deposited to the European Bioinformatics Institute GWAS Catalog (https://www.ebi.ac.uk/gwas/) under accession no. GCSTXXXXX.

GENCODE v.41 basic gene annotation GTF file from https://www.gencodegenes.org/human/stats_41.html

UKBB LD reference panel available from Amazon S3 bucket s3://broad-alkesgroup-ukbb-ld/UKBB_LD/

UKBB functional annotations available from Amazon S3 bucket https://broad-alkesgroup-ukbb-ld.s3.amazonaws.com/UKBB_LD/baselineLF_v2.2.UKB.polyfun.tar.gz

PsychENCODE TWAS weights are available from http://resource.psychencode.org/.

PsychENCODE isoTWAS weights are available from https://zenodo.org/record/6795947#.Y8mi2-zMLBI.

Summary statistics on Alzheimer's disease[3], dementia with Lewy bodies[140] and amyotrophic lateral sclerosis[142] are publicly available at the European Bioinformatics Institute GWAS Catalog under accession no. GCST90027158, GCST90001390 and GCST90027163, respectively.

Summary statistics on bipolar disorder[1], schizophrenia[2], major depressive disorder[143], attention deficit hyperactivity disorder (ADHD)[94] and alcohol abuse disorder[145] are publicly available on the psychiatric genomics consortium (PGC) website (https://www.med.unc.edu/pgc/results-and-downloads).

The MEGASTROKE consortium, launched by the International Stroke Genetics Consortium, has published the stroke[141] summary statistics at https://www.megastroke.org. Full insomnia[144] summary statistics for UKB and the top 10,000 SNPs for 23andMe are available at https://ctg.cncr.nl/software/summary_statistics/.

Chapter 3 Figures

Figure 3.1. Overview of samples included in the CSF QTL mapping study.

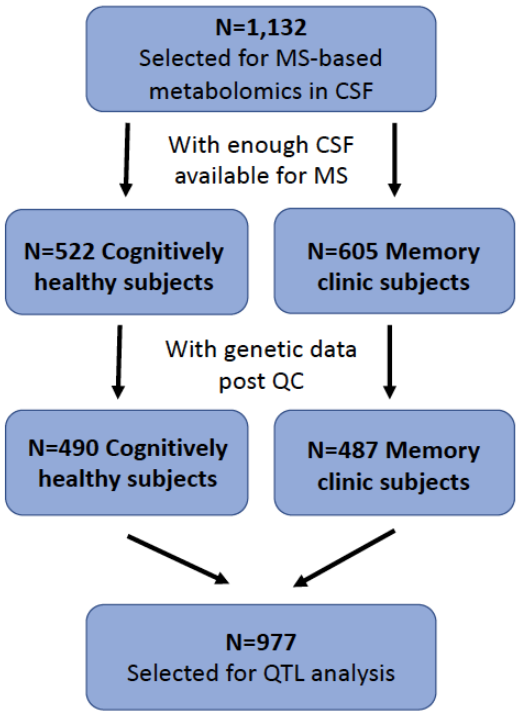


Figure 3.2. **Manhattan plot and genetic architecture of mQTL associations**

(A) Manhattan plot of the CSF metabolites that had at least one genome-wide significant mQTL association (*P*<6e-11) in the total sample (n=977). Each point depicts a distinct mQTL association between genetic variant and CSF metabolite levels. Green represents CSF metabolites measured using HILIC-QTOF MS/MS (biogenic amines, 1,618 metabolites), blue represents metabolites measured with GC-TOF MS (primary metabolism, 393 metabolites) and red represents CSH-QTOF MS/MS (complex lipids, 3,532 metabolites). Only CSF metabolites with annotation and *P*<6e-11 have a label. Novel CSF mQTL are depicted in bold and with an asterisk. The red line depicts the significance threshold of *P*<6e-11 (Bonferroni correction for 754 independent CSF signals), the yellow line depicts *P*<5e-8. Only results with a *P*<1e-8 and

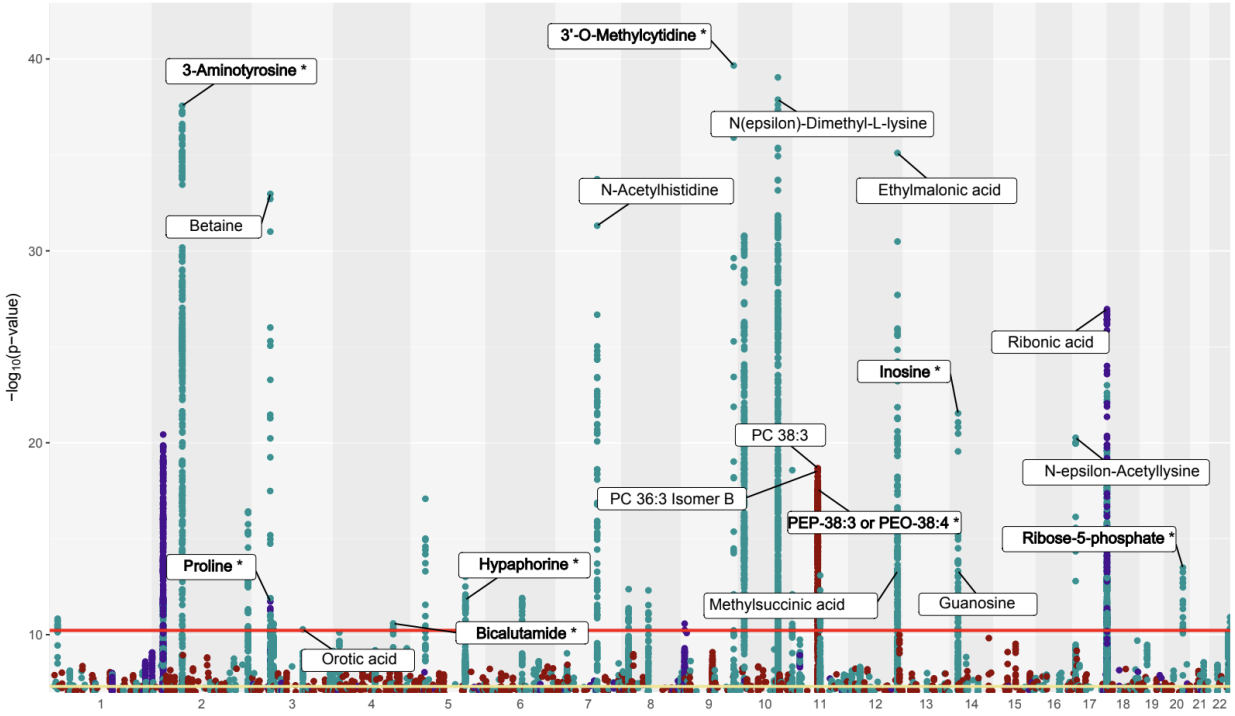
P>1e-40 are plotted.

