ANNOTATING GENETIC RISK VARIANTS TO TARGET GENES USING Hi-C COUPLED MAGMA (H-MAGMA)

Nancy Sey

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Neuroscience in the School of Medicine

Chapel Hill
2022

Approved by:

Joyce Besheer

Hyejung Won

Karen Mohlke

Lisa Tarantino

Douglas Phanstiel

ABSTRACT

Nancy Sey: Annotating genetic risk variants to target genes using Hi-C Coupled MAGMA (H-MAGMA)
(Under the direction of Hyejung Won)

An outstanding goal in modern genomics is to systematically predict the functional outcome of non-coding variation associated with complex traits. To bridge the gap between non-coding variation and its functional impact, we developed Hi-C Coupled Multi-Marker Analysis of GenoMic Annotation (H-MAGMA), a framework that converts SNP associations into gene-level associations based on chromatin interaction profiles to assign variants to their target genes. Applying this approach, we identified key biological pathways implicated in a wide range of brain disorders and showed its utility in complementing other functional genomic resources such as expression quantitative trait loci (eQTL)-based variant annotation. We applied H-MAGMA to five psychiatric and four neurodegenerative disorders. We identified that H-MAGMA detects risk genes associated with brain disorders. Additionally, we identified excitatory neurons as the critical cell types underlying psychiatric disorders compared to neurodegenerative disorders. Furthermore, we identified that genes associated with psychiatric disorders are expressed during early brain development, while those associated with neurodegenerative disorders are expressed in later years. Next, we utilized H-MAGMA to pinpoint genes associated with cigarette smoking and alcohol use traits. We next characterized the underlying biological processes and critical cell types underlying substance use traits. We found that pathways including ethanol metabolic process and alcohol catabolic

process to be associated with alcohol use traits, while response to nicotinic and acetylcholinergic pathways were identified for cigarette smoking traits. Moreover, we identified dopaminergic, GABAergic, and serotonergic neurons in the midbrain as relevant cell types that may contribute to substance use etiology. Lastly, we provide a detailed protocol for generating the H-MAGMA variant-gene annotation file and provide additional annotation files for 28 tissues and cell types, with the hope of contributing a resource for researchers.

To Olivia, with great love and admiration

ACKNOWLEDGEMENTS

Completing this work would not have been possible without the help of all the incredible people I have met on this journey. To my advisor, Hyejung Won, thank you for taking a gamble on me when you accepted me as your first graduate student back in 2019. I have grown tremendously as a scientist and as a person through your incredible mentorship. Thank you for challenging me with a computational project even before I knew what `read.table`() meant. Thank you for encouraging me to seek out various opportunities when I didn't believe I had a shot at them. Lastly, thank you for helping me build up self-esteem by gently nudging me to give all those presentations I was too nervous to give. Your mentorship, kindness, and thoughtful nature is something I will forever cherish. To my lab colleagues, thank you for your support and help these past years. To Jessica - I am going to miss our afternoon teatime. To the friends I have made in graduate school, especially, Janay, Sam, Minna, Felix, and Ian, thank you all for the beautiful friendship and allowing me the space to vent when things got challenging and stressful. You have all been strong advocates and for that I will forever be grateful.

To my family and friends, this journey could not have begun without your love and support. Thank you mummy and daddy for being my rock. You have both been incredible role models and prayer warriors. Thank you for all the sacrifices you have made for Grace, Theresa, and I and for teaching us the importance of education. To my darling sisters - you are the best sisters anyone could ever hope for. Thank you for your continuous support, pep talks, and being my cheerleaders. Grace - I truly appreciate the tutoring and all the pastries.

Lastly, to the Muyengwa sisters, thank you for making me a part of your family. Dzidzai my "wife" - your words of encouragement and prayers, as well as delicious dinners have propelled me to where I am, and I can't wait to be the sugar mummy you deserve! Vimbai, my exuberant sunshine, I can't explain how impactful your friendship has been. Thank you for the wine nights.

The work outlined in this dissertation was supported through grants from the National Institute of Drug Abuse (R21DA051921); the National Institute of Mental Health (R00MH113823, DP2MH122403); the NARSAD Young Investigator Award from the Brain and Behavior Research Foundation; the National Science Foundation Graduate Research Fellowship Program (DGE-1650116); the Howard Hughes Medical Institute's James H. Gilliam Fellowship for Advanced Study Program; and the National Institute of General Medical Sciences (5T32GM067553.

TABLE OF CONTENTS

LIST OF FIGURES x

LIST OF TABLES xi

LIST OF ABBREVIATIONS xii

CHAPTER 1: GENERAL INTRODUCTION 1

CHAPTER 2: H-MAGMA FOR IMPROVED PREDICTION OF BRAIN DISORDER RISK GENES BY INCORPORATING BRAIN CHROMATIN INTERACTION PROFILES 7

Introduction 7

Results 9

H-MAGMA 9

Developmental trajectories of risk genes associated with brain disorders 11

Pathways implicated in brain disorders 12

Cell-type specificity 15

Cell-type specific gene mapping 16

Shared genetic architecture among brain disorders 17

Biological pathways underlying pleiotropy 18

Discussion 19

Methods 21

Figures 32

Tables 38

CHAPTER 3: CHROMATIN ARCHITECTURE IN ADDICTION CIRCUTRY IDENTIFIES RISK GENES AND POTENTIAL BIOLOGICAL MECHANISMS UNDERLYING CIGARETTE SMOKING AND ALCOHOL USE TRAITS 39

Introduction 39

Results 41

Epigenetic landscape of cortical and midbrain dopaminergic neurons 41

CN and DN H-MAGMA identifies genes and biological pathways underlying cigarette smoking and alcohol use traits 43

Cellular expression profiles of cigarette smoking and alcohol use risk genes convey cell types associated with substance use 47

Shared genetic architecture among substance use 49

Drug repurposing analysis 50

Discussion 51

Methods 54

Figures 66

Tables 74

CHAPTER 4: DETAILED PROTOCOL OF H-MAGMA AND EXPANDING THE TOOL TO NON-BRAIN CELL-TYPES 76

Introduction 76

Procedure 85

Anticipated results 96

Data availability 97

Figures 98

Tables 100

CHAPTER 5: GENERAL DISCUSSION 106

REFERENCES 113

LIST OF FIGURES

Figure 2.1. Schematic of the H-MAGMA approach....32

Figure 2.2. Comparison between H-MAGMA and cMAGMA....33

Figure 2.3. Spatiotemporal dynamics of brain-disorder risk genes....34

Figure 2.4. Developmental trajectories of brain disorder risk genes derived from cMAGMA....35

Figure 2.5. Cellular expression profiles of brain disorder-risk genes....36

Figure 2.6. Shared molecular mechanisms of psychiatric disorders....37

Figure 3.1. Gene regulatory landscape in cortical and dopaminergic neurons....66

Figure 3.2. Analysis overview of the present study....67

Figure 3.3. Heritability enrichment of cigarette smoking and alcohol use traits in dopaminergic and cortical cell types....68

Figure 3.4. Genes and pathways associated with cigarette smoking and alcohol use traits ...69

Figure 3.5. Cellular and brain regional expression profiles of cigarette smoking and alcohol use traits....71

Figure 3.6. Genetic correlation and overlapping genes between cigarette smoking and alcohol use traits....72

Figure 3.7. Pleiotropic genes highlight shared neurobiological bases of cigarette smoking and alcohol use....73

Figure 4.1. Schematic of the protocol....98

Figure 4.2. Number of PD risk genes at different FDR thresholds....99

LIST OF TABLES

Table 2.1. Biological processes enriched for brain disorders 38

Table 3.1. Hi-C libraries used in this chapter.....74

Table 3.2. Source of cis-regulatory elements used in this chapter.....75

Table 4.1. Exonic and promoter coordinates corresponding to steps 4 and 5 of Chapter 4..100

Table 4.2. Exonic and promoter SNPs corresponding to steps 9 and 11 of Chapter 4 101

Table 4.3. Hi-C annotated SNPs corresponding to step 20 of Chapter 4.....102

Table 4.4. Variant-gene annotation file corresponding to step 28 of Chapter 4 103

Table 4.5. Parkinson's disorder output file corresponding to steps 29 and 33 of Chapter 4 104

Table 4.6. Troubleshooting steps associated with Chapter 4 105

LIST OF ABBREVIATIONS

AD	Alzheimer's disease
ADHD	Attention deficiency hyperactivity disorder
ALS	Amyotrophic lateral sclerosis
ASD	Autism spectrum disorder
BD	Bipolar disorder
CN	Cortical neurons
CPD	Cigarettes per day
CRE	Cis-regulatory element
DAR	Differentially accessible regions
DEG	Differentially expressed genes
DLPFC	Dorsolateral prefrontal cortex
DN	Dopaminergic neurons
DPW	Drinks per week
eQTL	Expression quantitative trait loci
FDR	False discovery rate
GO	Gene ontology
GWAS	Genome wide association studies
H-MAGMA	Hi-C coupled MAGMA
LD	Linkage disequilibrium
LDSC	Linkage disequilibrium score regression
iPSC	Induced pluripotent stem cells
MAGMA	Multimarker analysis of genomic annotation

MDD	Major depressive disorder
MS	Multiple sclerosis
NAc	Nucleus accumbens
ND	Nicotine dependence
PAU	Problematic alcohol use
PD	Parkinson's disorder
PFC	Prefrontal cortex
RRHO	Rank rank hypergeometric overlap
SCZ	Schizophrenia
scRNA-seq	Single cell RNA sequence
SN	Substantia nigra
SNP	Single nucleotide polymorphism
SUD	Substance use disorder
TAD	Topologically associating domains
VTA	Ventral tegmental area

CHAPTER 1: GENERAL INTRODUCTION

Brain disorders including psychiatric, and substance use disorders are often characterized by disruptions in a person's cognition, mood, and behavior. According to a national survey, approximately 1 in 5 Americans aged 12 or older suffered from at least one psychiatric or substance use disorder, underscoring the public health significance of understanding their etiology[1]. However, despite their public health and economic burden, treatment options for both lack behind other diseases which can be attributed to several factors including insufficient understanding of their underlying neurobiology[2]. It is therefore pertinent to investigate the neurobiological mechanisms associated with the various psychiatric and substance use disorders to improve insights into novel therapeutic targets.

A growing body of evidence suggests that genetic variations in individuals account for differences in a diagnosis of psychiatric and substance use disorder, suggesting that the genetic contribution to the variability in psychiatric disorders is substantial[3]. However, challenges remain in finding genes that might mediate variations in brain disorders. Genome-wide Association Studies (GWAS) provide an avenue to identify variations such as Single Nucleotide Polymorphisms (SNPs) associated with complex human traits such as psychiatric disorders[4]. Through GWAS, we can identify variations at genetic loci that occur more frequently in individuals with a particular trait compared to individuals who do not exhibit the trait. Since its introduction, GWAS have vastly advanced our understanding of the genetic basis of complex disorders such as Schizophrenia[5] and Bipolar disorders[6]. However, despite this advancement, the biological implications of these variants

are not well characterized because the majority of them reside in noncoding regions of the genome. Indeed, an estimated 90% of SNPs identified through GWAS reside in noncoding regions, including promoter and enhancer regions of the genome with unknown biology, suggesting that they might influence gene regulation[7].

Several techniques have been developed in the field to derive biologically meaningful interpretation of GWAS findings. For example, transcriptome-wide association studies (TWAS) and summary data PrediXcan (S-PrediXcan) were developed to incorporate GWAS and gene expression information to link genetic variants to target genes[8,9]. Similarly, VEGAS was developed to link genetic risk variants to genes of interest based on permutations[10]. In addition to these tools, conventional gene-based analysis from GWAS findings has utilized Multi-marker Analysis of GenoMic Annotation (MAGMA) to assign variants identified from GWAS to their target genes[11]. MAGMA is a bioinformatic tool used to convert single nucleotide polymorphism (SNP)-level P-values identified from GWAS for traits to gene-level P-values in order to identify target genes associated with the trait. MAGMA remains widely used compared to the aforementioned techniques because the tool is user friendly, efficient, and can be run for any trait with available GWAS summary statistics. Despite its practical function in identifying target genes associated with traits, MAGMA relies on positional mapping, typically linking non-coding variants to the nearest genes. However, functional genomic resources consistently point out that the gene regulatory landscape is much more complex than the linear genome. For example, distal regulatory elements may influence gene regulation via forming three-dimensional (3D) structure of the genome (intricate folding of DNA in the nucleus)[12], meaning that it is possible for variants to interact with distal genes, a characteristic that is not factored into conventional MAGMA. This has

necessitated the need for an advanced approach to identify target genes underlying psychiatric disorders to fully understand its genetic and biological components.

Various techniques exist in the field to study the spatial organization of chromatin in cells. Among these include Hi-C genome-wide chromosome capture technique (Hi-C), which detects chromatin interaction in the nuclei[13]. Using Hi-C we can identify regions in the genome that physically interact with each other in a 3D space by measuring the frequency of interaction of their fragments[14,15]. This enabled us to improve the gene-based analysis tool to fully capture genes important to delineating traits. Thus, the overarching goal of this dissertation was to develop a gene-mapping tool based on functional genomics evidence. To extend the capacity of MAGMA, we developed Hi-C coupled MAGMA (H-MAGMA), a novel gene mapping tool that improves on MAGMA, by annotating non-coding SNPs to their target genes based on chromatin structure.

Introduction of H-MAGMA

In Chapter 2, we introduce H-MAGMA and provide a rationale for developing the tool. We then apply H-MAGMA to several brain disorders including Schizophrenia and Alzheimer's which remain the most burdensome brain disorders worldwide[1]. Since the advancement in genomic studies and promise of GWAS, researchers in psychiatric genomics have applied GWAS to several brain disorders and have identified hundreds of loci associated with brain disorders[16]. However, given the importance of non-coding variants, and that they make up a large proportion of GWAS findings, linking these loci to genes to derive biologically relevant information remains a challenge. Therefore in this chapter, we introduce how using a map of chromatin interaction in the brain could be used to detect novel risk genes via H-MAGMA by applying it to five psychiatric disorders (Schizophrenia[5], Autism[17],

ADHD[18], Major Depressive Disorder[19], and Bipolar disorder[6]) and four neurodegenerative disorders (Alzheimer's disease[20], Parkinson's diseases[21], Amyotrophic lateral sclerosis[22], and Multiple sclerosis[23]). We further investigate the biology of the risk genes by identifying their biological, molecular, and cellular components critical to each disorder.

Expanding H-MAGMA to substance use disorders

In Chapter 3 we use H-MAGMA to produce a coherent analysis of genetic variants of cigarette smoking and alcohol use traits. Heritability estimates of substance use ranges from 40-60%, indicating that genetics plays a crucial role in substance use[24]. While GWAS of cigarette smoking and alcohol use have identified genetic variants associated with substance use phenotypes, most of the variants lie in non-coding regions of the genome, making it a challenge to detect their associated genes to decipher how they increase vulnerability of substance use disorder. Additionally, GWAS have identified both genetic variants associated with consumption/use as well as clinical diagnoses of a use disorder. While both informative, consumption/use has been shown to somewhat differ from clinical diagnoses of a use disorder. For instance, a GWAS on both alcohol consumption and alcohol use disorder observed that despite their genetic overlap (genetic correlation = 0.60), alcohol consumption and alcohol use disorder exhibited different trait and disease associations[25]. Therefore, to better characterize the functional impact of genetic variants associated with cigarette smoking and alcohol use traits, it is pertinent to characterize both consumption/use and a use disorder. Thus, Chapter 3 investigates the functional impact of substance use variants. We integrate our H-MAGMA framework to GWAS of heaviness of smoking (measured by the number of cigarette smoked per day [CPD])[26], Nicotine dependence (ND)[27], Problematic alcohol use (PAU)[28], and heavy drinking (measured by the number of drinks per week

[DPW])[26] to identify genes associated with each trait. Additionally, a rich body of research on the neurobiology of substance use has pinpointed the critical role of neurons in understanding substance use vulnerability[29]. Specifically, brain regions including the prefrontal cortex, a region that is associated with higher order cognition and executive functioning has been shown to influence substance use[29,30]. Furthermore, prior research has shown that when exposed to rewarding stimuli such as a substance of abuse, dopaminergic neurons in the midbrain project to other parts of the reward system, resulting in a cascade of processes that might result in developing addiction[31]. Taken together, these findings highlight the important role of cortical and dopaminergic neurons in understanding substance use vulnerability. Given that gene regulatory mechanisms are highly tissue-specific, we therefore use Hi-C datasets from cortical and dopaminergic neurons to delineate the biological impact of genetic variations associated with substance use. We describe the characteristics of cigarette smoking and alcohol use risk genes by (1) identifying known biological functions using gene ontology analysis, (2) identifying specific cell types enriched for each trait, and (3) describing shared biological mechanisms between cigarette smoking and alcohol use.

Expanding H-MAGMA beyond brain cell types

Lastly, Chapter 4 builds upon the successes of Chapters 2 and 3 to expand H-MAGMA beyond brain cell types. In Chapter 2, we introduce H-MAGMA built from bulk tissue from the adult and fetal brains. Chapter 3 extends H-MAGMA to specific brain cell types including cortical and dopaminergic neurons. Given that the current H-MAGMA files are only available for brain cell types, this limits its application to non-brain disorders. Thus, to address this deficiency and build a more comprehensive tool to contribute to the field of genetics as a whole, Chapter 4 expands H-MAGMA to multiple tissue and cell types

including liver, lung, pancreas, and gastric tissue[32]. Additionally, we provide a detailed protocol on how users can develop H-MAGMA for any other tissue or cell types of interest using publicly available datasets.

CHAPTER 2: H-MAGMA FOR IMPROVED PREDICTION OF BRAIN-DISORDER RISK GENES BY INCORPORATING BRAIN CHROMATIN INTERACTION PROFILES[1]

Introduction

Genome-wide association studies (GWAS) have provided insight into the genetic etiology of multiple brain disorders. However, extracting biological mechanisms from GWAS data is a challenge, which is largely because most common risk variants reside in noncoding regions of the genome[33].

MAGMA was initially developed to extract biological insights from GWAS by linking risk variants to their cognate genes[11]. It aggregates single nucleotide polymorphism (SNP) associations to gene-level associations while correcting for confounding factors such as gene length, minor allele frequency and gene density[11]. While MAGMA is a powerful tool and is broadly used, there is room for improvement. MAGMA assigns SNPs to the nearest genes, which has two major pitfalls. First, it is becoming increasingly recognized that noncoding SNPs can regulate distal genes via long-range (>10 kb) regulatory interactions, whereby distal enhancers are brought into contact with the gene promoter[34,35]. Second, MAGMA does not consider tissue-specific regulatory relationships, whereas disease-risk SNPs are enriched in regulatory elements of the disease-relevant tissue[36,37]. To overcome the limitations in MAGMA, we modified the MAGMA approach to create H-MAGMA to assign noncoding SNPs to their cognate genes based on long-range interactions

[1] Reproduced with permission from Nature Springer. Sey, N.Y.A. *et al.* A computational tool (H-MAGMA) improves prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* (2020) doi:10.1038/s41593-020-0603-0.

in disease-relevant tissues measured by Hi-C. H-MAGMA advances conventional MAGMA (hereafter referred to as cMAGMA) by incorporating relevant functional genomic evidence and allowing developmental-stage-specific and cell-type-specific gene mapping. H-MAGMA also differs from traditional Hi-C-guided gene mapping, as it employs the genome-wide mapping capability of MAGMA. While traditional Hi-C-guided gene mapping restricts its analysis to genome-wide significant (GWS) loci[38], H-MAGMA can leverage signals from subthreshold loci that explain a significant proportion of heritability[39].

H-MAGMA was constructed from four classes of brain-derived Hi-C datasets that include human cortical tissue across two developmental stages (prenatal and postnatal) and two brain cell types (neurons and astrocytes), enabling developmental-specific and cell-type-specific gene mapping. We applied H-MAGMA to five psychiatric disorders (attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), schizophrenia (SCZ), bipolar disorder (BD) and major depressive disorder (MDD)) and four neurodegenerative disorders (amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), Alzheimer's disease (AD) and Parkinson's disease (PD)) to generate gene-level summary statistics (Fig. 2.1). By comparing H-MAGMA with cMAGMA, we found that noncoding SNPs often interact with distal genes, thereby necessitating the use of functional genomic evidence in assigning SNPs to cognate genes. We also found a significant overlap between H-MAGMA and two widely used expression quantitative trait loci (eQTL)-based gene-mapping tools: Bayesian framework for colocalization (coloc) analyses[40] and transcriptome-wide association studies (TWAS)[9]. Gene-level association statistics from H-MAGMA closely resembled genetic relationships among brain disorders, which enabled subsequent

analyses to identify biological pathways, developmental windows and cell types critical for each brain disorder.

Results

H-MAGMA

Since our primary goal was to identify neurobiological mechanisms underlying brain disorders, we leveraged two Hi-C datasets obtained from human brain tissue—one from the developing cortex[35] and the other from the adult dorsolateral prefrontal cortex[34] (DLPFC)—to generate gene–SNP pairs that served as an input file for H-MAGMA (Fig. 2.1a). Exonic and promoter SNPs were directly assigned to their target genes based on their genomic location, while intronic and intergenic SNPs were assigned to their cognate genes based on chromatin interactions (Fig. 2.1a). We also generated a cMAGMA input file that utilized the same set of genes and SNPs as H-MAGMA, whereby all intronic and intergenic SNPs were annotated by positional mapping and with a generous gene definition that included 35-kb upstream and 10-kb downstream of each gene.

A major source of discrepancy between H-MAGMA and cMAGMA was noncoding variants because promoter and exonic SNPs were assigned to the same genes in both frameworks. We therefore tested how often intronic and intergenic SNPs were mapped to the nearest genes as predicted by cMAGMA. We found that only 20% of intronic SNPs and 5% of intergenic SNPs interact with nearest genes based on Hi-C (Fig. 2.1b; Fig 2.2a). Because Hi-C-based gene mapping cannot capture proximal interactions within 10 kb[35], we additionally used an eQTL resource from the human DLPFC[34], from which we found that 56% of intronic SNPs and 76% of intergenic SNPs did not show any association with nearest genes (Fig. 2.2a). The majority of noncoding SNPs associated with nearest genes showed

additional association with distal genes, as 80% of intronic SNPs and 87% of intergenic SNPs showed associations with distal genes (Fig. 2.1b). These results highlight the importance of using functional genomic evidence in assigning noncoding SNPs to genes.

We reasoned that H-MAGMA would provide neurobiologically relevant target genes for GWAS by linking noncoding variants to their cognate genes via brain-derived chromatin interaction profiles. We therefore applied the framework to nine brain GWAS, including five neuropsychiatric disorders and four degenerative disorders (Fig.2.1a). The number of brain-disorder risk genes (false discovery rate (FDR) < 0.05) was comparable between H-MAGMA and cMAGMA (Fig. 2.2b), whereas the number of SNPs assigned per gene was threefold higher for cMAGMA (~244 SNPs per gene) than H-MAGMA (~73 SNPs per gene; Fig. 2.2c). In total, cMAGMA and H-MAGMA linked ~7.4 million and ~4.0 million SNPs to genes, respectively (Fig. 2.2d).

Up to 60% of disorder risk genes were selective to H-MAGMA (genes identified by H-MAGMA but not by cMAGMA), which suggests that gene annotation guided by functional genomics can help identify novel genes and pathways (Fig. 2.2b). H-MAGMA-selective genes were significantly enriched for heritability in all nine brain disorders, thereby demonstrating the increase in power of H-MAGMA (Fig. 2.1c; Fig.2.2e).

Using SCZ GWAS as a representative example, we next compared H-MAGMA with the eQTL-based gene annotation tools coloc and TWAS. Coloc tests whether GWAS SNPs and eQTL in a certain GWS locus share the same causal variant[40], whereas TWAS impute the genotype–expression relationship based on the eQTL association statistics and derives expression–trait associations by correlating the imputed gene expression to the trait[41]. We found that a significant proportion of genes identified by eQTL-based gene mapping were

also detected by H-MAGMA (Fig. 2.1d; 74.9% of coloc genes, Fisher's exact test, odds ratio (OR) = 4.34, 95% confidence intervals (CI) = 3.22–5.90, $P = 1.76 \times 10{-}26$; 72.6% of TWAS genes, Fisher's exact test, OR = 12.14, 95% CI = 10.20–14.49, $P = 3.94 \times 10{-}206$). H-MAGMA detected a much larger number of genes that were associated with SCZ, which explained a significant proportion of heritability (4.7% of SNPs explained 38.93% of heritability, enrichment = 8.23, enrichment $P = 7.17 \times 10{-}53$).

Developmental trajectories of risk genes associated with brain disorders

Since three-dimensional chromatin loops are highly tissue-specific[35], it is important to decide which Hi-C datasets are appropriate to identify target genes for each disorder. To address this, we first measured the heritability enrichment of each disorder using tissue-specific regulatory elements. Consistent with the previous findings[37], psychiatric disorders showed strong enrichment in brain tissues, while degenerative disorders lacked brain-specific enrichment. Within brain tissue, psychiatric disorders showed stronger heritability enrichment in the fetal brain than in the adult brain, which highlights their neurodevelopmental origin (Fig. 2.3a). Fetal enrichment was more robust in neurodevelopmental disorders such as ADHD and ASD than in adult-onset disorders such as BD, SCZ and MDD.

To confirm that this result was based entirely on regulatory enrichment, we used an alternative gene-centric approach. Genes associated with each brain disorder were identified based on fetal and adult brain H-MAGMA, and their expression values were compared between prenatal and postnatal stages. There was a clear distinction between psychiatric and degenerative disorders. Genes associated with psychiatric disorders were highly expressed during prenatal stages, while genes associated with degenerative disorders were highly

expressed in postnatal brains (Fig. 2.3b, c). The only exception was MS, which displayed prenatal enrichment. This distinction between psychiatric and degenerative disorders was less clear in cMAGMA: ASD-associated and BD-associated genes were postnatally enriched, whereas AD-associated genes did not display postnatal enrichment and ALS-associated genes were prenatally enriched (Fig. 2.4).

Next, we plotted the developmental expression trajectories of brain-disorder risk genes (Fig. 2.3b, c). Genes associated with ASD, SCZ and MDD showed remarkably similar expression patterns, with a peak at developmental stage 5 (16–19 post-conception week (PCW)). BD-associated and ADHD-associated genes gradually increased during the prenatal stage, with a peak at developmental stage 6 (19–22 PCW). Developmental stages 5 and 6 represent mid-gestation, the period during which upper layer neurons are generated and neuronal differentiation, including axonogenesis and dendritic arborization, takes place[42,43]. This result highlights mid-gestation as a critical window during neurodevelopment that may confer risk to multiple psychiatric disorders, which is consistent with recent results from cross-disorder GWAS[44,45]. Conversely, degenerative disorders showed distinct expression trajectories. Genes associated with degenerative disorders, except MS, constantly and gradually increased during both prenatal and postnatal stages, which suggests that these genes may become more susceptible to damage with aging. This result suggests that there is a strong neurodevelopmental predisposition for psychiatric disorders, which contrasts with degenerative disorders, which have a postnatal origin.

Pathways implicated in brain disorders

To identify biological pathways underlying psychiatric- and degenerative-disorder risk, we conducted a gene ontology (GO) analysis on gene-level association statistics from

H-MAGMA. We ranked genes based on *Z*-scores so that genes with higher *Z*-scores (more significantly associated with a given disorder) are located at the top of the list. We then tested whether a given gene set is overrepresented at the top of the list by performing an incremental enrichment analysis. This approach allowed us to identify biological pathways associated with a given trait regardless of the power of GWAS and to characterize the biological pathways reflecting the gene set as a whole rather than using arbitrarily defined genes with a specific *P* value threshold.

All brain disorders showed enrichment for pathways involved in transcriptional and translational regulation (for example, transcriptional regulators, RNA splicing, and DNA damage and repair pathways; Table 2.1). This is in line with a previous finding that transcriptional dysregulation may mediate the risk for developing brain disorders[46]. Neuronal differentiation and neuronal apoptotic pathways were also enriched in all brain disorders. Neurogenesis was enriched in the majority of disorders except ASD and BD, which is consistent with an increasing number of studies elucidating the role of neurogenesis, differentiation and neuronal apoptosis in brain disorders[47,48]. Unsurprisingly, neurotransmitter and synaptic pathways were implicated in multiple brain disorders, which supports decades of studies highlighting the importance of synaptic function in psychiatric disorders[49].

There were interesting distinctions among brain disorders. For example, all brain disorders showed postsynaptic associations, while a selected set of disorders (ADHD, SCZ, MDD and MS) also exhibited presynaptic associations. Furthermore, while the majority of brain disorders displayed enrichment in glutamatergic signaling, ASD, SCZ and ALS displayed enrichment in GABAergic signaling. ASD-associated genes were enriched for

acetylcholinergic and serotonergic signaling, which reflects the known biology of ASD[50,51]. SCZ-associated and BD-associated genes were also enriched for acetylcholinergic signaling, which supports previous studies reporting that altered cholinergic signaling contributes to SCZ and BD pathogenesis[35,52]. MS-associated genes were enriched for dopaminergic signaling, the disruption of which has been associated with immune malfunction in MS[53]. These results collectively highlight synaptic dysfunction in brain disorders, albeit we also detected distinctions among disorders based on neurotransmitters and presynaptic and postsynaptic associations.

We observed pronounced immune-related processes for degenerative disorders in contrast to psychiatric disorders. In support of this finding, multiple aspects of glial cell development were also associated with brain disorders, with stronger enrichment in degenerative disorders (Table 2.1). Moreover, all degenerative disorders showed associations with genes involved in myelination and oligodendrocyte function, which suggests that there is a potential role of oligodendrocytes in neurodegeneration. In line with this, single-cell transcriptomic profiles in AD postmortem brains suggested that oligodendrocytes have altered molecular profiles[54]. Together with heritability enrichment, this finding of enriched immune response in degenerative, but less so in psychiatric disorders, hints at a possible explanation for genetic distinctions between psychiatric and degenerative disorders[16].

Additional interesting findings include amyloid-β enrichment for AD and PD, and tau enrichment for MS and PD (Table 2.1), which supports the importance of amyloid-β and tau pathology in degenerative disorders[55,56]. We also observed Wnt/β-catenin pathway enrichment for a number of brain disorders, including ASD, SCZ, MDD, PD, MS and ALS. Wnt/β-catenin signaling is a key pathway for neurogenesis and cortical pattern specification,

and its dysregulation has been observed in several psychiatric disorders[57]. Notably, genes involved in vocalization were associated with ASD, the diagnostic criteria of which include impairment in vocalization[58]. We also identified brain regions (for example, the cortex, the hippocampus, the substantia nigra and the hypothalamus) associated with multiple brain disorders. This is intriguing, as we used cortical Hi-C data.

Cell-type specificity

Brain disorders often exhibit different cellular signatures and vulnerability, which highlights the need to identify critical cell types for brain disorders to develop proper therapeutic strategies. For example, ASD postmortem brains exhibit cell-type-specific gene expression signatures such as upregulation of glial genes and downregulation of neuronal genes[59]. Meanwhile, common variation in SCZ maps onto specific groups of cells, including pyramidal neurons and medium spiny neurons[60]. Finally, microglia are increasingly being recognized as a central cell type contributing to the etiology of AD[61].

To identify central cell types that mediate the risk for brain disorders, we next assessed cell-type-specific expression profiles of brain-disorder risk genes. One striking difference between psychiatric and degenerative disorders was that psychiatric-disorder-associated genes coalesced in neurons, while degenerative-disorder-associated genes were highly expressed in glia (microglia for AD and MS, astrocytes for ALS and PD). Since psychiatric disorders showed a neurodevelopmental origin, we also measured cell-type-specific expression profiles of psychiatric-disorder-associated genes in the developing cortex and found convergence onto outer radial glia and excitatory neurons. This selective enrichment in excitatory neurons prevailed across development, as adult neuronal expression

profiles for psychiatric-disorder-associated genes also indicated excitatory neuronal enrichment.

While cMAGMA gave a similar result to H-MAGMA, there were important discrepancies, which included astrocytic expression of ASD-associated genes, lack of astrocytic expression of PD-associated and ALS-associated genes and lack of endothelial expression of MS-associated genes. Given the growing evidence of astrocyte-mediated neurodegeneration in ALS and PD[62,63] the emerging role of the blood–brain barrier in MS[64] and the lack of genetic association signals of an astrocytic co-expression network in ASD[65] this result indicates that H-MAGMA can provide cellular etiology that can be missed by cMAGMA.

Cell-type specific gene mapping

As we detected a remarkable cellular specificity for both psychiatric and degenerative disorders, we next sought to identify disorder risk genes in a cell-type-specific manner. To this end, we built an H-MAGMA framework based on Hi-C interactions from induced pluripotent stem cell (iPSC)-derived neurons and astrocytes[66]. Neuronal and astrocytic H-MAGMA data were subsequently used to decode psychiatric- and degenerative-disorder GWAS, respectively (Fig. 2.5a). We found that a significant proportion of genes (20–40%) were detected in a cell-type-specific fashion.

Cell-type-specific H-MAGMA recapitulates biological processes, cell-type specificities and developmental trajectories of brain homogenate H-MAGMA. For example, brain-disorder risk genes derived from cell-type-specific H-MAGMA were involved in transcriptional regulation, neurogenesis and synaptic transmission. Meanwhile, degenerative-

disorder risk genes showed pronounced enrichment for glial development and inflammatory responses.

Cell-type-specific H-MAGMA further recapitulated cellular expression profiles of disease risk genes. For example, we observed excitatory neuronal expression of psychiatric-disorder risk genes, microglial expression of AD-associated and MS-associated genes, and astrocytic expression of PD-associated and ALS-associated genes (Fig. 2.5a). As astrocytes gain inflammatory profiles with aging[67], we further assessed age-associated astrocytic expression of degenerative-disorder risk genes derived from astrocytic H-MAGMA. We found that AD-associated and PD-associated genes were expressed in mature astrocytes, while ALS-associated genes were highly expressed in fetal astrocytes. MS-associated genes were highly expressed in glioblastoma, which is consistent with the emerging view that astrocyte-mediated neuroinflammation is a key contributor to the pathogenesis of MS[68]. Furthermore, psychiatric-disorder-associated genes showed prenatal enrichment with a peak during mid-gestation, while degenerative-disorder-associated genes were postnatally enriched with a gradual increase in expression across a lifespan (Fig. 2.5b, c). A remarkable difference between cell-type-specific and brain homogenate H-MAGMA was the postnatal expression of MS-associated genes from astrocytic H-MAGMA, which was not detected in brain homogenate.

Shared genetic architecture among brain disorders

We next assessed whether the gene-level association statistics obtained from H-MAGMA can be used to elucidate shared genetic architecture among brain disorders. Since the number of genes significantly associated with a given disorder differs based on the sample size and power of GWAS, we used a rank–rank hypergeometric test of overlap

(RRHO), which is a threshold-free algorithm for comparing two genomic datasets[69]. Genes were ranked based on *Z*-scores from the H-MAGMA output and ranked lists between two disorders were compared to identify the gene-level overlap between them. We then compared this gene-level overlap with genetic correlations calculated by linkage disequilibrium (LD) score regression (LDSC)[70].