(B) Using the GENCODE basic gene annotation, we first subsetted to 19,100 protein-coding genes, and we counted exons and UTRs as defined within the file. Promoters were defined as the region 10Kb upstream of a gene TSS, introns as regions within gene annotations not already covered by exons, and intergenic regions as any region not covered by any previous annotations. The y-axis denotes the number of independent SNPs, with some SNPs counted more than once for multiple metabolites.

(C) Histogram of polygenicity depicting number of metabolites associated with each independent locus.

(D) Histogram of pleiotropy depicting the number of independent loci associated with each metabolite.

A.



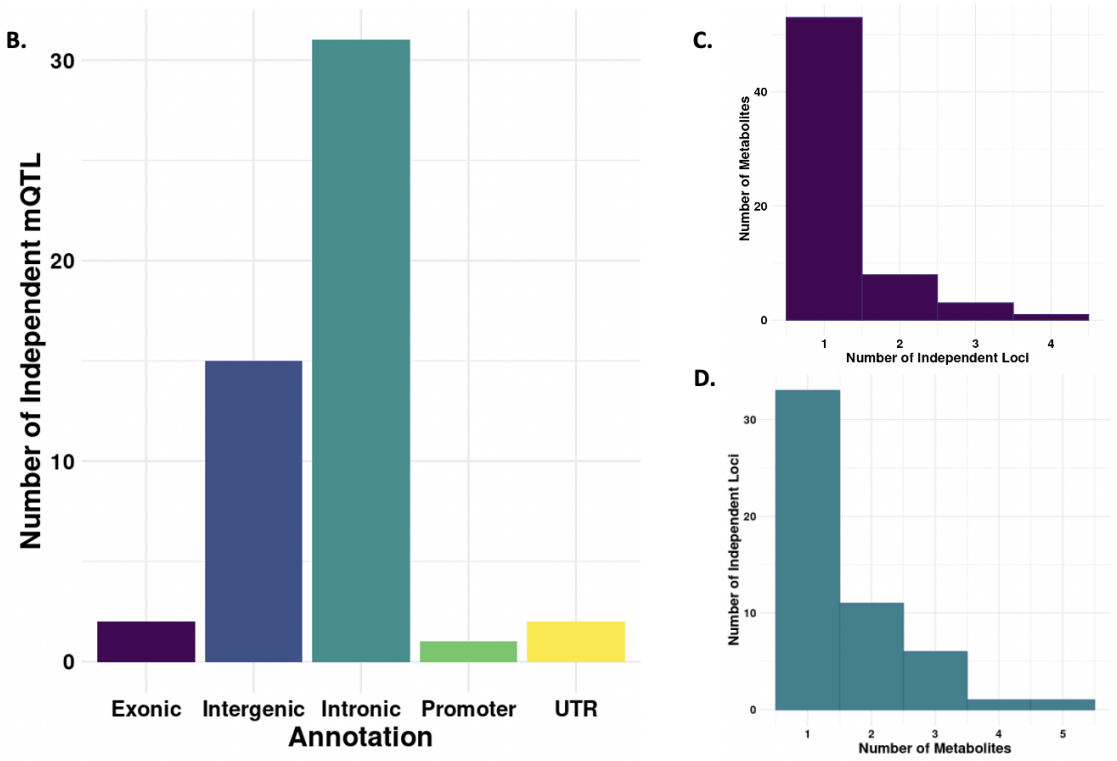

B.
Number of Independent mQTL
30
20
10
0
Exonic
Intergenic
Intronic
Promoter
UTR
Annotation
C.
Number of Metabolites
40
20
0
1
2
3
4
Number of Independent Loci
D.
Number of Independent Loci
30
20
10
0
1
2
3
4
5
Number of Metabolites

Figure 3.3. Heatmap of Z scores of CSF metabolites with at least one significant metabolome-wide significant association with one of the brain-related traits.

This visualization includes metabolite-phenotype associations with at least one significant (*P*-FDR<0.05) association. Blank squares indicate that CSF metabolite weights were not sufficiently predictive. # depicts *P*-FDR<0.10, * *P*-FDR<0.05, ** *P*-FDR<0.0001*** and *P*-FDR<0.00001.