Gene-level overlaps recapitulated the previously reported genetic architecture of brain disorders[16] such that psychiatric disorders exhibited strong overlaps in their ranked gene lists, whereas degenerative disorders did not display significant overlaps. Among psychiatric disorders, neurodevelopmental disorders (ADHD and ASD) and adult-onset psychiatric disorders (BD, SCZ and MDD) showed strong overlaps, which indicates that these disorders share neurobiological bases. The correlation between RRHO and genetic correlation was 0.79 ($P = 8.08 \times 10–9$), which demonstrates that gene-level association statistics from H-MAGMA reflect shared genetic architecture and can therefore be further used to decipher the biological mechanisms underlying shared genetic architecture among psychiatric disorders.

Biological pathways underlying pleiotropy

Cross-disorder GWAS of eight psychiatric disorders recently identified more than 100 GWS loci that increase the risk for multiple disorders, which provides further evidence of widespread pleiotropy among psychiatric disorders[44]. Shared genetic etiology across psychiatric disorders may underlie concerted developmental expression trajectories and cellular expression profiles of psychiatric-disorder-associated genes (Figs. 2.3 and 2.5). Therefore, we examined genes shared in multiple psychiatric disorders ($n \geq 4$) to identify common molecular mechanisms of psychiatric disorders. In total, we found 1,841 genes (hereby referred to as pleiotropic genes) that are shared in more than four psychiatric

disorders. Notably, pleiotropic genes showed higher enrichment for genes mapped to pleiotropic cross-disorder GWS loci than those mapped to non-pleiotropic (disease-specific) GWS loci[44] (Fig. 2.6a).

Pleiotropic genes were involved in gene regulation, synaptic function and neuronal and dendritic development (Fig. 2.6b). They showed a distinct peak at mid-gestation, which is consistent with the overall developmental expression patterns of psychiatric-disorder-associated genes (Fig. 2.6c). Finally, pleiotropic genes showed strong excitatory neuronal enrichment for cortical projection neurons in cortical layers 2/3 (excitatory neuronal subtypes 1) and corticothalamic projection neurons in cortical layers 5/6 (excitatory neuronal subtype 7) (Fig. 2.6d).

Discussion

We introduce H-MAGMA, a novel gene mapping tool that builds on MAGMA to annotate non-coding variants to their target genes based on chromatin interaction. To examine the interrelationship between Hi-C and eQTL, we compared H-MAGMA-derived outputs with two eQTL-based gene-mapping tools, coloc and TWAS. Consistent with previous findings[34,41], we detected a substantial overlap. While eQTL-based gene mapping is undoubtedly a powerful approach, H-MAGMA can provide a complementary platform to understand the mechanism of GWAS for the following reasons. First, Hi-C can provide comprehensive genome-wide maps for tissues or cell types with limited access. One example is Hi-C datasets from iPSC-derived neurons and astrocytes that allow GWAS annotation in a cell-type-specific manner[66], which is currently not available with eQTL. Second, it has been recently shown that the variants associated with chromatin accessibility capture stimulus-sensitive signals and explain a significant proportion of heritability, even more so than

eQTL[71,72]. Supporting this claim, we found that H-MAGMA-derived genes explained a significant proportion of heritability in addition to eQTL-derived genes. These results collectively suggest that chromatin architecture such as Hi-C and chromatin accessibility can provide complementary regulatory phenotypes that may be missed by eQTL. It is of note that H-MAGMA also has shortcomings, as it does not capture gene regulatory mechanisms such as altered RNA splicing or the allelic effect (Hi-C cannot predict whether the SNPs will downregulate or upregulate the cognate genes). Leveraging multiple genomic resources, such as eQTL, spliceQTL, chromatin accessibility (ca) QTL and Hi-C, is therefore critical for annotating and interpreting GWAS.

An application of H-MAGMA to nine brain disorders GWAS enabled a systematic delineation of pathogenic mechanisms of brain disorders. For example, one important question in psychiatry is whether a critical window exists for the treatment of psychiatric disorders. Moreover, there is an ongoing debate regarding whether adult-onset disorders such as SCZ and depression have a neurodevelopmental origin. By comparing prenatal and postnatal expression trajectories, we found that genes associated with psychiatric disorders show remarkable developmental convergence onto mid-gestation, while genes associated with degenerative disorders were gradually increased across the life span, which reflects their increased burden with aging.

Another layer of convergence among psychiatric disorders was hinted at by cellular expression profiles. Psychiatric-disorder-associated genes were selectively expressed in excitatory neurons, while degenerative-disorder-associated genes showed more diverse cellular enrichment profiles. Similar cell-type specificity was reported by an interactome

study[73], which demonstrates the robustness of the result obtained when using an orthogonal approach.

These results demonstrate that the shared genetic basis of psychiatric disorders translates into shared neurobiological mechanisms. To further identify shared neurobiological bases among psychiatric disorders, we defined a set of pleiotropic genes that are associated with more than four psychiatric disorders. Pleiotropic genes were associated with neuronal development and synaptic plasticity, which suggests that inappropriate neuronal activity and regulation may act as key components in the pathogenesis of psychiatric disorders. Pleiotropic genes also displayed mid-gestational and excitatory neuronal enrichment, which summarizes the overall pattern of psychiatric-disorder-associated genes. Importantly, this characteristic was also observed for pleiotropic genes identified by a meta-analysis of eight psychiatric disorders[44].

Altogether, H-MAGMA can help develop neurobiologically relevant hypotheses from GWAS by incorporating higher-order chromatin interactions in a disease-relevant context.

Methods

Hi-C

Fetal brain Hi-C data were obtained from the paracentral cortex of three individuals of gestation week 17–18[35]. Adult brain Hi-C data were obtained from the DLPFC of three individuals (aged 36, 44 and 64 years)[34]. Neuronal and astrocytic Hi-C data were derived from human iPSCs obtained from two individuals (aged 15 and 31 years)[66].

GWAS

We used the following GWAS summary datasets: ADHD: $n = 20{,}183$ cases and 35,191 controls[18]; ASD: $n = 18{,}381$ cases and 27,969 controls[17]; BD: $n = 20{,}352$ cases and 31,538 controls[6]; SCZ: $n = 11{,}260$ and 24,542 controls[5]; MDD: $n = 246{,}363$ cases and 561,190 controls[19]; AD: $n = 71{,}880$ cases and 383,378 controls[20]; PD: $n = 37{,}700$ cases and 1,400,000 controls[21]; MS: $n = 4{,}888$ cases and 10,395 controls[23]; and ALS: $n = 12{,}577$ cases and 23,475 controls[22]. Since we used publicly available GWAS summary statistics, no data points were excluded from analysis, no statistical methods were used to predetermine the sample size, and data collection and analysis were not performed blinded to the conditions of the experiments.

Development of H-MAGMA

Exonic and promoter SNPs were directly assigned to their target genes based on their genomic location using a gene model, Gencode v26 (https://www.gencodegenes.org/human/release_26lift37.html), and a promoter was defined as 2-kb upstream of the transcription start site (TSS) of each gene isoform. Intronic and intergenic SNPs were assigned to their cognate genes based on chromatin interactions with promoters and exons as previously described[34,35]. Briefly, we generated a background Hi-C interaction profile by pooling 9 million imputed SNPs from SCZ GWAS summary statistics[74]. Using this background Hi-C interaction profile, we fit the distribution of Hi-C contacts at each distance from each chromosome using the package fitdistrplus (https://cran.r-project.org/web/packages/fitdistrplus/index.html). Significance for a given Hi-C contact was calculated as the probability of observing a stronger contact under the fitted Weibull distribution matched by chromosome and distance. Hi-C contacts with FDR < 0.01

were selected as significant interactions. Significant Hi-C interacting regions were overlapped with Gencode v26 exon and promoter coordinates to identify exon-based and promoter-based interactions. We used exon-based and promoter-based interactions because our previous study[34] comparing Hi-C data with eQTL demonstrated the gene regulatory potential of exon-level interactions. Hi-C data from brain homogenate (fetal and adult human brain) and brain cells (human iPSC-derived neurons and astrocytes) were used to generate MAGMA input files that describe gene–SNP pairs. Input files can be found in the GitHub repository at https://github.com/thewonlab/H-MAGMA.

Gene annotation for cMAGMA

We generated an input file for cMAGMA that was comparable to H-MAGMA. We used the same gene model (Gencode v26) and SNP list used for H-MAGMA, and allowed a window of 35-kb upstream and 10-kb downstream of each gene as previously described[5,46]. Subsequently, any intronic and nearby intergenic SNPs were assigned to the genes based on positional mapping. This input file can be found in the GitHub repository at https://github.com/thewonlab/H-MAGMA.

Noncoding SNP annotation

We first grouped noncoding SNPs into intronic and intergenic SNPs. Proximal genes were defined by positional mapping as follows: for intronic SNPs, genes in which SNPs are located were defined as proximal genes; for intergenic SNPs, nearest genes were defined as proximal genes. Intronic and intergenic SNPs were then overlapped with the SNPs annotated by Hi-C (Hi-C noncoding SNPs: SNPs that interact with gene promoters and exons) and eQTL (eQTL noncoding SNPs: SNPs that have associations with gene expression). For Hi-C noncoding SNPs, we compared proximal genes with genes that physically interact with the

SNPs. For eQTL noncoding SNPs, we compared proximal genes with e-genes (genes that show eQTL associations). We assessed how often physically interacting genes and/or e-genes for a given SNP contain proximal (nearest) genes and whether SNPs show any interactions or associations with distal (non-nearest) genes (Fig. 2.1b).

Running MAGMA

For both H-MAGMA and cMAGMA, we used the MAGMA analysis pipeline as the default setting as follows:

`magma_v1.07b/magma --bfile g1000_eur -pval <GWAS summary statistics> use=rsid,p ncol=N --gene-annot <MAGMA input annotation file> --out<output file>.` Here, g1000_eur denotes the reference data file for a European ancestry population. This file can be downloaded from https://ctg.cncr.nl/software/magma. Detailed instructions can be found in the GitHub repository at https://github.com/thewonlab.

Comparison between H-MAGMA and cMAGMA

We compared disorder risk genes identified by H-MAGMA with those identified by cMAGMA using the package Vennerable in R. We reported the proportion of H-MAGMA-selective genes by calculating the number of genes only identified by H-MAGMA divided by the total number of genes identified by H-MAGMA. Since H-MAGMA results were available from the fetal and adult brain Hi-C data, we used genes that were significantly associated in either the fetal or the adult dataset using the function 'union' function in R (hereby referred to as union disorder risk genes).

We next obtained SNPs mapped to H-MAGMA-selective genes using H-MAGMA input files from the fetal brain and the adult brain (H-MAGMA SNPs) and the cMAGMA input file (cMAGMA SNPs). We also obtained H-MAGMA-selective SNPs by excluding cMAGMA SNPs from H-MAGMA SNPs to ensure that the heritability enrichment we observed was not due to the exonal and promoter SNPs that are shared between H-MAGMA and cMAGMA. We then measured heritability explained by H-MAGMA SNPs and H-MAGMA-selective SNPs using stratified LDSC with the baseline-LD model (S-LDSC)[37].

Comparison between H-MAGMA and eQTL-based gene mapping algorithms

To compare H-MAGMA with eQTL-based tools, we used previously reported SCZ risk genes obtained through TWAS[75] and coloc[34]. Both TWAS and coloc were performed on SCZ GWAS[5] using the largest eQTL resource obtained from the adult human DLPFC[34]. We restricted our H-MAGMA results to those derived from the adult brain so that we could match the developmental period (adult) and brain region (DLPFC) with the eQTL database. TWAS identified 708 SCZ-associated genes (TWAS SCZ genes) for which imputed expression values correlated with SCZ (FDR < 0.05). Coloc identified 255 SCZ-associated genes (coloc SCZ genes) for which eQTL co-localized with SCZ GWS loci (posterior probability 4 (PP4) > (PP0 + PP1 + PP2 + PP3)).

While H-MAGMA uses the whole genome as the genetic background, coloc and TWAS require a more carefully defined background. Because coloc is a GWS loci-centric approach, e-genes within GWS loci ± 1 Mb were considered as background (3,632 genes). Conversely, TWAS is a genome-wide approach and uses *cis*-heritable genes as background (13,396 genes). We therefore intersected H-MAGMA SCZ association results with coloc and TWAS background, from which 1,576 and 2,801 H-MAGMA SCZ genes (FDR < 0.05) were

selected and compared with coloc and TWAS SCZ genes, respectively. By comparing H-MAGMA SCZ genes and coloc and TWAS SCZ genes, we obtained 3,004 H-MAGMA-selective genes (genes identified by H-MAGMA but not by TWAS and/or coloc). SNPs mapped to H-MAGMA-selective genes were subsequently identified via the H-MAGMA input file from the adult brain (H-MAGMA SNPs). Finally, heritability enrichment of H-MAGMA SNPs was calculated using S-LDSC to demonstrate that H-MAGMA genes without eQTL support still explain a significant proportion of heritability.

Heritability enrichment for tissue-specific regulatory elements

To measure heritability enrichment of nine brain disorder GWAS in active genomic regions in each cell and tissue type, we used S-LDSC[37] with chromHMM-defined chromatin states[36]. Since chromatin profiling has not been performed in all cell or tissue types (for example, DNase hypersensitivity was missing for fetal brains, while chromatin immunoprecipitation sequencing for histone 3 lysine 27 acetylation was not performed in the adult DLPFC), we instead used genomic regions that are active in each cell and tissue type using chromatin states defined by chromHMM[76]. We defined active genomic elements by the regions marked as active TSSs (state 1), flanking active TSSs (state 2), genic enhancers (state 6) and enhancers (state 7), while repressive genomic elements were marked as heterochromatin (state 9), repressed polycomb (state 13), weak repressed polycomb (state 14) and quiescent (state 15) in the core 15-state model (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html). To further assess developmental-stage-specific heritability enrichment in the human brain tissue, we defined fetal active elements (elements that are active in the fetal brain and become repressive in the adult brain) and adult active elements (elements that are repressive in the fetal brain then

become active in the adult brain). The SNP annotation file can be downloaded from the GitHub repository at https://github.com/thewonlab/H-MAGMA. Heritability enrichment values in different cell and tissue types resulting from S-LDSC were then scaled to enable tissue-level comparison of enrichment values.

Gene selection

For assessing developmental expression profiles, cell-type-specific expression profiles and GO enrichment of disorder-associated genes, we used the following strategies to select genes. We restricted our analysis to only protein-coding genes because the majority of genes detected in the spatiotemporal transcriptomic atlas[42] , single-cell expression datasets[77–79] and GO terms were protein-coding genes, and because noncoding genes have much lower expression values compared with protein-coding genes, which can dilute signals. We also excluded genes within the major histocompatibility region due to the complexity of LD, which can override the overall pattern. Finally, we removed genes within chromosome X, as only a subset of GWAS had association statistics available in chromosome X.

Developmental and cellular expression profiles

Analyzing developmental and cell-type-specific expression levels required the selection of significantly associated genes for each disorder. We calculated adjusted *P* values based on the Benjamini and Hochberg procedure using the function p.adjust in R. We then selected genes with two FDR thresholds (FDR < 0.01 for GWAS with >20 GWS hits for SCZ, BD, MDD and AD; FDR < 0.1 for GWAS with <20 GWS hits for ADHD, ASD, PD, MS and ALS) as significantly associated brain-disorder genes.

A spatiotemporal transcriptomic atlas from a previous publication[42] was used to obtain cortical expression profiles across multiple developmental stages. Fourteen

developmental stages were defined as follows: stage 1: 4 PCW ≤ age < 8 PCW; stage 2: 8 PCW ≤ age < 10 PCW; stage 3: 10 PCW ≤ age < 13 PCW; stage 4: 13 PCW ≤ age < 16 PCW; stage 5: 16 PCW ≤ age < 19 PCW; stage 6: 19 PCW ≤ age < 24 PCW; stage 7: 24 PCW ≤ age < birth; stage 8: birth ≤ age < 6 months; stage 9: 6 months ≤ age < 1 year; stage 10: 1 year ≤ age < 6 years; stage 11: 6 years ≤ age < 12 years; stage 12: 12 years ≤ age < 20 years; stage 13: 20 years ≤ age < 60 years; stage 14: age > 60 years.

Log-transformed expression values were centered to the mean expression level per sample using the function `scale(center = T, scale = F)` in R. Genes associated with brain disorders were selected for each brain sample, and their average centered expression values were calculated for each brain sample. To ensure that developmental expression trajectories are not dictated by the developmental stage from which Hi-C data were obtained, we used union disorder risk genes. To further verify the developmental trajectories in a cell-type-specific fashion, we used neuronal Hi-C for psychiatric disorders and astrocytic Hi-C for degenerative disorders. Prenatal versus postnatal expression values were compared using the lm function in R (for example, for a given disorder, lm(expression values ~ stages)).

We also used single-cell transcriptomic data from the adult brain[77,79] and the fetal brain[78] to identify cell-type-specific expression profiles of brain-disorder-associated genes. To measure astrocytic expression profiles across developmental stages, we used transcriptomic data from purified human astrocytes[80]. H-MAGMA results derived from fetal and adult brain Hi-C were used to assess cell-type-specific expression values in the fetal and adult brain, respectively. Furthermore, neuronal H-MAGMA was used to assess cell-type and neuronal-subtype enrichment of psychiatric-disorder risk genes, whereas astrocytic H-MAGMA was used to assess cellular expression profiles and age-associated expression

changes in astrocytes for degenerative disorders. We processed log-transformed expression values per cell or sample using the function `scale(center=T, scale=F)` in R. Average centered expression values of genes associated with brain disorders were calculated for each cell type.

GO analysis

We used the R package gProfileR (https://biit.cs.ut.ee/gprofiler/gost) for running GO analysis as it allows a ranked gene list, which resembles gene set enrichment analysis. Because it does not require a *P* value threshold to select significantly associated genes, it allows comparing GO terms for differently powered GWAS in a non-biased fashion. After ranking genes based on *Z*-scores generated by H-MAGMA, we ran GO analysis using the following command line:

```
gprofiler(<Ranked gene list>, organism="hsapiens",
ordered_query=T, significant=T, max_p_value=0.05,
min_set_size=15, max_set_size=600, min_isect_size=5,
correction_method="fdr", hier_filtering="moderate",
custom_bg=background gene set, include_graph=T,
src_filter="GO").
```

RRHO

We assessed the genetic relationship between two disorders (*r*g) by using genetic correlation analysis of LDSC[70]. To provide similar metrics based on gene-level association statistics, we compared ranks between two datasets (for example, H-MAGMA outcomes from two disorders) using the R package RRHO

(https://www.bioconductor.org/packages/release/bioc/html/RRHO.html) with the following command line:

```
RRHO(<Ranked gene list 1>, <Ranked gene list 2>,
outputdir=<output directory>, alternative="enrichment",
BY=TRUE, log10.ind=TRUE).
```

To compare gene-level overlaps (RRHO output) with genetic correlations (calculated by LDSC), *P* values from RRHO were converted into *Z*-scores using the following command line:

```
Zscore = qnorm(10^(-Pvalues), lower.tail=FALSE).
```

We then compared resulting RRHO *Z*-scores with *r*g values from the genetic correlation analysis using Pearson's correlation. This correlation coefficient provides a metric to compare a genetic relationship between two disorders measured at the SNP level (*r*g) versus the gene level (RRHO *Z*).

Identification of pleiotropic genes

RRHO outputs two gene sets consisting of the most upregulated and downregulated genes, with most upregulated genes referring to a list of genes that are associated with both conditions and most downregulated genes referring to a list of genes that are not associated with both conditions. Therefore, we employed the most upregulated genes as a gene list that is shared between two disorders, hence representing pleiotropic genes. We then generated pleiotropic genes shared in at least four disorders by intersecting the RRHO most upregulated genes between the following disorder pairs (ADHD versus ASD, BD, SCZ or MDD; ASD versus BD, SCZ or MDD; BD versus SCZ or MDD; and SCZ versus MDD). Since

psychiatric-disorder-associated genes showed neurodevelopmental and neuronal enrichment, we used fetal brain and neuronal H-MAGMA results. We merged the gene sets using the union function in R and obtained uniquely identified genes. The code is provided in the GitHub repository at https://github.com/thewonlab/H-MAGMA. In the end, we obtained 1,841 genes that were shared in more than four disorders and we defined them as pleiotropic genes. These genes were compared with the genes mapped to pleiotropic versus non-pleiotropic GWS loci from a meta-analysis of eight psychiatric disorders [44]. We next performed GO, developmental expression and cell-type expression analyses on the pleiotropic genes as described above.

Contributions

H.W. designed the H-MAGMA framework. N.Y.A.S. and H.F. applied cMAGMA and H-MAGMA to nine brain disorders. J.C.M. compared H-MAGMA with cMAGMA. H.F. and W.M. performed LDSC and genetic correlation analyses. N.Y.A.S. conducted developmental trajectories analyses, RRHO and functional characterization of pleiotropic genes. B.H. analyzed astrocyte RNA sequencing data and compared H-MAGMA with eQTL-based tools. P.R., K.J.B. and S.A. contributed the Hi-C data from iPSC-derived neurons and astrocytes. N.Y.A.S. and H.W. wrote the manuscript.

Figures

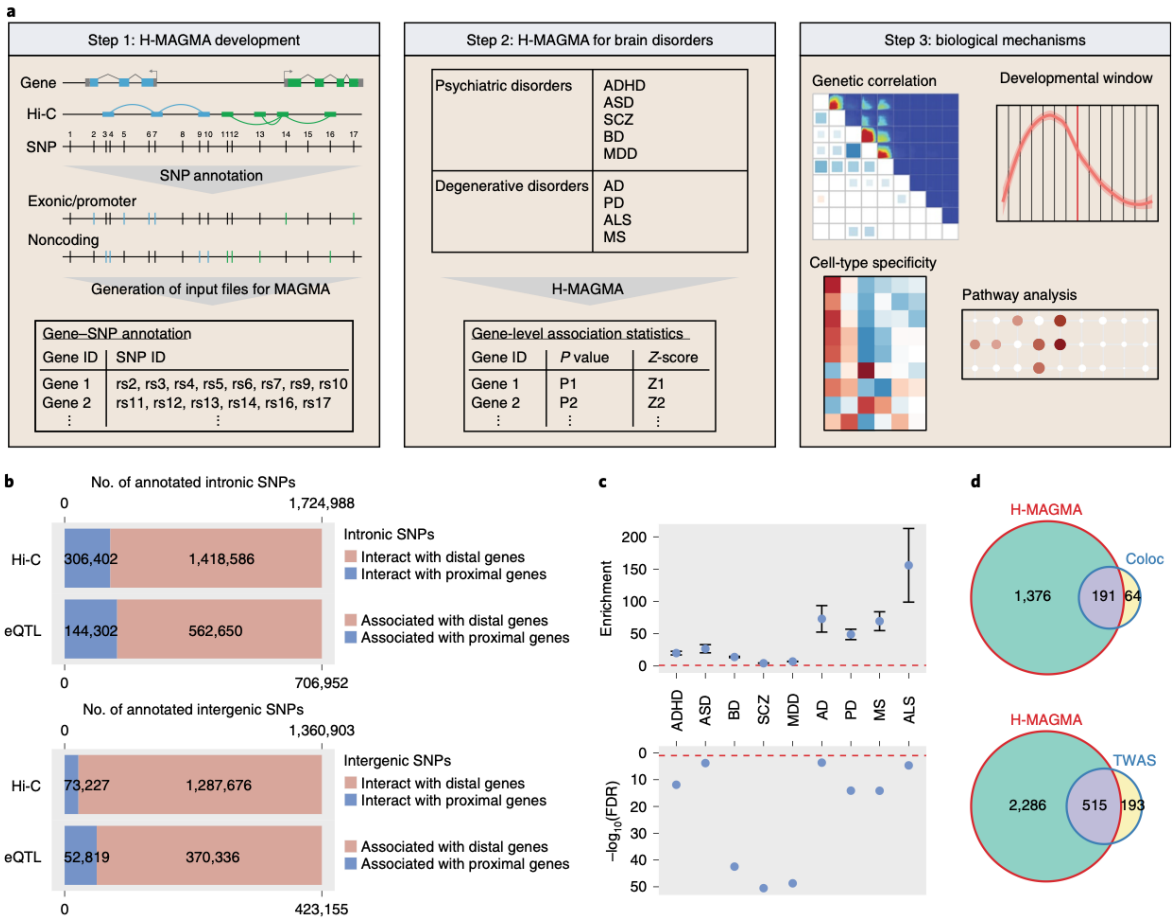
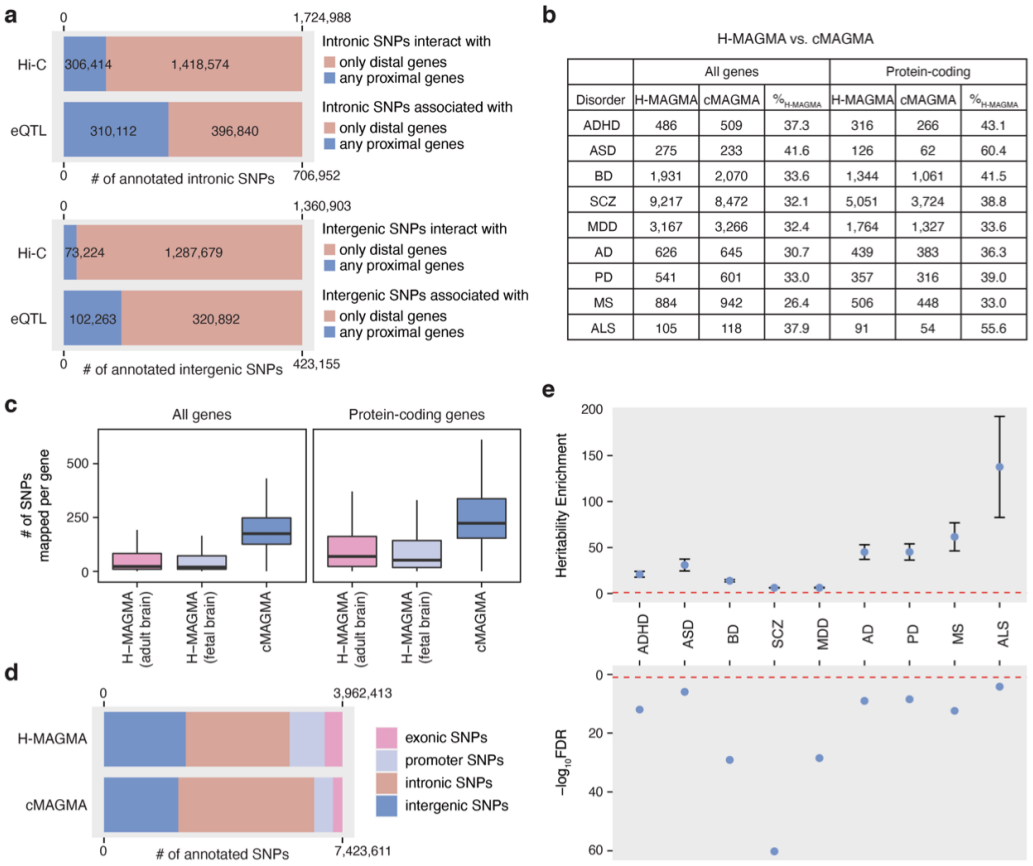


Figure 2.1. Schematics of the H-MAGMA approach. (a) H-MAGMA leverages chromatin interaction profiles (Hi-C) to assign intergenic and intronic SNPs to cognate genes. We applied this framework to five psychiatric disorders and four degenerative disorders using Hi-C datasets from the fetal brain and adult brain. In return, H-MAGMA provides gene-level association statistics, which were used to elucidate biological mechanisms underlying brain disorders. **(b)** Intronic and intergenic SNPs were often annotated to distal genes. **(c)** SNPs mapped to H-MAGMA-selective genes explained a significant proportion of heritability. Top: the heritability enrichment ± standard error, whereby enrichment denotes the proportion of heritability/proportion of SNPs. The red broken line indicates enrichment = 1. Bottom: the FDR of heritability enrichment. The red broken line indicates FDR = 0.05. **(d)** Overlap between SCZ-associated genes identified by H-MAGMA, TWAS and coloc.



	All genes			Protein-coding		
Disorder	H-MAGMA	cMAGMA	%H-MAGMA	H-MAGMA	cMAGMA	%H-MAGMA
ADHD	486	509	37.3	316	266	43.1
ASD	275	233	41.6	126	62	60.4
BD	1,931	2,070	33.6	1,344	1,061	41.5
SCZ	9,217	8,472	32.1	5,051	3,724	38.8
MDD	3,167	3,266	32.4	1,764	1,327	33.6
AD	626	645	30.7	439	383	36.3
PD	541	601	33.0	357	316	39.0
MS	884	942	26.4	506	448	33.0
ALS	105	118	37.9	91	54	55.6

Figure 2.2 Comparison between H-MAGMA and cMAGMA. **(a)** The number and proportion of intronic and intergenic SNPs annotation to proximal and distal genes. SNPs mapped to proximal genes may also have distal associations, while SNPs mapped to distal genes do not have any association with proximal genes. **(b)** The number of brain disorder risk genes (genes that are significantly associated with each brain disorder at a threshold of FDr<0.05) predicted by H-MAGMA and cMAGMA. % H-MAGMA denotes the percentage of H-MAGMA selective genes (genes that were identified by H-MAGMA but not by cMAGMA). **(c)** The number of SNPs assigned to each gene for H-MAGMA and cMAGMA. Center, median; box=1st-3rd quartiles (Q); minima, Q1 - 1.5 x interquartile range (IQr); maxima, Q3 + 1.5 x IQr. **(d)** The number and proportion of SNPs annotated to the cognate genes by H-MAGMA and cMAGMA. **(e)** H-MAGMA selective SNPs (SNPs assigned to H-MAGMA selective genes in H-MAGMA – SNPs assigned to H-MAGMA selective genes in cMAGMA) explain a significant proportion of heritability. Top graph: Heritability enrichment ± standard error; enrichment denotes proportion of heritability/proportion of SNPs; red broken line, enrichment=1. Bottom graph: false discovery rate (FDR) of heritability enrichment: red broken line, FDr=0.05.

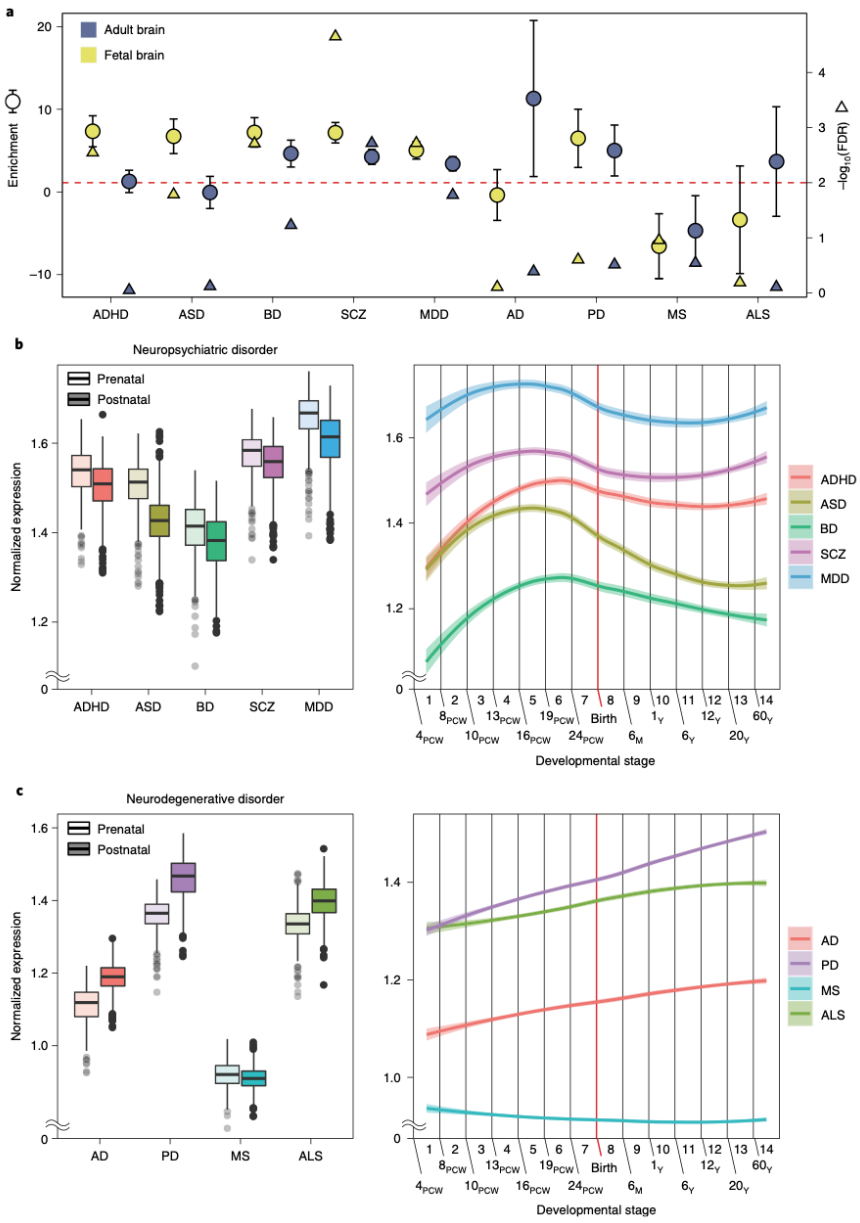


Figure 2.3. Spatiotemporal dynamics of brain-disorder risk genes. (a) Heritability enrichment of brain disorders in active regulatory elements of the fetal brain and adult brain. Enrichment ± standard error (circles) and significance of heritability enrichment (triangles) are depicted. **(b,c)** Developmental expression trajectories of brain-disorder risk genes for neuropsychiatric (b) and neurodegenerative **(c)** disorders. For the boxplots (left panels), n = 410 and 453 for prenatal and postnatal samples, respectively. The center lines represent the median, and the boxes represent the first and third quartiles (Q), whereby the minima is Q1 − 1.5 × the interquartile range (IQr) and the maxima is Q3 + 1.5 × IQr. For the locally estimated scatterplot smoothing (LOESS) plots (right panels), smooth curves are shown with 95% confidence bands. M, month; Y, year.