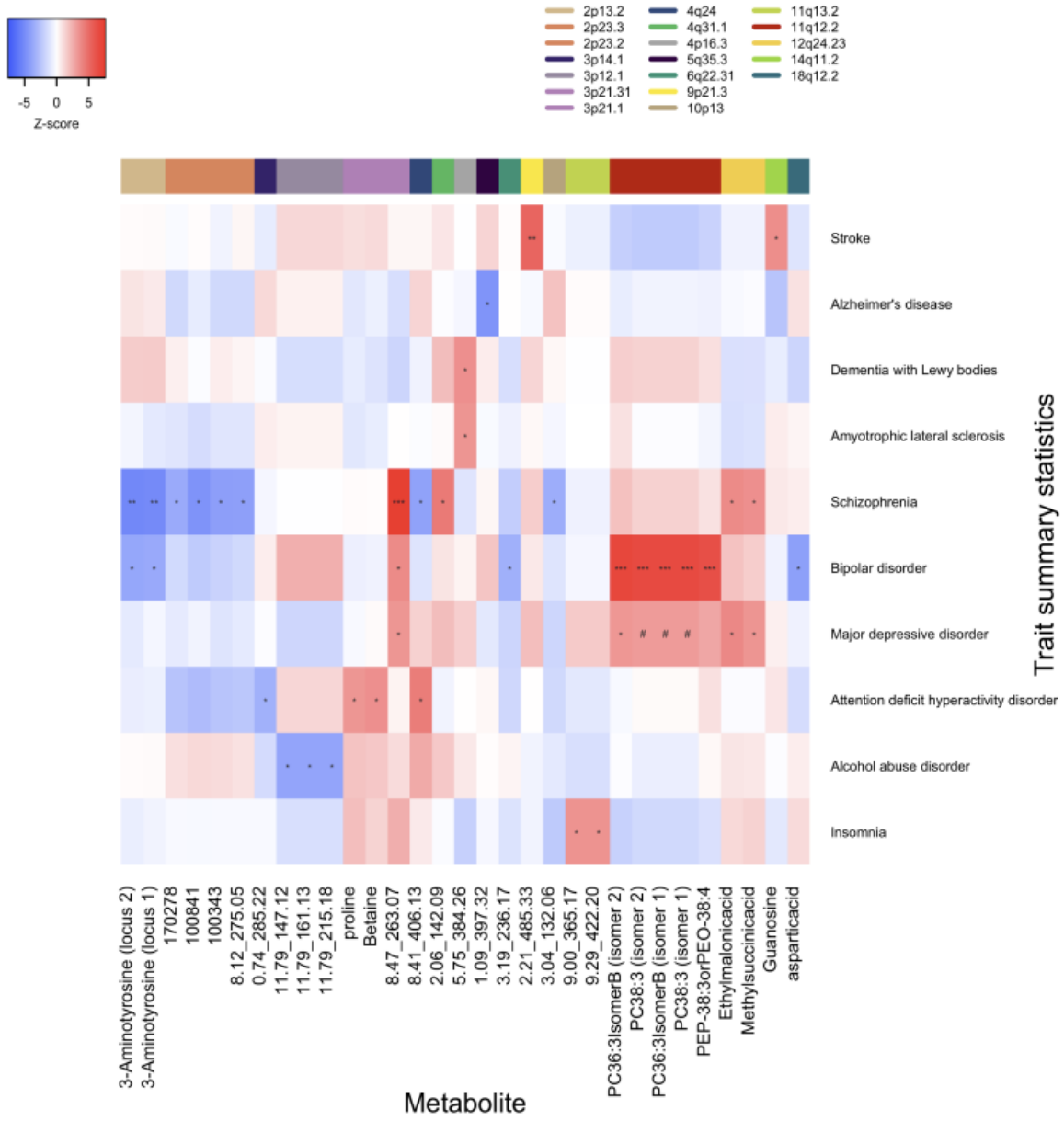
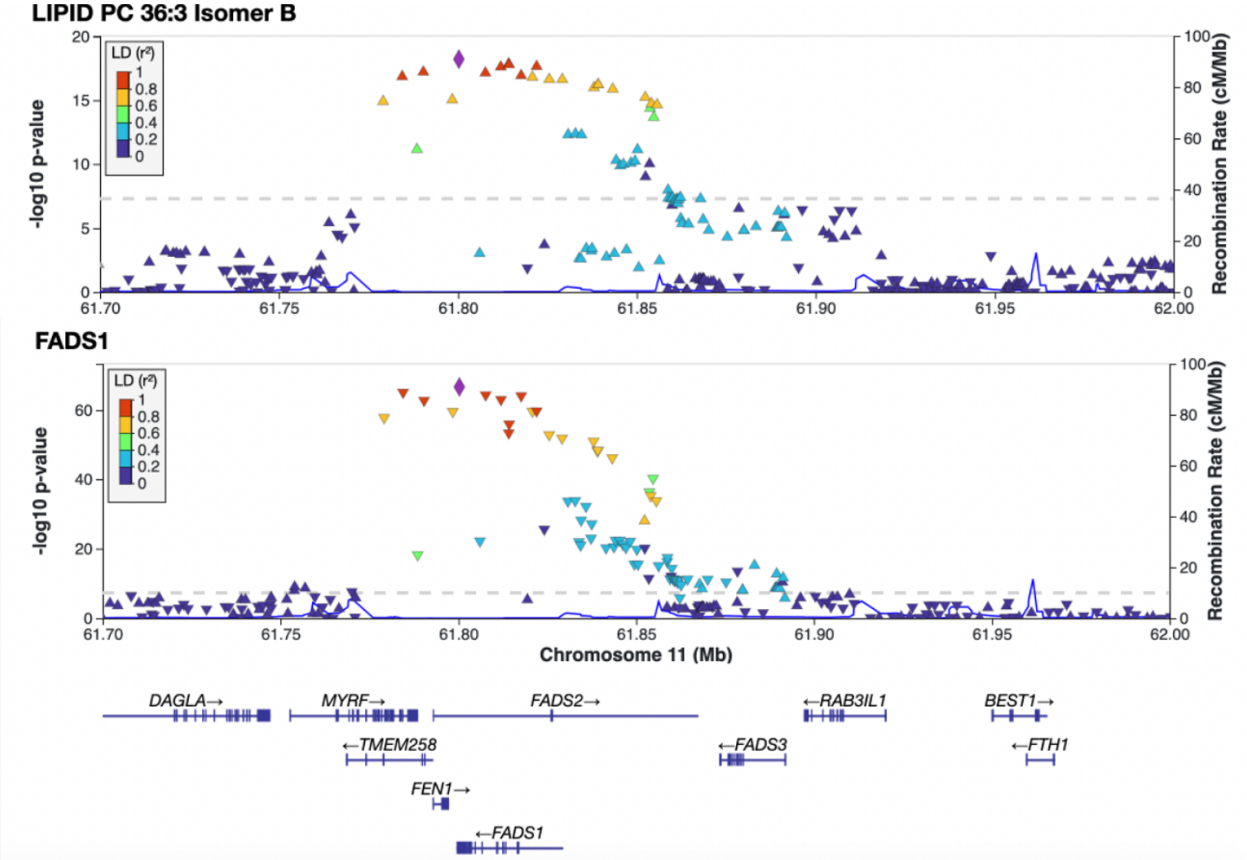


Figure 3.4. Phosphatidylcholine QTL colocalization with *FADS1* locus.

Locus zoom plots, top panel shows associations with lipid PC 36:3 isomer B levels at the *FADS1*/*FADS2*/*TMEM258* locus and their LD relative to the top SNP; bottom panel shows eQTL SNPs for *FADS1*. The diamond shape denotes the lead SNP whereas the orientation of the triangles denote the direction of the effect. The gray dashed line indicates the significance threshold of p=5e-8.



Conclusions

The past decade of genomic research has seen significant progress in unraveling the genetic basis of psychiatric disorders like bipolar disorder and schizophrenia, and neurodegenerative conditions such as Alzheimer's disease. However, the exact molecular mechanisms by which common genetic variation leads to risk of these phenotypes still remains elusive. Research into AD has uncovered the large effects of the *APOE4* allele on risk of developing the disease, though specific causal pathways, explaining how the aggregation of amyloid-beta and phosphorylated tau tangles leads to cognitive decline, remain unclear. For psychiatric disorders, the added challenge of having no clear biomarkers by which to distinguish diagnoses or measure effects of treatment, and relatedly the lack of genetic risk loci with substantial effect, has hindered the development of treatment options for patients.