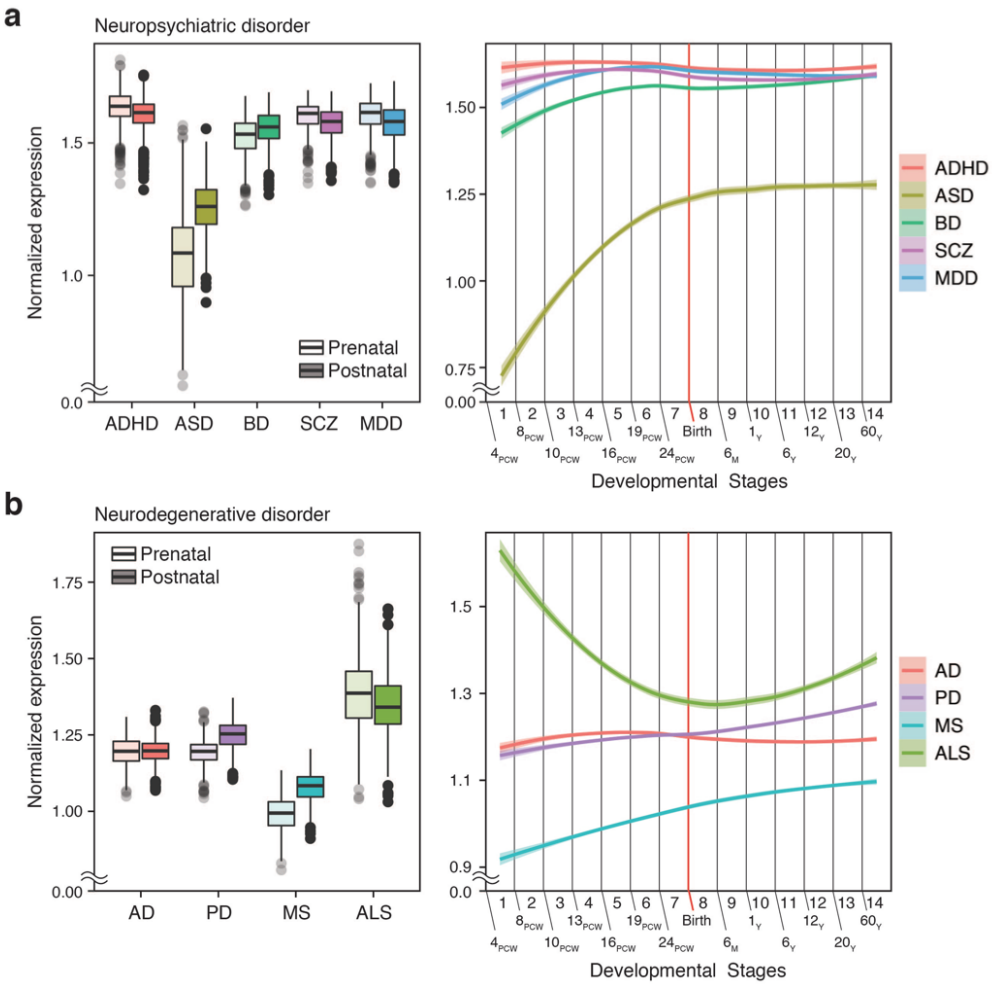


Figure 2.4. Developmental trajectories of brain disorder risk genes derived from cMAGMA. PCW, post-conception week; M, month; Y, year. (Left) N = 410 and 453 for prenatal and postnatal samples, respectively. Center, median; box=Q1-Q3; lower whisker, Q1 - 1.5 x IQr; upper whisker, Q3 + 1.5 x IQr. (right) LOESS smooth curve with 95% confidence bands.

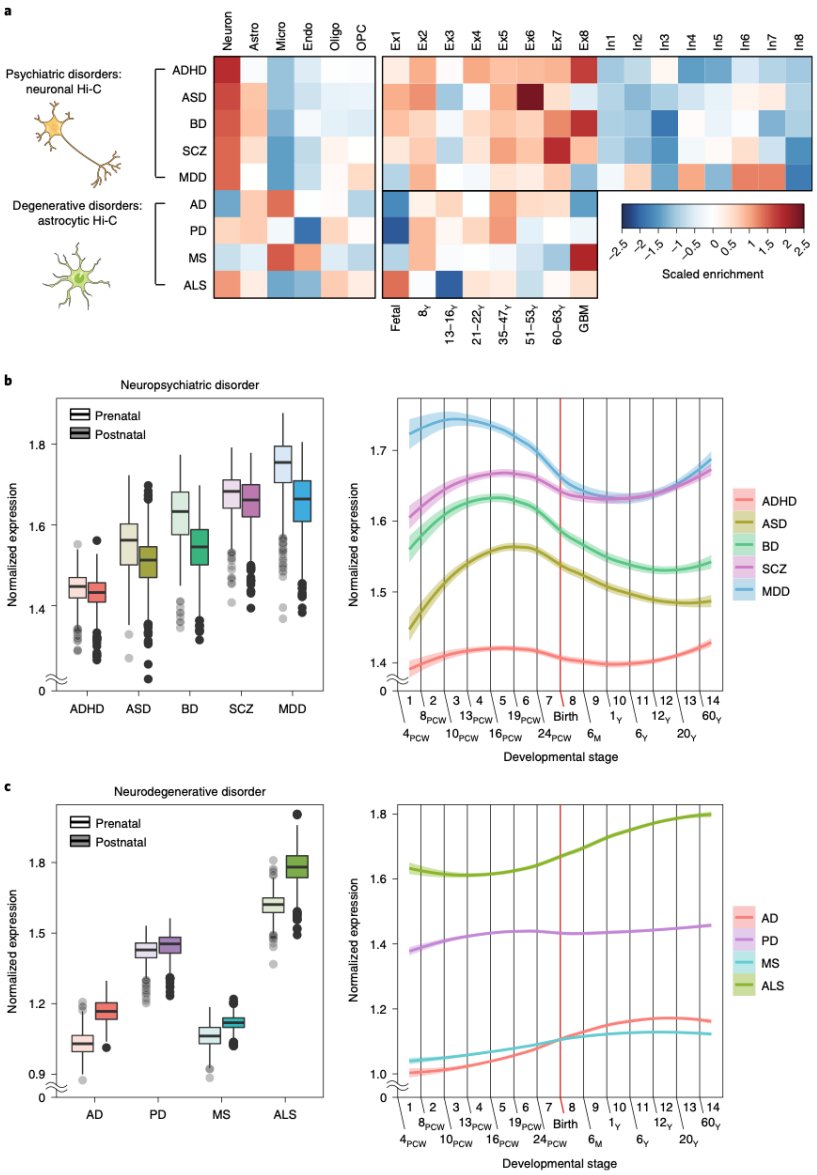


Figure 2.5. Cellular expression profiles of brain-disorder risk genes. (a) We used neuronal and astrocytic H-MAGMA to annotate psychiatric disorder and degenerative disorder GWAS, respectively. Psychiatric-disorder-associated genes are highly expressed in neurons, while neurodegenerative-disorder- associated genes exhibit glial signatures. Astro, astrocytes; endo, endothelial cells; ex, excitatory neurons; GBM, glioblastoma multiforme tumor; in, inhibitory neurons; micro, microglia; oligo, oligodendrocytes; OPC, oligodendrocyte progenitor cells. (**b,c,**) Developmental expression trajectories of psychiatric-disorder- associated genes (b) and degenerative-disorder-associated genes (c). For the boxplots (left panels), $n = 410$ and 453 for prenatal and postnatal samples, respectively. The center lines represent the median, and ranges are as for Fig. 2b,c. LOESS plots show smooth curves with 95% confidence bands.

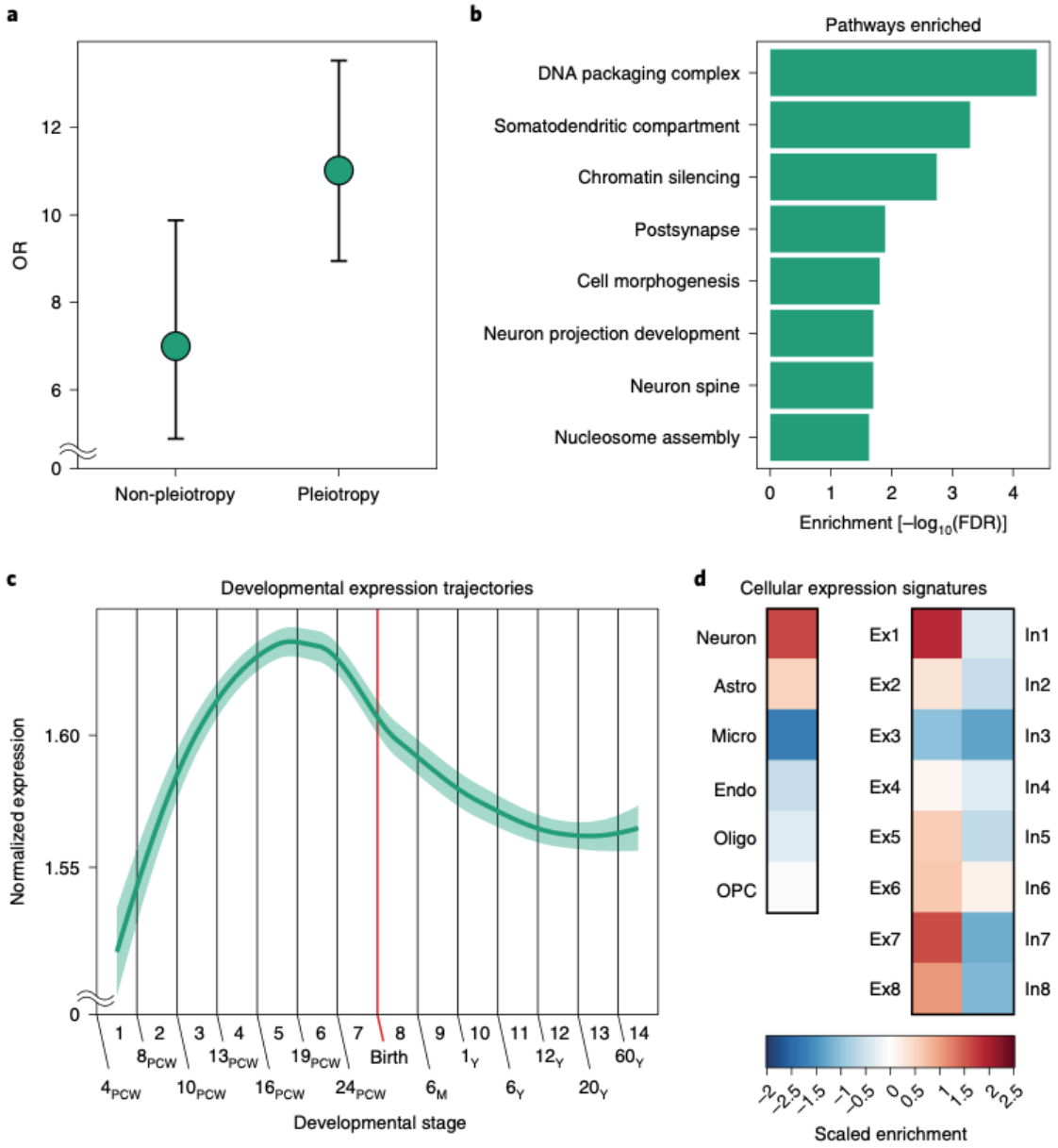


Figure 2.6. Shared molecular mechanisms of psychiatric disorders. (a) Comparison between pleiotropic genes and genes mapped to non-pleiotropic and pleiotropic GWS loci. Or and 95% CI are shown. **(b)** GO enrichment of pleiotropic genes. **(c)** A developmental expression trajectory of pleiotropic genes as shown using LOESS smooth curve with 95% confidence bands. **(d)** Cell-type-specific expression profiles of pleiotropic genes.

Tables

Table 2.1. Biological processes enriched for brain disorders

	ADHD	ASD	BD	SCZ	MDD	AD	PD	MS	ALS
Transcriptional regulators	✓	✓	✓	✓	✓	✓	✓	✓	✓
DNA damage/ repair	✓	✓	✓	✓	✓	✓	✓	-	✓
RNA splicing	✓	✓	✓	✓	✓	✓	✓	✓	✓
Neurogenesis	✓	-	-	✓	✓	✓	✓	✓	✓
Neuronal differentiation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Neuronal apoptosis	✓	✓	✓	✓	✓	✓	✓	✓	✓
Glutamatergic	✓	✓	✓	✓	✓	✓	✓	✓	✓
GABAergic	-	✓	✓	✓	-	-	-	-	✓
Synaptic	**Pre**/Post	Post	**Post**	**Pre**/Post	**Pre**/Post	Post	Post	Pre/**Post**	**Post**
Neurotransmitter	Dopamine, Serotonin, **Acetylcholine, Monoamine, Nitric oxide**	Acetylcholine, Nitric oxide, **Serotonin**	**Acetylcholine, Nitric oxide**	Dopamine, **Acetylcholine, Nitric oxide**	Monoamine	-	**Nitric oxide, Norepinephrine**	**Dopamine**	-
Glial cells/ astrocytes	**Differentiation**	**Development, Gliogenesis**	Gliogenesis	-	Astrocyte development, Glial guided migration	**Differentiation, Migration, Proliferation**	Differentiation, **Proliferation**	Projection, **Migration, Proliferation**	Migration
Oligodendrocytes	-	✓	-	✓	✓	**✓Myelination**	✓Myelination	**✓Myelination**	✓Myelination
Brain development	Cerebral cortex	**Telencephalon**	Hippocampus, **Substantia nigra, Telencephalon**	-	Cerebral cortex, Telencephalon, **Substantia nigra**	Forebrain, **Cerebral cortex**	Forebrain, **Limbic system, Midbrain, Telencephalon**	**Cerebral cortex, Hypothalamus, Forebrain**	**Hindbrain, Forebrain, Substantia nigra**
Interferon	✓	-	✓	✓	✓	✓	✓	✓	✓
Antigen processing	✓	✓	✓	✓	-	✓	✓	✓	✓
Interleukin	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cytokine	-	✓	-	✓	✓	✓	✓	✓	-
T cell	✓	✓	✓	✓	✓	✓	✓	✓	✓
B cell	-	✓	-	✓	-	✓	-	-	-
Toll-like receptor	✓	✓	✓	✓	✓	-	✓	-	✓
Aging	✓	✓	-	-	-	✓	-	-	✓
Ischemia	-	-	-	-	-	-	-	✓	✓
Others	ERBB	Vocalization, Wnt, BAF, APP	Glucocorticoid, **APP, Synaptic plasticity**	**Amyloid-β, BAF, β-catenin/Wnt, ERBB**	β-catenin, **ERBB**	Amyloid-β, ERBB, **Glucocorticoid**	Wnt, **Amyloid-β, ERBB, Tau**	ERBB2, **β-catenin/ Wnt, Tau**	Wnt

GO terms detected in H-MAGMA but not in cMAGMA are marked in bold; dashes indicate that no enrichment was found. APP: β-amyloid precursor.

CHAPTER 3: CHROMATIN ARCHITECTURE IN ADDICTION CIRCUITRY IDENTIFIES RISK GENES AND POTENTIAL BIOLOGICAL MECHANISMS UNDERLYING CIGARETTE SMOKING AND ALCOHOL USE TRAITS[2]

Introduction

The National Survey on Drug Use and Health in 2018 estimated that 27.3 million individuals were daily cigarette smokers, and 16.6 million individuals were heavy alcohol users[81]. Cigarette smoking and alcohol use are the 1st and 3rd leading causes of mortality and morbidity, accounting for 480,000 and 88,000 deaths per year in the United States, respectively[82,83]. Despite their public health burden, treatment options for nicotine and alcohol use disorders are limited. However, existing treatments can be improved, and new treatments can be developed with a better understanding of the underlying neurobiology of addiction. Genome-wide association studies (GWAS) on smoking and alcohol use traits have demonstrated that common variation explains a significant proportion of phenotypic variance of substance use[25]. Nearly 400 genomic loci were found to have an impact on smoking and/or alcohol use traits from GWAS sample sizes of up to 1.2 million[26–28]. However, the vast majority of associated variants reside in non-coding DNA, and their target genes and relevant neurobiological mechanisms are poorly understood. Examining higher-order chromatin architecture is crucial to understanding the functional consequences of non-coding variation by linking variants to distal genes based on chromatin interaction profiles[35,84]. Whereas the

[2] Reproduced with permission from Nature Springer. Sey, Nancy Y A et al. Chromatin architecture in addiction circuitry identifies risk genes and potential biological mechanisms underlying cigarette smoking and alcohol use traits. *Mol. Psyc*. (2022) doi:10.1038/s41380-022-01558-y

three-dimensional (3D) genomic landscape of the human brain has advanced our understanding of neurobiological mechanisms underlying psychiatric disorders[85,86], such approaches have been essentially lacking in explaining the genetic architecture of substance use disorders (SUD).

To understand the functional impact of common variants associated with cigarette smoking and alcohol use, we applied Hi-C coupled MAGMA (H-MAGMA)[86] to GWAS of smoking and alcohol use traits and identified their putative target genes for further characterization[26–28]. Smoking and alcohol use traits likely affect neural circuits that underlie addiction and include the prefrontal cortex (PFC), nucleus accumbens (NAc), amygdala, and midbrain dopaminergic cell groups such as ventral tegmental area (VTA) and substantia nigra (SN)[29,87]. We reasoned that the characterization of chromatin architecture across the brain reward circuitry is critical to understanding the gene regulatory mechanisms associated with substance use. With neurons being the major drivers of substance use behaviors, we profiled chromatin architecture from cortical neurons (CNs) in the dorsolateral PFC (DLPFC)[88] and dopaminergic neurons (DNs) in the midbrain[89]. We then built H-MAGMA inputs from CNs and DNs and applied them to GWAS summary statistics of smoking and alcohol use traits. In particular, given the recent work on a potential difference in genetic architecture between substance consumption and clinical diagnosis of use disorder[25], we mapped genetic variants associated with consumption or use (drinks per week [DPW][26] and cigarettes per week [CPD][26]) versus use disorder (problematic alcohol use [PAU][28] and nicotine dependence [ND][27]) to their associated risk genes. Our analysis of substance use risk genes identified key biological pathways, primary cell types, and brain circuitry that might confer risk for substance use. In addition, we characterized genes and pathways shared

between cigarette smoking and alcohol use traits to provide a core neurobiological basis of addiction.

Results

Epigenetic landscape of cortical and midbrain dopaminergic neurons

Neural circuitry underlying addiction involves, among others, dopaminergic cell groups in the midbrain, including VTA and SN, as well as neuronal populations in the PFC[29] (Fig. 3.1a). However, the gene regulatory landscape in these two brain regions and its implication in the genetics of cigarette smoking and alcohol use traits have not been studied. To understand the relationship between the reward circuitry and genetic underpinnings of substance use, we evaluated enrichment of genetic risk factors for four traits associated with alcohol use (PAU[28] and DPW[26]) and cigarette smoking (ND[27] and CPD[26]) in *cis*-regulatory elements (CREs) of the midbrain and PFC[90] using stratified LD score regression (LDSC)[37]. We defined CREs using publicly available data and annotations derived from epigenomic assays and demonstrated that every trait showed significant heritability enrichment for CREs in the midbrain and PFC (Fig. 3.1b, Fig. 3.2 and Fig. 3.3).

Midbrain DNs have long been hypothesized to be the major player of the brain reward circuitry[29,91]. Thus, we investigated whether midbrain DN-CREs explained the heritability enrichment of cigarette smoking and alcohol use traits. Indeed, genetic risk factors for substance use traits were enriched in chromatin accessible regions of DNs derived from human induced pluripotent stem cells[92] (hiPSC, Fig. 3.3a). Given the cellular heterogeneity of the PFC, we also evaluated heritability enrichment of substance use traits in CREs of four major cell types (neurons, astrocytes, microglia, and oligodendrocytes) in the cortex[93]. Neurons showed the strongest heritability enrichment of substance use traits among

the four cell types (Fig. 3.3b). These results collectively suggest that the gene regulatory relationships in CNs and DNs may provide rich information about genetic underpinnings of substance use traits.

We next sought to compare gene regulatory relationships between CNs and DNs. Substantial differences in chromatin architecture have been observed across different cell types in the human brain[66,88,93], but less information is available for the chromatin architecture in different brain regions and/or neuronal subtypes. To interrogate differences in chromatin architecture between CNs and DNs, we identified differential chromatin accessibility peaks between CNs and DNs[92] using DiffBind[94]. Differentially accessible regions (DARs) in CNs (CN-DARs) were then mapped to their target genes based on CN Hi-C data recently generated by our group[88] (Fig. 3.1c). Since DN Hi-C data with comparable read depths were not available, we generated high-resolution chromosome conformation maps from the midbrain DNs to link DN-DARs to the corresponding genes (Fig. 3.1c and Fig. 3.2). To examine the correlation between neuronal subtype-specific chromatin architecture and gene expression signatures, we then measured cell-type specific expression profiles of the genes linked to CN- and DN-DARs. Genes linked to CN-DARs were highly expressed in cortical pyramidal neurons of the telencephalon (GLU1–3, 6–8), whereas genes linked to DN-DARs were highly expressed in midbrain dopaminergic (DOP2) and cholinergic neurons (CHO1) as well as subcortical-projection glutamatergic neurons in the telencephalon (GLU5, 13–17) [95] (Fig. 3.1d).

We also found evidence of different enhancer wiring between CNs and DNs. For example, *FOXA2* and *NR4A2*, master regulators for dopaminergic neuronal specification and differentiation[96–98], displayed different regulatory connections between CNs and DNs.

FOXA2 was linked to two proximal enhancers in DNs as compared to one distal enhancer in CNs (Fig. 3.1e). *NR4A2* was linked to multiple distal enhancers only in DNs, but not in CNs (Fig. 3.1f). On the contrary, genes that encode synaptic scaffolding proteins at glutamatergic synapses (e.g. *SHANK3* and *DLGAP2*) displayed enhancer-promoter connectivity only in CNs, but not in DNs (Fig. 3.1g-h).

We next compared topologically associating domains (TADs) between CNs and DNs. Consistent with previous reports[99], TADs were largely conserved between CNs and DNs. However, we noted some differences in TAD boundary strengths (defined by *binSignal*) between CNs and DNs. For example, *EN1 is a* critical survival factor for DN differentiation and maintenance[100]. We found that *EN1* was located at the TAD boundary whose strength is stronger in DNs than in CNs. *FOXA2* also showed strengthened TAD boundaries in DNs, which corresponds to the confinement of loops in proximal space as evidenced in Fig. 3.1e**.** Importantly, these genes were more highly expressed in DNs than in CNs. On the contrary, *CREBBP* and *WNT3A*, genes with elevated expression in CNs, were located in TADs that were enlarged in CNs. Therefore, these results indicate that different neuronal subtypes involved in substance use traits display distinct chromatin architecture that is coupled with transcriptional regulation.

CN and DN H-MAGMA identifies genes and biological pathways underlying cigarette smoking and alcohol use traits

To investigate the functional impact of common variants associated with cigarette smoking and alcohol use traits, we next employed H-MAGMA to assign genetic variants to their target genes based on long-range chromatin interaction[86]. Heritability enrichment results suggested roles for CNs and DNs in cigarette smoking and alcohol use traits (Fig. 3.1b and

Fig. 3.3). Therefore, we generated H-MAGMA input files from CN Hi-C data previously reported in Hu et al.[88] and newly generated Hi-C libraries from the midbrain DNs (hereafter referred to as CN and DN H-MAGMA, respectively). We applied H-MAGMA to PAU, DPW, ND, and CPD, and identified risk genes for each trait using a false discovery rate (*FDR*) threshold of 5% (Fig. 3.4a-b). We detected a small number of risk genes for ND in comparison to other GWAS, which can be attributed to the smaller sample size of the ND GWAS.

Both CN and DN H-MAGMA identified *CHRNA3*, whose role in increased risk of smoking[27] has been well established, to be associated with CPD (Fig. 3.4c). Similarly, *ADH1B*, a gene that encodes alcohol metabolizing enzymes, was found to be associated with DPW via both CN and DN H-MAGMA, illustrating H-MAGMA's ability (Fig. 3.4d**)**. Additionally, some of the risk genes identified by CN and DN H-MAGMA have been directly characterized using rodent models of substance use. For example, *DRD2*, a gene that encodes a dopaminergic receptor, was identified by CN and DN H-MAGMA to be associated with all traits. Deletion of its ortholog in mice altered alcohol preferences and alcohol-induced ataxia[101]. Moreover, individual variability in *DRD2* was found to be associated with response to nicotine replacement therapies[102]. We also identified *GABRG1* association with DPW from DN H-MAGMA. In a previous study investigating the role of GABA receptors in alcohol consumption, *Gabrg1* knockout mice exhibited decreased alcohol consumption during an operant conditioning testing and home cage drinking assessment[103]. Lastly, *BDNF* was identified by CN and DN H-MAGMA to be associated with PAU and DPW. Downregulation of this gene in the dorsal striatum has been shown to increase alcohol consumption in rats[104].

While differential gene expression has been detected in the postmortem brain samples of substance abusers compared with controls[105,106], the extent to which this expression signature is predisposed by genetic risk factors remains unclear. To address this, we compared H-MAGMA gene-level scores to differentially expressed genes from human brain tissue and human iPSC-derived neural cultures after exposure to nicotine and alcohol, respectively. In this comparison, DN H-MAGMA results for DPW showed a significant association with gene expression signatures from hiPSC-derived forebrain neural cells exposed to alcohol (DEG logFC ~ H-MAGMA Z-score + number of SNPs, $\beta=2.31\times10^{-2}$, $p<2.0\times10^{-16}$)[107]. In parallel, we identified a significant association between CN H-MAGMA results for ND and gene expression signatures from the adult prefrontal cortex of active smokers (DEG logFC ~ H-MAGMA Z-score + number of SNPs, $\beta=3.10\times10^{-3}$, $p<2.0\times10^{-16}$)[106]. Among the genes that were differentially expressed in ND was *CHRNB4*, a gene that encodes a nicotinic acetylcholine receptor subunit. Together, these results suggest that genetic risk factors may contribute to the gene expression signatures of cigarette smoking and alcohol use traits.

Next, we mapped risk genes identified from CN and DN H-MAGMA to biological pathways using gene ontology (GO) analysis. Rather than using a specific *FDR* threshold, we ran ranked-based GO analysis using the Z-score of H-MAGMA output files. Since we used two separate H-MAGMA inputs to assign common variants to their target genes, we obtained two GO results for each trait – one for CN H-MAGMA risk genes and the other for DN H-MAGMA risk genes. We then classified GO terms as CN-specific or DN-specific if they represented biological pathways unique to CN or DN H-MAGMA, respectively.

We validated previous findings that ethanol metabolic processes and response to alcohol were associated with PAU and DPW (Fig. 3.4e-f)[25,28], and that cholinergic and nicotinic pathways were associated with ND and CPD[26,27] (Fig. 3.4g-h). Notably, we also identified alcohol catabolic processes for ND and nicotinic pathways for PAU. Likewise, we further identified GO terms relating to other substances of abuse. For instance, GO terms for PAU included response to morphine (Fig. 3.4e), while GO terms for CPD included response to cocaine (Fig. 3.4h). Taken together, these findings underscore potential genetic overlap and interplay among different substances of abuse.

We identified several similarities across cigarette smoking and alcohol use traits. For example, neuronal processes such as neuronal migration and apoptosis were associated with cigarette smoking and alcohol use, which is in line with studies that have pinpointed the disruption of neuronal migration and neurotransmission in response to substance use[108,109]. We also observed myelination and gliogenesis to be associated with DPW and CPD, respectively, hinting at the role of neuron-glia interactions in substance use traits. Several immune processes including T and B cell activation were shown to be associated with cigarette smoking and alcohol use, which corroborates the relationship between substance use and suppressed immunity (Fig. 3.4g)[110,111]. We also identified a potential role of protein folding that has been shown to contribute to the stress response[112]. A potential link between substance use and neurodegeneration emerged, such as amyloid-beta metabolic processes for CPD and tau protein binding for DPW. Lastly, pain perception was associated with DPW and CPD, consistent with prior research linking pain perception and the reward circuitry[113].

We also observed distinct biological processes between cigarette smoking and alcohol use traits. For instance, long term synaptic depression (Fig. 3.4e), as well as learning and

memory (Fig 3.4f), were characteristic of alcohol use traits but not cigarette smoking, highlighting the important role of synaptic plasticity and memory consolidation in the mechanism of alcohol use[114,115]. We also found GO terms relating to sleep and wake cycle for alcohol use traits, which support a rich body of evidence suggesting that prolonged alcohol use and misuse can cause deleterious effects on sleep quality[116]. Cigarette smoking traits also exhibited distinct associations not observed in alcohol use traits. For instance, we noted lung development to be associated with both ND and CPD which supports epidemiological findings of lung morbidities linked to cigarette smoking[117].

Discrete biological processes were also observed between CN and DN H-MAGMA. Ethanol metabolism and alcohol response were enriched for alcohol use traits in a DN-specific manner (Fig. 3.4e-f). In contrast, the potential link between neurodegeneration and substance use was specific to CNs. These results suggest that the neurobiological basis of cigarette smoking and alcohol use traits may need to be studied in a brain region- and neuronal subtype-specific manner.

Cellular expression profiles of cigarette smoking and alcohol use risk genes convey cell types associated with substance use

Since CNs and DNs display heterogeneity and act in synchrony with multiple cell types, we leveraged single-cell RNA sequencing (scRNA-seq) datasets to further refine neuronal subtypes that confer risk of substance use. We first evaluated cellular expression profiles of cigarette smoking and alcohol use risk genes identified from CN H-MAGMA using scRNA-seq data from the human cortex[77]. We not only recapitulated our findings that genetic risk variants underlying cigarette smoking and alcohol use are highly expressed in

neurons, but also observed that the risk genes were highly expressed in excitatory neurons (Fig. 3.5a). Specifically, we found PAU, DPW, and CPD risk genes to be highly expressed in layer 5 pyramidal neurons (Ex5) that project to both cortical and subcortical areas including the striatum and the midbrain and layer 4 neurons (Ex2) that receive sensory signals from the thalamus, a region that has been shown to be integral to addiction by modulating arousal and motivation[118].

Comparably, we examined expression profiles of cigarette smoking and alcohol use risk genes identified from DN H-MAGMA in midbrain cell types using scRNA-seq data from the human embryonic ventral midbrain[119]. Risk genes for all traits except for ND were highly expressed in DNs, providing additional evidence to support the impact of DNs in modulating substance use via the reward-circuitry[29] (Fig. 3.5a). Within the DN lineage, we found elevated expression of risk genes in intermediate DNs (DA1) for all traits, suggesting that they may be more vulnerable to substance use. Moreover, risk genes showed elevated expression in midbrain GABAergic neurons which have been shown to regulate a diverse set of processes including motor control and inhibition of dopaminergic cells, thereby modulating the reward-circuitry[120]. Similarly, risk genes' expression in serotonergic neurons is consistent with their reported involvement in substance use vulnerability[121].

Next, we extended our approach to a brain-wide fashion by assessing brain regional expression profiles of cigarette smoking and alcohol use risk genes. We leveraged extensive scRNA-seq data from the mouse nervous system to determine brain regions with high expression values of risk genes identified by CN and DN H-MAGMA[95]. Both cigarette smoking and alcohol use risk genes were highly expressed in cortical and midbrain regions as

expected (Fig. 3.5b). We also found strong expression in the hippocampus for ND risk genes, highlighting the role of hippocampus-dependent learning in ND[122]. Furthermore, thalamic expression was observed for PAU, DPW, and CPD risk genes, which is consistent with high levels of expression in Ex2 that receives thalamic inputs (Fig. 3.5a) and points to the role of sensory perception in drug-seeking behaviors[123]. Finally, our results highlight the amygdala for elevated expression of risk genes associated with cigarette smoking and alcohol use. The association of risk variants with the amygdala underscores the role of emotional processing in substance use due to its projections to other parts of the reward-circuitry[29].

Shared genetic architecture among substance use

Individuals often become dependent on multiple substances, and these comorbidities may be driven by shared genetic signals[28,124]. We hypothesized that biological characterization of pleiotropic genes between cigarette smoking and alcohol use traits would identify neurobiological mechanisms underlying the shared genetic architecture of substance use traits.

We first calculated genetic correlations and gene-level overlap across cigarette smoking and alcohol use traits using LDSC and rank-rank hypergeometric overlap (RRHO) test, respectively (Fig. 3.6). We found that RRHO of DN H-MAGMA outputs gives stronger gene-level overlaps than that of CN H-MAGMA. For example, 119 and 3,120 genes were shared between PAU and CPD using CN and DN H-MAGMA, respectively (Figure 3.7a and Fig. 3.6b-c). These results suggest that DNs may play a central role in explaining comorbidity in substance use. Because the PAU and CPD showed a significant genetic correlation (genetic correlation = 0.19) and gene-level overlap (RRHO Z-score = 12.16), we

selected shared genes between PAU and CPD in DN H-MAGMA to serve as pleiotropic genes (Fig. 3.7a). Pleiotropic genes were enriched for synaptic function and cell junction organization (Fig. 3.7b), suggesting that alterations in synaptic organization may influence core features of substance use. We further evaluated cellular expression profiles of pleiotropic genes in the human embryonic ventral midbrain. We again found elevated expression of pleiotropic genes in dopaminergic, GABAergic, and serotonergic neurons in the midbrain, indicating their potential function in substance use biology **(**Fig. 3.7c**)**.

Based on our hypothesis that pleiotropic genes between cigarette smoking and alcohol use traits may represent risk genes shared across multiple SUD, we next examined whether they are dysregulated in response to other substances. We overlapped our pleiotropic genes with differentially expressed genes (DEGs) in the rat NAc after cocaine treatment[125]. We found a significant proportion of our pleiotropic genes was dysregulated in response to cocaine (Fig. 3.7d). We also compared the cellular expression profiles of pleiotropic genes in saline versus cocaine treatment conditions[125]. We found that pleiotropic genes were downregulated in response to cocaine in *Drd1*- and *Drd2*-expressing MSN (Fig. 3.7e). Taken together, these results indicate that pleiotropic genes derived from cigarette smoking and alcohol use traits can provide insights into the core neurobiological mechanism of substance abuse.

Drug repurposing analysis

A fundamental issue facing the treatment of SUD is the limited number of effective medications available. Although medications such as Naltrexone[126] and Nicotine Replacement Therapies (NRT)[127] have been traditionally used to treat alcohol use disorder

and nicotine addiction, respectively, their efficacies are lacking or produce severe adverse outcomes, rendering the need for new treatment. To address this challenge, we used the Drug Signature and Drug Matrix databases of EnrichR[128], a comprehensive gene analysis tool to identify potential drug candidates for SUD based on genetic evidence. We identified several significantly enriched drug candidates for cigarette smoking and alcohol use traits. Among these included mood stabilizers and selective serotonin reuptake inhibitors such as Fluoxetine, Citalopram, and Imipramine, consistent with their potential therapeutic benefits in some patients diagnosed with nicotine or alcohol dependency. We further identified enrichment for antipsychotics such as Chlorpromazine and Clozapine, pointing to some degree of convergence of addiction-relevant risk genes with molecular pathways implicated in other types of psychiatric illnesses. These findings speak to the well-documented epidemiological[129,130] and genetic[131,132] evidence supporting the comorbidity between psychiatric illnesses and substance use.

Discussion

We interrogated chromatin interaction profiles of CNs and DNs, two primary neuronal subtypes involved in the neurocircuitry of addiction, to map GWAS risk variants of cigarette smoking and alcohol use traits to their target genes. While we identified cigarette smoking and alcohol use traits to be significantly enriched in CREs in the midbrain and PFC, it is possible for the enrichment pattern to be dependent on the sample size of GWAS used. We next built enhancer-promoter interaction landscapes in CNs and DNs by combining Hi-C and ATAC-seq and demonstrated brain region- and neuronal subtype-specific gene regulatory relationships. We then employed these profiles to perform CN and DN H-MAGMA, which was used to identify risk genes and neurobiological pathways underlying

PAU, DPW, ND, and CPD. Investigation into the biological pathways underlying cigarette smoking and alcohol use risk genes revealed the important role of drug catabolic process and alcohol metabolic process in substance use. Notably, we found that substance use risk genes were enriched for pathways associated with other neurodegenerative disorders such as tau protein binding for DPW and amyloid-beta metabolic process for CPD. The association between substance use and neurodegenerative disorders has been observed in a mouse model of Alzheimer's disease where alcohol exposure was shown to heighten neuronal and behavioral deficits related to Alzheimer's disease[133]. Thus, our results provide additional evidence to support that substance use and neurodegenerative disorders may share underlying genetic risk factors[134,135] and that risk variants associated with alcohol use may exacerbate neurodegenerative disorders by disrupting protein metabolism. We also identified an association between cigarette smoking and food intake which is in line with the previous reports linking weight gain with smoking cessation[136,137].