More recently, the integration of omics-data in genome-wide contexts has provided valuable insights into the functional consequences of genetic risk loci. GWAS have been the backbone of complex trait research (as opposed to their Mendelian[173] counterparts) since the first genome-wide studies were implemented in the early 2000s.[174,175] Such studies are hypothesis-generating, not candidate gene driven, attempting to assay the entire genome in an agnostic manner to find potentially novel loci associated with a given trait. Genotyping for GWAS is not entirely unbiased, however, since the majority of SNP-arrays used are ascertained for common variants in European populations,[176] excluding potentially relevant alleles that are common in other ancestries. There has been a recent trend in the field toward whole-genome sequencing in place of genotyping to better capture all possible variations, and though this is currently much more expensive than genotyping, as technology improves such costs decrease. Similarly, genome-wide association studies tend to focus on common alleles (with minor allele frequencies no less than 1%), though rare genetic variation has also proved insightful for

elucidating mechanisms of common disease,[177,178] providing another avenue of exploration through whole-genome sequencing studies. There is also an ever-present need to increase the sample sizes of such studies, especially in the context of psychiatric phenotypes for which the genetic effects on the phenotype are very small. Overall, GWAS have been successful in identifying numerous genetic variants associated with complex traits, but understanding the biological mechanisms of the non-coding variants underlying these associations has proven challenging, thus interest in functional molecular data has blossomed.

Transcriptomics attempts to connect non-coding variation back to the gene as the main unit of interest, focusing on how common genetic variants influence the expression of specific genes. While this has proved insightful in many contexts, studying brain-related phenotypes remains challenging due to the difficulty in accessing brain tissue in living donors.[15] Gene expression is known to vary for different tissues or even cell types within tissues, though here we leverage the knowledge that cis-effects of genetic variants on gene expression tend to be correlated.[16,17] Interesting genes like *LINC00933* from the fibroblast project and *CAMKK2* from the immune cell types project were found to be genetically regulated by and colocalized with a locus associated with bipolar disorder, thus are potential candidates for potential drug targets and *in-vitro* follow-up studies.

The investigation of chromatin accessibility in fibroblasts using ATAC-seq also yielded insights into the regulatory landscape of the genome. Notably, the study found twice the number of associations between chromatin accessibility and complex diseases compared to gene expression, supporting the notion that gene expression is more context-dependent than chromatin accessibility.[14,29] Moreover, the study demonstrated an improvement in fine-mapping of QTL, identifying regulatory variants that were previously overlooked by traditional closest-gene methods, wherein the gene most proximal to the SNP of interest is considered its

regulatory target. We found that not all QTL SNPs and open chromatin regions necessarily regulate the nearest gene, emphasizing the importance of considering long-range enhancer-promoter interactions in gene regulation. These discoveries contribute to our understanding of the intricate relationship between chromatin accessibility, gene regulation, and disease, offering new avenues for investigating the molecular basis of complex traits and diseases.

Similarly, metabolomics, the study of small molecules involved in cellular processes, has emerged as a valuable tool in psychiatric genetics.[179] In this study, we found 54 novel mQTL associations and we replicated 11 mQTL associations previously reported in CSF and other biofluids. Colocalization with brain-eQTL revealed that most metabolites seemed to be regulated by the gene locus that encodes the protein for which they are substrates or products, though the implementation of isoTWAS also revealed other potential mechanisms not detectable at the gene-level. One particularly interesting finding was for several lipid levels associated with the *FADS* locus, which has also been implicated as a risk region for bipolar disorder.[1,153] Lipid dysregulation has been suspected to play a role not only in psychiatric disorders,[180,181] but also in neurodegenerative disease as well.[154,182]

While the work presented in this dissertation has yielded novel insights into the genetic architecture of and genes involved in neuropsychiatric phenotypes, there is still much more work needed in order to see improvements in clinical outcomes for patients. Regarding future directions, a study assessing chromatin interactions via a next-generation sequencing technology like chromatin capture Hi-C[183] could be applied to validate the SNP-chromatin-gene pathways identified in the fibroblast project *in vitro*. Neuronal iPSCs (induced pluripotent stem cells) derived from the fibroblasts[184] also could be used to validate that the trait-associated genes found in fibroblasts or blood are similarly regulated in neurons. Since psychiatric disorders are hypothesized to involve neurodevelopmental dysregulation, neuronal iPSCs from

patients and controls is a suitable model for such processes.[185] One challenge with studying neuropsychiatric disease *in vitro*, however, is how to capture the manifestation of complex symptoms like changes in mood or hallucinations. While these exact endophenotypes cannot be studied with cellular models, cells are useful for studying other phenotypes thought to be relevant to bipolar disorder and schizophrenia, including electrical changes in synapses due to aberrations in calcium signaling.[186,187] Cells have also been useful in the context of drug screening for psychiatric disorders[188] and for Alzhimer's disease.[189] It is also possible to capture proxies of complex behavior, like sleep disturbances in bipolar disorder through changes in circadian rhythms.[52,53] Animal models such as zebrafish[14,53] and mice[153] have also been used to illustrate the impact of genetic or environmental perturbations in the context of psychiatric disorders. Future studies focusing on the genes prioritized in this dissertation could involve perturbations of these genes, for instance over/underexpression through CRISPR-Cas editing, assessing the differences in specific cellular endophenotypes described here. Future studies could also involve direct stimulation of iPSCs with the metabolic compound(s) that were found to be associated with neuropsychiatric traits, examining whether cells derived from patients or controls behave differently in response to such stimulation.

Furthermore, much of the missing heritability problem could lie in variants and QTLs not studied here, including rare variants (minor allele frequency < 1%), structural variation (such as short tandem repeats, copy number variations, etc), and *trans*-acting QTL (>1Mb or inter-chromosomal). As technologies like whole genome sequencing, long-read sequencing, and single cell sequencing become cost-effective, large-scale generation of such genetic data is becoming increasingly possible. While the small effects of the GWAS risk loci and QTL are challenging to interpret, drugs acting on the downstream effects of such loci can have substantial impact.[190] Given that the development of drugs and other therapeutics is often a lengthy process, another use of the information gleaned in these studies involves risk

predictions. Neuropsychiatric phenotypes are highly polygenic, and methods to develop risk scores based on the additive effects of genetic variation are already underway,[1,191] though at this time these tools also require much improvement to be clinically applicable.[20]

One crucial aspect in human genetics as a field is the need for diverse samples. Historically, genetic studies have predominantly focused on populations of European ancestry, which has limited the generalizability of findings to other populations,[192] especially in the context of risk scores for polygenic disorders.[20,176] This lack of diversity hinders our ability to fully comprehend the genetic underpinnings of neuropsychiatric disease and other broad ranges of phenotypes. Recognizing this limitation, there is now a growing emphasis on collecting and analyzing data from diverse populations.[193,194] Some genetic variants may be specific to certain populations, while others may have a shared causal impact across different ancestries.[195] By incorporating data from multiple diverse ancestries, we can identify both shared and population-specific genetic contributors to various traits, leading to a more comprehensive understanding of the underlying molecular mechanisms.