We next surveyed the cellular expression profiles of cigarette smoking and alcohol use risk genes to refine cortical and midbrain neuronal subtypes that confer risk for substance use. Within CNs, we found that cigarette smoking and alcohol use risk genes were highly expressed in glutamatergic neurons, providing an additional level of support for the neuronal basis of addiction[29]. We have previously shown that risk genes of psychiatric disorders were also enriched for glutamatergic neurons[86]). Based on prior epidemiological studies reporting higher substance use among individuals with mental health issues, these results suggest a potential cellular basis of comorbidity between substance use and psychiatric disorders[124]. We also identified potential divergence between ND and CPD such that risk genes associated with ND were enriched in inhibitory neurons, in contrast to the observed excitatory

enrichment for CPD risk genes. While this may hint at a distinct biological pattern underlying use (CPD) versus a use disorder (ND), caution should be exercised as this finding could also be influenced by the smaller number of genes associated with ND in comparison to CPD due to the smaller sample size of ND GWAS. Our cellular expression profiles within the DN lineage showed enrichment in intermediate DNs (DA1), suggesting early development as a critical time period linked to heritable risk of substance use that manifest later in life[106,138]. Finally, we leveraged cigarette smoking and alcohol use risk genes to identify the brain circuitry of addiction based on the hypothesis that defining brain regions most relevant to substance use may help derive better targeted approaches to treating SUD. In addition to the cortical and midbrain enrichment, we found enrichment for the amygdala and thalamus, reinforcing that multiple brain regions are important for understanding substance use and addiction.

To further characterize how cigarette smoking and alcohol use risk genes can expand our understanding of substance use and addiction, we generated a list of pleiotropic genes between PAU and CPD. In contrast to individual risk genes being more focused on individual substance use traits, we reasoned that pleiotropic genes would provide us with the opportunity to identify principal pathways associated with addiction. Therefore, we generated pleiotropic genes using both CN and DN H-MAGMA output files. DNs, but not CNs, showed strong gene-level overlap between PAU and CPD, conveying that DNs might be the central cell type that mediates pleiotropy. Based on our hypothesis that pleiotropic genes might translate beyond just cigarette smoking and alcohol use traits, we compared them with DEGs in response to cocaine[125]. Indeed, we showed that pleiotropic genes were likely to be

dysregulated in response to cocaine in the rat NAc, demonstrating that these genes may be more susceptible to a wide range of substance use.

Lastly, we took advantage of EnrichR to identify potential drug candidates to treat nicotine dependence and alcohol use disorder. We found potential drug candidates including those already on the market to treat various psychiatric illnesses such as depression and schizophrenia, further supporting a shared genetic architecture between psychiatric illnesses and substance use. These findings could be further prioritized by incorporating pathway analyses and literature review to corroborate the association between potential drug candidates' mechanism of action and substance use. Additionally, prioritized genes and drug candidates could be validated in model organism experiments. Together, we demonstrate that H-MAGMA built from brain region- and neuronal subtype-specific chromatin architecture can successfully identify risk genes and biologically relevant processes associated with cigarette smoking and alcohol use.

Methods

Postmortem brains and nuclei sorting

Tissue blocks dissected from the substantia nigra pars compacta (SNpc) along with surrounding regions of the ventral tegmental area (VTA) were dissected from midbrain slices from adult specimens collected by the Human Brain Collection Core (HBCC) at the National Institutes of Health. All procedures were approved by the Institutional Review Boards of the participating institutions.

Our protocol for the sorting of nuclei extracted from ventral midbrain/substantia nigra nuclei, including RNA-seq based quality checks confirming the dopaminergic phenotype of Nurr1+/NeuN+ double-positive nuclei, has been described in two earlier publications which

identified Nurr1+/NeuN+ nuclei to be enriched for dopaminergic genes over other neuronal subtypes[89,139]. The anti-Nurr1 antibody was produced in rabbits (Sigma-Aldrich, N4663) and the anti-NeuN antibody (EMD Millipore, MAB377X) was from mice.

Brain tissue was cut and dounced with 5 mL of lysis buffer with RNase inhibitor, then transferred to an ultracentrifuge tube, followed by immediately adding 9 mL of sucrose buffer underlaid beneath the solution. The samples were then spun at 24,000 rpm in an ultracentrifuge for 1 hour at 4°C. Next, the pellet was resuspended with 1mL of 0.1% BSA in DBPS, which was subsequently left on ice for 5–10 minutes. Pre-conjugated Nurr1 primary antibody (N4664) that had been incubated with the secondary antibody (Alexa 647) for an hour was then added to the nuclei suspension. Subsequently, 1.5 uL of NeuN antibody conjugated with Alexa 488 was added. Samples were wrapped in foil and rotated for 2 hours at 4°C. After 2 hours of incubation, DAPI was added to the reaction. The nuclei suspension is immediately taken to be processed on a FACSAria flow cytometry sorter, with all gates modified to eliminate debris and divide cells effectively, resulting in an apparent separation of nuclei populations through their fluorescent cell signal.

Dopaminergic neuronal Hi-C library generation

For each sample, 5,000 to 12,964 dopaminergic neuronal nuclei (NeuN+/Nurr1+) were processed through the Arima-HiC Kit User Guide for Mammalian Cell Lines (A51008, San Diego, CA) according to the manufacturer's instructions with the anti-Nurr1 antibody produced in rabbit (Sigma-Aldrich, N4663). For the present study, we generated new Hi-C libraries from five brain donors. Note that an earlier pilot study[89] had generated midbrain Arima Hi-C libraries from one donor (an independent donor not included in the present

study) which had produced better quality control indices as compared to an alternative low input protocol (Tn5-Hi-C).

Genomic DNA from Hi-C processed sorted nuclei was purified using the Beckman Coulter AMPure® SPRIselect Beads (Indianapolis, IN). Subsequently, samples were sonicated utilizing the Covaris S220 (Woburn, MA), then size selected and purified using Beckman Coulter AMPure® SPRIselect Beads (Indianapolis, IN) to target for 300–500 base pair sized fragments. Samples were then enriched in biotin using the Arima-HiC Kit for Library Preparation, alongside the Swift Biosciences® Accel-NGS® 2S Plus DNA Library Kit (San Diego, CA). Afterward, the Swift Biosciences Accel-NGS 2S Plus DNA library kit (21024, Ann Arbor, MI) was utilized for end-repair and adapter ligation. Unique indices were ligated to each sample using the Swift Biosciences 2S Indexing Kit (26148). DNA libraries were amplified and purified using the Kapa Hyper Prep Kit (NC0709851, Wilmington, MA) and Beckman Coulter AMPure® SPRIselect Beads according to the manufacturer's instructions. The resulting Hi-C libraries were sequenced through Illumina NovaSeq 6000 (150 bp paired-end sequencing) at a depth between 200-300 million reads per sample.

Hi-C analysis

We applied HiC-Pro (v2.11.1)[140] to the DN Hi-C[89]. In brief, we used Bowtie2 (v2.3.5.1)[141] with *--very-sensitive -L 30 --score-min L,-0.6,-0.2 --end-to-end --reorder* to align Hi-C reads to hg19 from UCSC database, and obtained unique mapped read pairs (valid pairs). Valid pairs were then used to generate Hi-C contact matrices at 10kb and 40kb resolutions. Hi-C contact matrices were subsequently normalized using Iterative Correction and Eigenvector decomposition (ICE) built in HiC-Pro. FitHiC2 (v2.0.7)[142] was then used to call chromatin interactions with *-U 2000000 -L 20000 -r 10000 -p 2*. Significant promoter-

anchored interactions, declared as chromatin contacts within 1Mb at an *FDR* threshold <1%, were defined based on overlap with gene promoter regions (2kb upstream and 1kb downstream of transcription start sites [TSS]). CN Hi-C data was obtained from Hu et al.[88].

TopDom (v0.9.1)[143] with default arguments was used to define topologically associating domains (TADs) from normalized 40kb contact matrices. TopDom firstly computed the average contact frequency (defined as a value of *binSignal*) between upstream and downstream regions for each bin. The *binSignal* values demarcate TAD boundaries such that it shows local minimum in a TAD boundary while it is relatively high within a TAD domain. We then used the pheatmap (v1.0.12) package to visualize chromatin contact maps. Information about samples and Hi-C libraries for DN is described in Table 3.1.

Gene regulatory relationships of cortical and dopaminergic neurons

RNA-seq and ATAC-seq data from CNs and DNs (cortical glutamatergic neurons) were obtained from GEO (GSE129017)[92]. We used FastQC (v.0.11.8)[144] to check the quality of RNA-seq and ATAC-seq reads.

For RNA-seq analysis, clean reads were mapped to the human reference genome (hg19) from the UCSC database with HISAT2 (v.2.2.1)[145] using default parameters. We assembled and quantified transcripts using StringTie (v.2.1.2)[146]. Normalized expression values (fragments per kilobase of exon model per million reads mapped, FPKM) of DN marker genes were compared between CNs and DNs.

For ATAC-seq analysis, we applied Bowtie2 (v2.3.5.1)[141] with *--very-sensitive* to map clean reads from ATAC-seq to hg19 from the UCSC database. After filtering out mitochondrial reads, duplicate reads were further removed by Picard (v.2.20.1, http://broadinstitute.github.io/picard/) by MarkDuplicates function. We then ran MACS2

(v.2.1.0.20150731)[147] with *--nolambda --nomodel* to call open chromatin regions. Blacklisted regions from ENCODE were removed from the MACS2-called peaks. Finally, we used DiffBind (v.2.13.1)[94] to analyze differentially open chromatin regions between CN and DN ATAC-seq data. Differentially open chromatin regions were selected based on *FDR<0.05*. Differential ATAC-seq peaks between CNs and DNs were intersected with promoter-anchored interactions to identify enhancer-promoter interactions as shown in Table 3.2.

GWAS datasets

We used the largest publicly available GWAS datasets from European ancestries of cigarette smoking and alcohol use traits. Datasets used were: Problematic alcohol use (PAU)[28], N = 435,563; Drinks Per Week (DPW)[26], N = 941,280; Nicotine Dependence (ND)[27], N = 78,067; Cigarettes Per Day (CPD)[26], N = 337,334.

LD score regression analysis

Stratified LD score regression (LDSC)[37] was used to estimate the enrichment of SNP-based heritability for PAU, DPW, ND, and CPD GWAS. *Cis*-regulatory elements (CREs) of the PFC and substantia nigra (SN) were defined as regions marked as active transcriptional start site (TSSs, state 1), flanking active TSSs (state 2), genic enhancers (state 6) and enhancers (state 7) in the chromHMM core 15-state model[90]. We acquired CREs of CNs and DNs by converting chromatin accessibility peaks reported in Zhang et al[92] to hg19 using liftOver. CREs of different cell types in the cortex were obtained by merging H3K27ac and H4K3me3 peaks reported from Nott et al[93]. Genetic variants were annotated to corresponding CREs, and SNP-based heritability enrichment was calculated using the GWAS summary statistics mentioned above.

H-MAGMA and gene selection

H-MAGMA input files were generated from the midbrain DN Hi-C data (DN H-MAGMA). Briefly, exonic and promoter SNPs were assigned to the genes in which they reside, while intronic and intergenic SNPs were coupled to their target genes based on significant chromatin interactions detected in DNs. For the adult brain and CN H-MAGMA, we used H-MAGMA input previously generated from bulk tissue and NeuN-positive cells sorted from the dorsolateral prefrontal cortex (DLPFC)[34,88], respectively. These input files are available in the GitHub repository at https://github.com/thewonlab/H-MAGMA.

Using these input files, we ran Hi-C coupled MAGMA (H-MAGMA) v.1.08[11] as previously described with the following code[86].

magma_v1.08/magma -–bfile g1000_eur –pval <GWAS summary statistics> use=rsid, p ncol=N -–gene-annot <MAGMA input annotation file> -–out<output file>

H-MAGMA converts SNP-level *p-values* into gene-level *p-values*, from which we selected protein-coding genes that are significantly associated with cigarette smoking and alcohol use traits at *FDR<0.05*. Since we used both cortical and dopaminergic Hi-C datasets, we obtained two gene sets, one from running CN H-MAGMA and the other from running DN H-MAGMA. These genes were used for subsequent functional analyses. We generated locus plots using the R package plotgardener[148,149].

Comparison of H-MAGMA results with differential expression signatures

We employed a linear regression model to test for association between H-MAGMA gene-level scores and differential gene expression signatures from human tissue after exposure to nicotine or alcohol using the following equation:

lm(DEG absolute log2 Fold Change ~ Z-scores calculated by H-MAGMA + number of SNPs assigned to each gene) Fold changes for differential expression signatures in response to cigarette smoking and alcohol use were obtained from Semick et al.[106] and Jensen et al.[107], respectively.

Gene ontology

We performed gene ontology (GO) analyses to identify biological pathways underlying cigarette smoking and alcohol use traits. Rather than using a selected set of genes with a specific *FDR* cutoff, we ran a rank-based gene ontology analysis using the Bioconductor package g:Profiler (v.0.7.0)[150]. Briefly, genes were ranked based on Z-scores calculated by H-MAGMA, such that genes more significantly associated with a given trait are listed at the top. Biological pathways over-represented by the highly ranked genes were selected.

```
gprofiler(<Ranked gene list>,organism="hsapiens",
ordered_query=T,significant=T,max_p_value=0.05,min_set_size=15
, max_set_size=600, min_isect_size=5, correction_method="fdr",
hier_filtering="strong",custom_bg=background gene set,
include_graph=T, src_filter="GO")
```

Cellular expression

We identified cellular expressions of cigarette smoking and alcohol use risk genes using publicly available single-cell RNA sequencing data (scRNA-seq)[77,79,119,151]. Given that GWAS power can influence the number of significant genes for a given trait, we used two different *FDR* thresholds to select cigarette smoking and alcohol use risk genes. We used

FDR<0.1 for GWAS for ND, given that there were <20 genome-wide significant loci; *FDR<0.05* for PAU, DPW, and CPD given >20 genome-wide significant loci. Next, we used scRNA-seq data from the human cortex to annotate cell-type specific and neuronal subcluster specific expressions of CN H-MAGMA risk genes for PAU, DPW, ND, and CPD[77,79]. Upon gene selection, we scaled expression profiles of each cell using the scale (x, center=T, scale=F) function in R and calculated the average expression of H-MAGMA risk genes in a given cell. Cell types (e.g. Neurons, Astrocytes, Microglia, Endothelial, Oligodendrocytes) and neuronal subclusters (e.g. excitatory and inhibitory neurons) with the highest average expression values were identified as central cell types underlying cigarette smoking and alcohol use traits. Similarly, we annotated DN H-MAGMA risk genes to midbrain cell-types identified from scRNA-seq from the human embryonic ventral midbrain during development[119]. Midbrain cell types include Radial glial (Rgl), Neuroblast, Progenitors (consisting of medial floorplate, lateral floorplate, midline, and basal plate progenitors), Neuronal progenitors (NProg), Oligodendrocyte progenitor cells (OPC), Dopaminergic neurons, Endothelial, GABAergic neurons, Microglial, Oculomotor and trochlear nucleus (OMTN), Pericytes, Red nucleus, and Serotonergic neurons. Additionally, we sought to identify specific dopaminergic clusters enriched for cigarette smoking and alcohol use risk genes. To achieve this, we annotated DN H-MAGMA risk genes to the dopaminergic lineage identified from the human embryonic ventral midbrain[119]. Lastly, we ran a linear regression model (H-MAGMA Z-score ~ cellular expression + mean expression across cell types + number of SNPs mapped by H-MAGMA) to evaluate statistical significance of cellular expression of risk genes[152].

Regional expression pattern

We measured brain regional expression profiles of cigarette smoking and alcohol use risk genes using a comprehensive dataset of the mouse nervous system from Zeisel et al. 2018[95]. Of the 24 brain regions represented, we analyzed the following regions: Cortex, Hippocampus, Amygdala, Striatum, Thalamus, Hypothalamus, Midbrain, Cerebellum, and Spinal cord. We generated a new list of risk genes for cigarette smoking and alcohol use traits by combining CN and DN H-MAGMA risk genes using `union(x,y)` in R to ensure that our findings were not being dominated by a specific H-MAGMA gene set. Next, we scaled each brain regional expression profile using `scale(x, center=T, scale=F)` in R and calculated the average expression of H-MAGMA risk genes. Regions with relatively enriched expression were identified as brain regions associated with cigarette smoking and alcohol use traits.