In conclusion, the field of neuropsychiatric genomics continues to make strides in understanding the genetic basis of disorders of the brain. The integration of multiple omics levels has significantly enhanced our understanding of GWAS risk loci by uncovering the functional consequences and molecular mechanisms of disease-associated loci. We used sources like fibroblasts, blood, and CSF rather than brain tissue, showing that relevant cis-genetic effects can be detected despite some cell type specificity of QTL. Finally, we emphasize that the inclusion of diverse samples in genetic studies is crucial for capturing the full spectrum of genetic variation and improving the generalizability of findings. These advancements bring us closer to personalized and targeted approaches for the prevention, diagnosis, and treatment of neuropsychiatric disorders.

References

1. Mullins, N. et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat. Genet. 53, 817–829 (2021).
2. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 604, 502–508 (2022).
3. Bellenguez, C. et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat. Genet. 54, 412–436 (2022).
4. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. Am. J. Hum. Genet. 88, 294 (2011).
5. Karlsson, I. K. et al. Measuring heritable contributions to Alzheimer's disease: polygenic risk score analysis with twins. Brain Communications 4, (2022).
6. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 390, 1211–1259 (2017).
7. Schmitt, A., Malchow, B., Hasan, A. & Falkai, P. The impact of environmental factors in severe psychiatric disorders. Front. Neurosci. 8, (2014).
8. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 398, (2021).
9. Dembek, C. et al. The economic and humanistic burden of bipolar disorder in adults in the United States. Ann. Gen. Psychiatry 22, 1–10 (2023).
10. The economic burden of bipolar I disorder in the United States in 2015. J. Affect. Disord. 226, 45–51 (2018).
11. Michalak, E. E., Yatham, L. N., Kolesar, S. & Lam, R. W. Bipolar disorder and quality of life: a patient-centered perspective. Qual. Life Res. 15, (2006).

12. 2023 Alzheimer's disease facts and figures. Alzheimers. Dement. 19, (2023).
13. Zeng, B. et al. Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. Nat. Genet. 54, 161–169 (2022).
14. Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. 50, 538–548 (2018).
15. Liharska, L. E. et al. A study of gene expression in the living human brain. medRxiv 2023.04.21.23288916 (2023) doi:10.1101/2023.04.21.23288916.
16. Qi, T. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nat. Commun. 9, 1–12 (2018).
17. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017).
18. Nedergaard, M. Neuroscience. Garbage truck of the brain. Science 340, 1529–1530 (2013).
19. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).
20. Martin, A. R. et al. Current clinical use of polygenic scores will risk exacerbating health disparities. Nat. Genet. 51, 584 (2019).
21. Yang, G., Mishra, M. & Perera, M. A. Multi-Omics Studies in Historically Excluded Populations: The Road to Equity. Clin. Pharmacol. Ther. 113, 541–556 (2023).
22. Goddard, K. A., Hopkins, P. J., Hall, J. M. & Witte, J. S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am. J. Hum. Genet. 66, (2000).
23. Pritchard, J. K. & Donnelly, P. Case-control studies of association in structured or admixed populations. Theor. Popul. Biol. 60, (2001).
24. Hellwege, J. et al. Population Stratification in Genetic Association Studies. Curr. Protoc. Hum. Genet. 95, 1.22.1 (2017).

25. Geoffroy, E., Gregga, I. & Wheeler, H. E. Population-Matched Transcriptome Prediction Increases TWAS Discovery and Replication Rate. iScience 23, (2020).
26. Keys, K. L. et al. On the cross-population generalizability of gene expression prediction models. PLoS Genet. 16, e1008927 (2020).
27. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. Lond. B Biol. Sci. 368, (2013).
28. Baca, S. C. et al. Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. Nat. Genet. 54, 1364–1375 (2022).
29. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet. 50, 424–431 (2018).
30. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. 109, 21.29.1 (2015).
31. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, (2006).
32. Hou, K. et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nat. Genet. 55, 549–558 (2023).
33. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, (2009).
34. Roytman, M., Kichaev, G., Gusev, A. & Pasaniuc, B. Methods for fine-mapping with chromatin and expression data. PLoS Genet. 14, e1007240 (2018).
35. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. 48, 245–252 (2016).
36. Bigdeli, T. B. et al. Contributions of common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry. Mol. Psychiatry 25, (2020).
37. Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a

Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 12, e1001779 (2015).

38. Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the Global Biobank Meta-analysis Initiative. Cell Genomics 2, 100180 (2022).
39. Martin, A. R. et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. 100, 635 (2017).
40. Mahadevan, J. et al. Analysis of whole exome sequencing in severe mental illness hints at selection of brain development and immune related genes. Sci. Rep. 11, 1–10 (2021).
41. Ren, X., Mao, A., Tan, S., Liu, J. & Wei, X. Analysis of the association between MICA gene polymorphisms and schizophrenia. J. Clin. Lab. Anal. 36, (2022).
42. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. Genome Biol. 18, 1–21 (2017).
43. Li, J., Yoshikawa, A., Alliey-Rodriguez, N. & Meltzer, H. Y. Schizophrenia risk loci from xMHC region were associated with antipsychotic response in chronic schizophrenic patients with persistent positive symptom. Transl. Psychiatry 12, 92 (2022).
44. Tamouza, R., Krishnamoorthy, R. & Leboyer, M. Understanding the genetic contribution of the human leukocyte antigen system to common major psychiatric disorders in a world pandemic context. Brain Behav. Immun. 91, 731 (2021).
45. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427 (2014).
46. Debnath, M., Berk, M., Leboyer, M. & Tamouza, R. The MHC/HLA Gene Complex in Major Psychiatric Disorders: Emerging Roles and Implications. Current Behavioral Neuroscience Reports 5, 179–188 (2018).
47. Gandal, M. J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science 362, (2018).
48. Kohn, Y. & Lerer, B. Excitement and confusion on chromosome 6q: the challenges of

neuropsychiatric genetics in microcosm. Mol. Psychiatry 10, (2005).

49. Dick, D. M. et al. Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the National Institute of Mental Health Genetics Initiative. Am. J. Hum. Genet. 73, 107–114 (2003).
50. Middleton, F. A. et al. Genomewide Linkage Analysis of Bipolar Disorder by Use of a High-Density Single-Nucleotide–Polymorphism (SNP) Genotyping Assay: A Comparison with Microsatellite Marker Assays and Finding of Significant Linkage to Chromosome 6q22. Am. J. Hum. Genet. 74, 886 (2004).
51. CRCP CGRP receptor component [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/27297.
52. Francia, M. et al. Fibroblasts as an in vitro model of circadian genetic and genomic studies. bioRxiv 2023.05.19.541494 (2023) doi:10.1101/2023.05.19.541494.
53. Yamazaki, S. & Takahashi, J. S. Real-time luminescence reporting of circadian gene expression in mammals. Methods Enzymol. 393, 288–301 (2005).
54. Villegas, J. & McPhaul, M. Establishment and culture of human skin fibroblasts. Curr. Protoc. Mol. Biol. Chapter 28, (2005).
55. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).
56. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. 7, 500.
57. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357.
58. Hitz, B. C. et al. The ENCODE Uniform Analysis Pipelines. bioRxiv (2023) doi:10.1101/2023.04.04.535623.
59. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, 1–9 (2008).

60. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 9, 1–5 (2019).
61. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).
62. Lam, M. et al. RICOPILI: Rapid Imputation for COnsortias PIpeLIne. Bioinformatics 36, 930–933 (2020).
63. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–1358 (2012).
64. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. 93, 278 (2013).
65. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383 (2014).
66. Zhang, J. & Zhao, H. eQTL Studies: from Bulk Tissues to Single Cells. ArXiv (2023).
67. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. Science 369, (2020).
68. Mandric, I. et al. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. Nat. Commun. 11, 5504 (2020).
69. Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nature Communications vol. 11 Preprint at https://doi.org/10.1038/s41467-020-19015-1 (2020).
70. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. Genome Biol. 22, 102 (2021).
71. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. 37, 773–782 (2019).
72. Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression

with prior derived from single-cell data. Genome Res. (2021) doi:10.1101/gr.268722.120.

73. Schwarz, T. et al. Powerful eQTL mapping through low coverage RNA sequencing. Preprint at https://doi.org/10.1101/2021.08.08.455466.
74. Khandaker, G. M., Dantzer, R. & Jones, P. B. Immunopsychiatry: important facts. Psychol. Med. 47, 2229–2237 (2017).
75. Werner, M. C. F. et al. Immune marker levels in severe mental disorders: associations with polygenic risk scores of related mental phenotypes and psoriasis. Transl. Psychiatry 12, 1–8 (2022).
76. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. Nat. Neurosci. 18, 199–209 (2015).
77. Le Clerc, S. et al. HLA-DRB1 and HLA-DQB1 genetic diversity modulates response to lithium in bipolar affective disorders. Sci. Rep. 11, 17823 (2021).
78. Chernecky, C. C. & Berger, B. J. Laboratory Tests and Diagnostic Procedures. (W.B. Saunders Company, 1997).
79. George-Gay, B. & Parker, K. Understanding the complete blood count with differential. J. Perianesth. Nurs. 18, 96–114; quiz 115–7 (2003).
80. Schmiedel, B. J. et al. Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type-dependent effects of disease-risk variants. Science immunology 7, (2022).
81. Chen, L. et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell 167, 1398–1414.e24 (2016).
82. Moffat, J. J., Ka, M., Jung, E.-M., Smith, A. L. & Kim, W.-Y. The role of MACF1 in nervous system development and maintenance. Semin. Cell Dev. Biol. 69, 9–17 (2017).
83. Bryois, J. et al. Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders. Preprint at https://doi.org/10.1101/2021.10.09.21264604.
84. Westra, H.-J. et al. Cell Specific eQTL Analysis without Sorting Cells. PLoS Genet. 11,

e1005223 (2015).

85. Schmiedel, B. J. et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell vol. 175 1701–1715.e16 Preprint at https://doi.org/10.1016/j.cell.2018.10.022 (2018).
86. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. 53, 1290–1299 (2021).
87. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. Preprint at https://doi.org/10.1101/024083.
88. Consortium, T. S. W. G. of T. P. G., The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S., Walters, J. T. R. & O'Donovan, M. C. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. Preprint at https://doi.org/10.1101/2020.09.12.20192922.
89. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. 50, 668–681 (2018).
90. Walters, R. K. et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat. Neurosci. 21, 1656–1669 (2018).
91. Johnson, E. C. et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. Lancet Psychiatry 7, 1032–1045 (2020).
92. Watanabe, K. et al. Author Correction: A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. 52, 353 (2020).
93. Jansen, P. R. et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nat. Genet. 51, 394–403 (2019).
94. Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat. Genet. 51, 63–75 (2019).
95. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat. Genet. 51, 404–413 (2019).

96. Akbarian, S. et al. The PsychENCODE project. Nature Neuroscience vol. 18 1707–1712 Preprint at https://doi.org/10.1038/nn.4156 (2015).
97. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat. Genet. 47, 1457–1464 (2015).
98. van der Harst, P. et al. Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369–375 (2012).
99. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).
100. Wang, Y.-F. et al. Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. Nat. Commun. 12, 772 (2021).
101. Eames, H. L., Corbin, A. L. & Udalova, I. A. Interferon regulatory factor 5 in human autoimmunity and murine models of autoimmune disease. Transl. Res. 167, 167–182 (2016).
102. Akkouh, I. A. et al. Exploring lithium's transcriptional mechanisms of action in bipolar disorder: a multi-step study. Neuropsychopharmacology vol. 45 947–955 Preprint at https://doi.org/10.1038/s41386-019-0556-8 (2020).
103. Krebs, C. E. et al. Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect. Psychol. Med. 50, 2575–2586 (2020).
104. Johnson, R. et al. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. Genome Medicine vol. 14 Preprint at https://doi.org/10.1186/s13073-022-01106-x (2022).
105. Zhou, W. et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genom 2, 100192 (2022).
106. Smyth, G. K. limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor 397–420 Preprint at https://doi.org/10.1007/0-

387-29362-0_23.

107. Munkholm, K., Peijs, L., Vinberg, M. & Kessing, L. V. A composite peripheral blood gene expression measure as a potential diagnostic biomarker in bipolar disorder. Transl. Psychiatry 5, e614 (2015).

108. Lisé, M.-F. et al. Myosin-Va-interacting protein, RILPL2, controls cell shape and neuronal morphogenesis via Rac signaling. J. Cell Sci. 122, 3810–3821 (2009).

109. Ling, N. X. Y. et al. Functional analysis of an R311C variant of Ca 2 -calmodulin-dependent protein kinase kinase-2 (CaMKK2) found as a de novo mutation in a patient with bipolar disorder. Bipolar Disorders vol. 22 841–848 Preprint at https://doi.org/10.1111/bdi.12901 (2020).

110. Alda, M. LITHIUM IN THE TREATMENT OF BIPOLAR DISORDER: PHARMACOLOGY AND PHARMACOGENETICS. Mol. Psychiatry 20, 661 (2015).

111. Grimes, C. A. & Jope, R. S. CREB DNA binding activity is inhibited by glycogen synthase kinase-3 beta and facilitated by lithium. J. Neurochem. 78, (2001).

112. Sakamoto, K., Karelina, K. & Obrietan, K. CREB: a multifaceted regulator of neuronal plasticity and protection. J. Neurochem. 116, 1–9 (2011).

113. Karege, F. et al. Association of AKT1 gene variants and protein expression in both schizophrenia and bipolar disorder. Genes Brain Behav. 9, 503–511 (2010).

114. Consortium, T. G. & The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science vol. 369 1318–1330 Preprint at https://doi.org/10.1126/science.aaz1776 (2020).

115. Caggiano, C. et al. Health care utilization of fine-scale identity by descent clusters in a Los Angeles biobank. Preprint at https://doi.org/10.1101/2022.07.12.22277520.

116. Wu, P. et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform 7, e14325 (2019).

117. Farioli-Vecchioli, S. et al. Btg1 is Required to Maintain the Pool of Stem and Progenitor

Cells of the Dentate Gyrus and Subventricular Zone. Front. Neurosci. 6, 124 (2012).

118. Kraus, W. E. et al. Metabolomic Quantitative Trait Loci (mQTL) Mapping Implicates the Ubiquitin Proteasome System in Cardiovascular Disease Pathogenesis. PLoS Genet. 11, e1005553 (2015).

119. Shin, S.-Y. et al. An atlas of genetic influences on human blood metabolites. Nat. Genet. 46, 543–550 (2014).

120. Prakash, S. Human metabolic individuality in biomedical and pharmaceutical research. Circulation. Cardiovascular genetics vol. 4 714–715 (2011).

121. Long, T. et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat. Genet. 49, 568–578 (2017).

122. Nag, A. et al. Genome-wide scan identifies novel genetic loci regulating salivary metabolite levels. Hum. Mol. Genet. 29, 864–875 (2020).

123. Schlosser, P. et al. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. Nat. Genet. 52, 167–176 (2020).

124. Yang, C. et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. Nat. Neurosci. 24, 1302–1312 (2021).

125. Tijms, B. M. & Visser, P. J. Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. Alzheimer's & Dementia vol. 16 Preprint at https://doi.org/10.1002/alz.037184 (2020).

126. Yan, J., Kuzhiumparambil, U., Bandodkar, S., Dale, R. C. & Fu, S. Cerebrospinal fluid metabolomics: detection of neuroinflammation in human central nervous system disease. Clin Transl Immunology 10, e1318 (2021).

127. Dong, R. et al. CSF metabolites associate with CSF tau and improve prediction of Alzheimer's disease status. Alzheimers. Dement. 13, e12167 (2021).

128. Sasayama, D. et al. Genome-wide quantitative trait loci mapping of the human cerebrospinal fluid proteome. Hum. Mol. Genet. 26, 44–51 (2017).

129. Hansson, O. et al. The genetic regulation of protein expression in cerebrospinal fluid. EMBO Mol. Med. 15, e16359 (2023).
130. Panyard, D. J. et al. Cerebrospinal fluid metabolomics identifies 19 brain-related phenotype associations. Preprint at https://doi.org/10.1101/2020.02.14.948398.
131. Tahir, U. A. et al. Whole Genome Association Study of the Plasma Metabolome Identifies Metabolites Linked to Cardiometabolic Disease in Black Individuals. Nat. Commun. 13, 4923 (2022).
132. Luykx, J. J. et al. Genome-wide association study of monoamine metabolite levels in human cerebrospinal fluid. Mol. Psychiatry 19, 228–234 (2014).
133. Frankish, A. et al. GENCODE 2021. Nucleic Acids Res. 49, D916–D923 (2021).
134. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. 51, 1339–1348 (2019).
135. Flynn, E. D. & Lappalainen, T. Functional Characterization of Genetic Variant Effects on Expression. Annu Rev Biomed Data Sci 5, 119–139 (2022).
136. Jo, B.-S. & Choi, S. S. Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. 13, 112 (2015).
137. Niu, H.-M. et al. Comprehensive functional annotation of susceptibility SNPs prioritized 10 genes for schizophrenia. Transl. Psychiatry 9, 1–12 (2019).
138. Yin, X. et al. Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. Nature Communications vol. 13 Preprint at https://doi.org/10.1038/s41467-022-29143-5 (2022).
139. Koshiba, S. et al. Identification of critical genetic variants associated with metabolic phenotypes of the Japanese population. Commun Biol 3, 662 (2020).
140. Chia, R. et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. Nat. Genet. 53, 294–303 (2021).

141. Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat. Genet. 50, 524–537 (2018).
142. van Rheenen, W. et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nat. Genet. 53, 1636–1648 (2021).
143. Howard, D. M. et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat. Neurosci. 22, 343–352 (2019).
144. Watanabe, K. et al. Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. Nat. Genet. 54, 1125–1132 (2022).
145. Sanchez-Roige, S. et al. Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. Am. J. Psychiatry 176, 107–118 (2019).
146. Sall, S., Thompson, W., Santos, A. & Dwyer, D. S. Analysis of Major Depression Risk Genes Reveals Evolutionary Conservation, Shared Phenotypes, and Extensive Genetic Interactions. Front. Psychiatry 12, 698029 (2021).
147. Bhattacharya, A. et al. Isoform-level transcriptome-wide association uncovers extensive novel genetic risk mechanisms for neuropsychiatric disorders in the human brain. medRxiv 2022.08.23.22279134 (2022) doi:10.1101/2022.08.23.22279134.
148. Fogli, A. et al. Decreased guanine nucleotide exchange factor activity in eIF2B-mutated patients. Eur. J. Hum. Genet. 12, 561–566 (2004).
149. Keller, B. O., Wu, B. T. F., Li, S. S. J., Monga, V. & Innis, S. M. Hypaphorine is present in human milk in association with consumption of legumes. J. Agric. Food Chem. 61, 7654–7660 (2013).
150. Yee, S. W. et al. Deorphaning a Solute Carrier 22 family member, SLC22A15, through functional genomic studies. FASEB J. 34, 15734 (2020).

151. Auwerx, C. et al. Exploiting the mediating role of the metabolome to unravel transcript-to-phenotype associations. Elife 12, (2023).
152. Ikeda, M. et al. A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. Mol. Psychiatry 23, 639–647 (2018).
153. Yamamoto, H. et al. GWAS-identified bipolar disorder risk allele in the FADS1/2 gene region links mood episodes and unsaturated fatty acid metabolism in mutant mice. Mol. Psychiatry (2023) doi:10.1038/s41380-023-01988-2.
154. Yin, F. Lipid metabolism and Alzheimer's disease: clinical evidence, mechanistic link and therapeutic promise. FEBS J. 290, 1420–1453 (2023).
155. Hammouda, S. et al. Genetic variants in FADS1 and ELOVL2 increase level of arachidonic acid and the risk of Alzheimer's disease in the Tunisian population. Prostaglandins Leukot. Essent. Fatty Acids 160, 102159 (2020).
156. Guo, K. et al. Transcription factor POU4F2 promotes colorectal cancer cell migration and invasion through hedgehog-mediated epithelial-mesenchymal transition. Cancer Science vol. 112 4176–4186 Preprint at https://doi.org/10.1111/cas.15089 (2021).
157. van der Flier, W. M. & Scheltens, P. Amsterdam Dementia Cohort: Performing Research to Optimize Care. J. Alzheimers. Dis. 62, 1091–1111 (2018).
158. Legdeur, N. et al. Resilience to cognitive impairment in the oldest-old: design of the EMIF-AD 90+ study. BMC Geriatr. 18, 289 (2018).
159. Boomsma, D. I. et al. Netherlands Twin Register: From Twins to Twin Families. Twin Res. Hum. Genet. 9, 849–857 (2006).
160. Konijnenberg, E. et al. The EMIF-AD PreclinAD study: study design and baseline cohort overview. Alzheimers. Res. Ther. 10, 75 (2018).
161. Albert, M. S. et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups

on diagnostic guidelines for Alzheimer's disease. Alzheimers. Dement. 7, (2011).

162. McKeith, I. G. et al. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. Neurology 89, (2017).

163. McKhann, G. M. et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers. Dement. 7, (2011).

164. Rascovsky, K. et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 134, 2456 (2011).

165. Tijms, B. M. et al. Unbiased Approach to Counteract Upward Drift in Cerebrospinal Fluid Amyloid-β 1-42 Analysis Results. Clin. Chem. 64, 576–585 (2018).

166. Duits, F. H. et al. The cerebrospinal fluid 'Alzheimer profile': easily said, but what does it mean? Alzheimers. Dement. 10, 713–723.e2 (2014).

167. Tesi, N. et al. Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease. Eur. J. Hum. Genet. 27, 244–253 (2018).

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

169. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics vol. 26 841–842 Preprint at https://doi.org/10.1093/bioinformatics/btq033 (2010).

170. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501 (2016).

171. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. 82, 1273–1300 (2020).

172. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of

complex trait heritability. Nat. Genet. 52, 1355–1363 (2020).

173. Freund, M. K. et al. Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. Am. J. Hum. Genet. 103, (2018).

174. Ozaki, K. et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat. Genet. 32, 650–654 (2002).

175. Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. Science 308, 385–389 (2005).

176. Martin, A. R. et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. Am. J. Hum. Genet. 108, (2021).

177. Wainschtein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat. Genet. 54, 263–273 (2022).

178. Singh, T. et al. Rare coding variants in 10 genes confer substantial risk for schizophrenia. Nature 604, 509 (2022).

179. Shih, P.-A. (betty). Metabolomics Biomarkers for Precision Psychiatry. Adv. Exp. Med. Biol. 1161, 101 (2019).

180. Cheon, S. Y. Impaired Cholesterol Metabolism, Neurons, and Neuropsychiatric Disorders. Exp. Neurobiol. 32, 57 (2023).

181. Burghardt, K. J., Gardner, K. N., Johnson, J. W. & Ellingrod, V. L. Fatty Acid Desaturase Gene Polymorphisms and Metabolic Measures in Schizophrenia and Bipolar Patients Taking Antipsychotics. Cardiovasc. Psychiatry Neurol. 2013, (2013).

182. Wong, M. W. et al. Dysregulation of lipids in Alzheimer's disease and their role as potential biomarkers. Alzheimers. Dement. 13, (2017).

183. Belton, J. M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 58, (2012).

184. Abdullah, A. I., Pollock, A. & Sun, T. The path from skin to brain: generation of functional neurons from fibroblasts. Mol. Neurobiol. 45, 586 (2012).

185. Tamburini, C. & Li, M. Understanding neurodevelopmental disorders using human pluripotent stem cell-derived neurons. Brain Pathol. 27, 508–517 (2017).
186. Hewitt, T. et al. Bipolar disorder-iPSC derived neural progenitor cells exhibit dysregulation of store-operated Ca2+ entry and accelerated differentiation. Mol. Psychiatry 1–14 (2023).
187. Miller, N. D. & Kelsoe, J. R. Unraveling the biology of bipolar disorder using iPS derived neurons. Bipolar Disord. 19, 544 (2017).
188. Lago, S. G., Tomasik, J. & Bahn, S. Functional patient-derived cellular models for neuropsychiatric drug discovery. Transl. Psychiatry 11, (2021).
189. Caldwell, A. B. et al. Endotype reversal as a novel strategy for screening drugs targeting familial Alzheimer's disease. Alzheimers. Dement. 18, 2117 (2022).
190. Lau, A. & So, H.-C. Turning genome-wide association study findings into opportunities for drug repositioning. Comput. Struct. Biotechnol. J. 18, 1639 (2020).
191. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–752 (2009).
192. Mooney, J. A. et al. Understanding the Hidden Complexity of Latin American Population Isolates. Am. J. Hum. Genet. 103, (2018).
193. Stevenson, A. et al. Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. BMJ Open 9, (2019).
194. Service SK et al. Distinct and shared contributions of diagnosis and symptom domains to cognitive performance in severe mental illness in the Paisa population: a case-control study. The lancet. Psychiatry 7, (2020).
195. Martin, A. R. et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. 107, 788–789 (2020).