Pleiotropic genes

To identify shared neurobiological mechanisms between cigarette smoking and alcohol use traits, we compared gene-level association statistics of PAU and CPD using the rank-rank hypergeometric overlap (RRHO, v.1.40)[69] R package. Because non-coding genes could result in spurious relationships, we restricted our analysis to protein-coding genes and ran RRHO with the following command line.

```
RRHO.result=RRHO(Gene list 1, Gene list 2,
outputdir="~/output/", alternative="enrichment",
labels=c("Gene list 1", "Gene list 2"), BY=TRUE, log.ind=TRUE,
plot=TRUE).
```

Overlapping genes between PAU and CPD as identified by RRHO output files served as pleiotropic genes for downstream analyses. To identify biological pathways underlying pleiotropic genes, we ran GO analyses as previously described[150]. Because RRHO does not provide a ranked gene list, we performed GO analyses on the unranked pleiotropic genes with the following command line.

```
gprofiler(<Unranked pleiotropic gene list>,
organism="hsapiens",ordered_query=F,significant=T,max_p_value=
0.05,min_set_size=15,max_set_size=800,min_isect_size=5,correct
ion_method="fdr",hier_filtering="moderate",custom_bg=backgroun
d gene set, include_graph=T, src_filter="GO)
```

Differentially expressed genes in response to cocaine

To test for cell-type specific changes of pleiotropic genes in response to cocaine, we first overlapped H-MAGMA genes with differentially expressed genes (DEGs) in the rodent nucleus accumbens (NAc) upon cocaine exposure[125]. DEGs from dopaminoceptive neurons expressing dopamine receptors (Drd1-MSNs, Drd2-MSNs1, Drd2-MSNs2, Drd3-MSNs) identified from the NAc were converted from the rodent HUGO Gene Nomenclature Committee (HGNC) symbol to their homologous human Ensembl gene IDs. Rodent genes that did not have a corresponding human Ensembl ID were removed from the analysis, resulting in a total of 12,437 cocaine background genes from the dopaminergic clusters. We next selected for DEGs at FDR adjusted $p\text{-}value<0.05$ from the dataset, resulting in a total of 608 significant dopaminergic DEGs from the cocaine background genes. These 608 significant DEGs were classified as cocaine DEGs for analysis. Because cocaine background genes differ from all H-MAGMA background genes, we generated a comparable H-

MAGMA risk gene set by overlapping pleiotropic genes with the cocaine background genes. Next, we compared the proportion of pleiotropic genes that overlapped with cocaine DEGs using the Venn(x) function in the Vennerable package (v.3.1.0.9000) in R. We also applied a Fisher's exact test to test for significance of overlap as follows:

```
fisher.test(matrix(c(overlapping set, Gene list 1-overlapping set, Gene list 2-overlapping set, cocaine background genes),2,2)
```

Lastly, to assess cell-type specific transcriptional changes of pleiotropic genes upon cocaine treatment, we compared transcriptional changes between saline vs. cocaine in the scRNA-seq data[125]. Briefly, we scaled each cell using the `scale(x, center=T, scale=F)` function in R and generated box plots comparing saline vs. cocaine treatment for each cluster (e.g. Astrocytes, Dopaminergic neurons, GABAergic neurons, Glutamatergic neurons, Metabotropic glutamate receptor [Grm8-MSN], Microglia, Mural cells, Oligodendrocytes, Polydendrocytes, Interneurons). To test if our risk genes behave differently after cocaine treatment, we compared cellular expression levels between saline and cocaine treatment using the `t.test (x1, x2)` function in R.

Drug enrichment analysis

We used EnrichR[128] to obtain a list of potential drug candidates for cigarette smoking and alcohol use risk genes. We limited our analysis to the Drug Signature (DsigDB) and Drug Matrix databases of EnrichR as they were the most comprehensive drug libraries available on the platform. Using cigarette smoking and alcohol use risk genes identified with the threshold of *FDR<0.05* for PAU, DPW, ND, and CPD, we adjusted drug-associated *p*-

values provided by EnrichR after multiple testing correction and selected for significant drugs approved by the Food and Drug Administration (FDA) and small molecules.

Figures

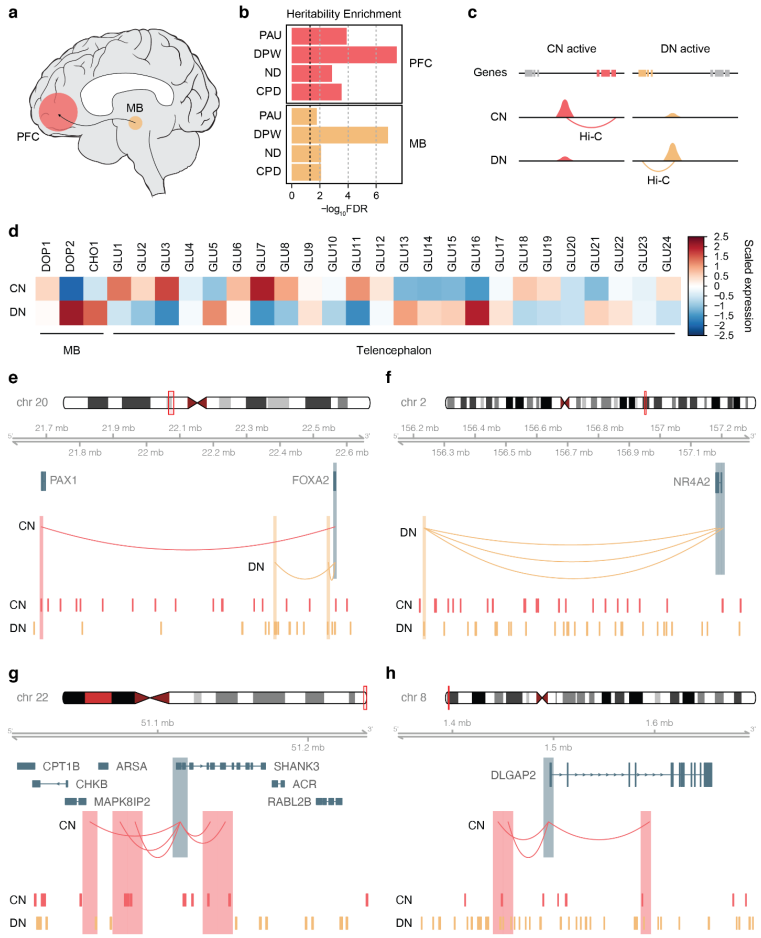


Figure 3.1. Gene regulatory landscape in cortical and dopaminergic neurons. (a) Brain reward circuitry encompasses the midbrain (MB) and its projection to the prefrontal cortex (PFC). **(b)** Stratified LDSC analysis determined that *cis*-regulatory elements (CREs) in the PFC and substantia nigra (SN) are enriched for genetic risk factors for problematic alcohol use (PAU), drinks per week (DPW), nicotine dependence (ND), and cigarettes per day (CPD). The black dotted line represents *FDR=0.05*.(**c)** Dopaminergic neuronal (DN) differentially accessible regions (DARs) were linked to their target genes using DN Hi-C data, while cortical neuronal (CN) DARs were linked to target genes using CN Hi-C data. **(d)** Genes mapped to DN-DARs were highly expressed in midbrain dopaminergic (DOP2) and cholinergic neurons (CHO1), while genes mapped to CN-DARs were highly expressed in telencephalic glutamatergic neurons (GLU1, 3, 7, 11). (**e-h)** Different enhancer connectivity between CNs and DNs for *FOXA2* (**e**), *NR4A2* (**f**), *SHANK3* (**g**), and *DLGAP2* (**h**) loci. Promoters of genes (*FOXA2*, *NR4A2*, *SHANK3*, *DLGAP2*) are highlighted in blue, while their interaction targets in CN and DN are highlighted in red and orange, respectively. CN- and DN-DAR are depicted in the bottom tracks.

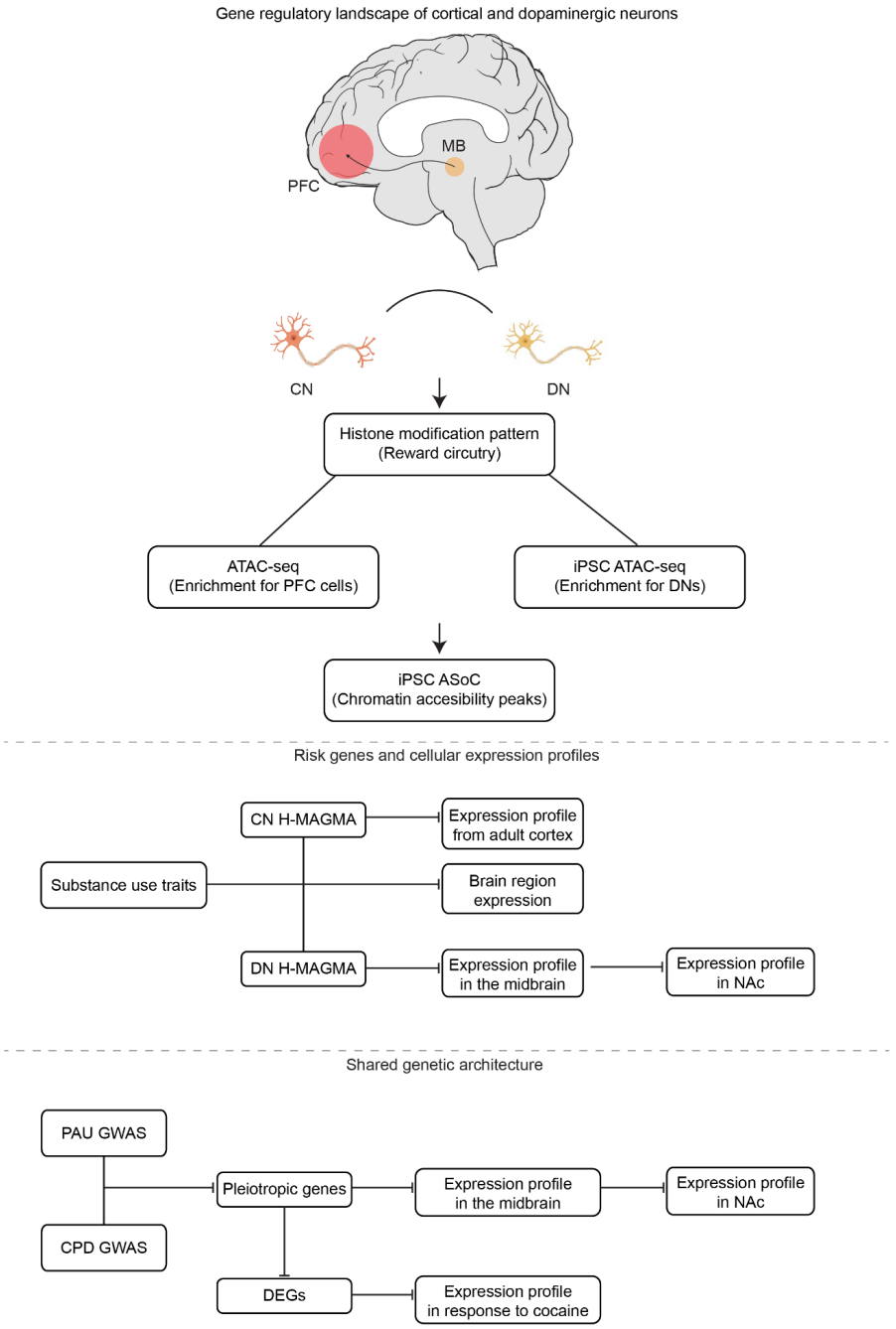


Figure 3.2. Analysis overview of the present study

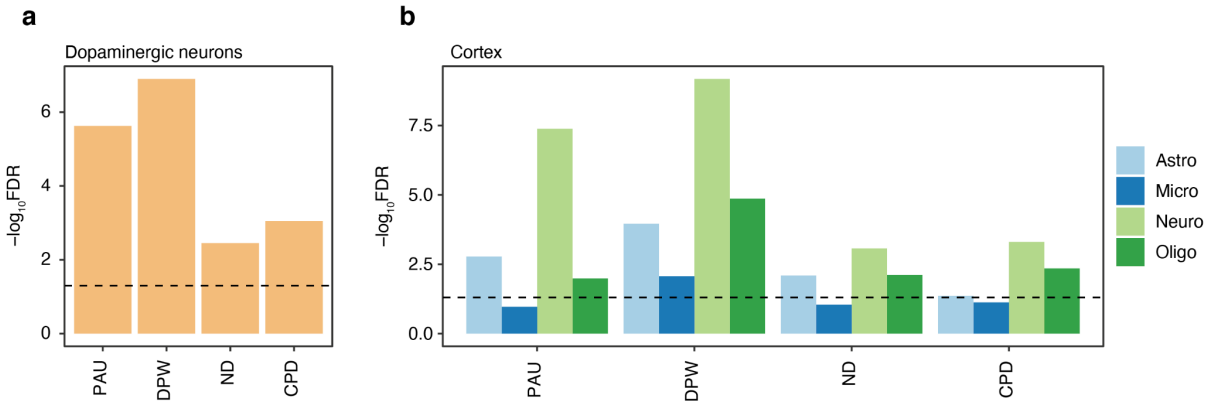


Figure 3.3. Heritability enrichment of cigarette smoking and alcohol use traits in dopaminergic and cortical cell types. (a) Heritability enrichment of cigarette smoking and alcohol use traits using stratified LDSC. Genetic risk variants associated with cigarette smoking and alcohol use traits are enriched for DN-CREs. (**b**) Cell-type specific heritability enrichment of cigarette smoking and alcohol use traits in the cortex. We observed neuronal enrichment for cigarette smoking and alcohol use traits. Dotted lines indicate *FDR=0.05*. Astro, astrocyte; Micro, microglia; Neuro, neuron; Oligo, oligodendrocyte.

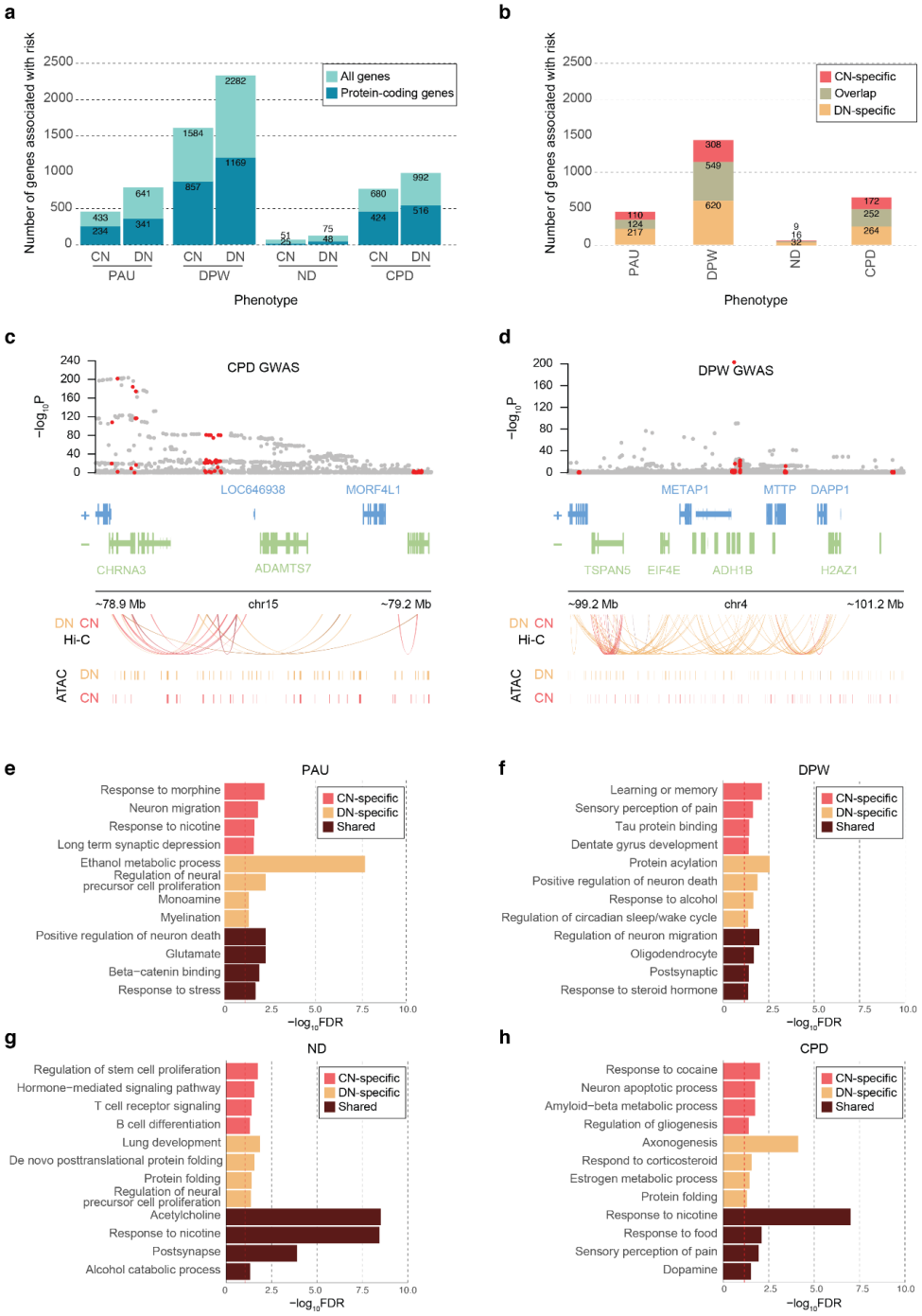


Figure 3.4. Pathways associated with cigarette smoking and alcohol use traits. (a) The number of risk genes for cigarette smoking and alcohol use traits based on H-MAGMA built from CN and DN Hi-C data (*FDR<0.05*). For each stacked bar plot, an upper bar plot and number in light blue denote all genes, whereas a lower layer and number in dark blue

correspond to protein-coding genes. (**b)** The number of protein-coding genes associated with smoking and alcohol use traits by CN H-MAGMA only (CN-specific), DN H-MAGMA only (DN-specific), and by both CN and DN H-MAGMA (overlap) at an *FDR* threshold of 0.05**. (c)** H-MAGMA identifies *CHRNA3* to be associated with CPD**.** From top to bottom, we display the GWAS variant association, gene model, Hi-C loops, and CREs. GWAS variant association, gray dots represent all SNPs in the locus, red dots represent SNPs annotated to *CHRNA3* via H-MAGMA (either DN or CN). Hi-C loops and CREs, red and orange lines represent the regulatory architecture of CNs and DNs, respectively. (**d)** H-MAGMA identifies *ADH1B* to be associated with DPW. Red dots in the GWAS association represent SNPs annotated to *ADH1B* via H-MAGMA (either DN or CN). **e-h**. Gene ontologies (GO) enriched for PAU **(e)**, DPW **(f)**, ND **(g)**, and CPD **(h)**. CN-specific GO terms represent terms unique to genes identified from H-MAGMA built on CN Hi-C data, while DN-specific GO terms represent terms unique to genes identified from H-MAGMA built from DN Hi-C data. Shared terms denote GO terms detected in both CN and DN H-MAGMA results. The dotted red line denotes *FDR=0.05*.

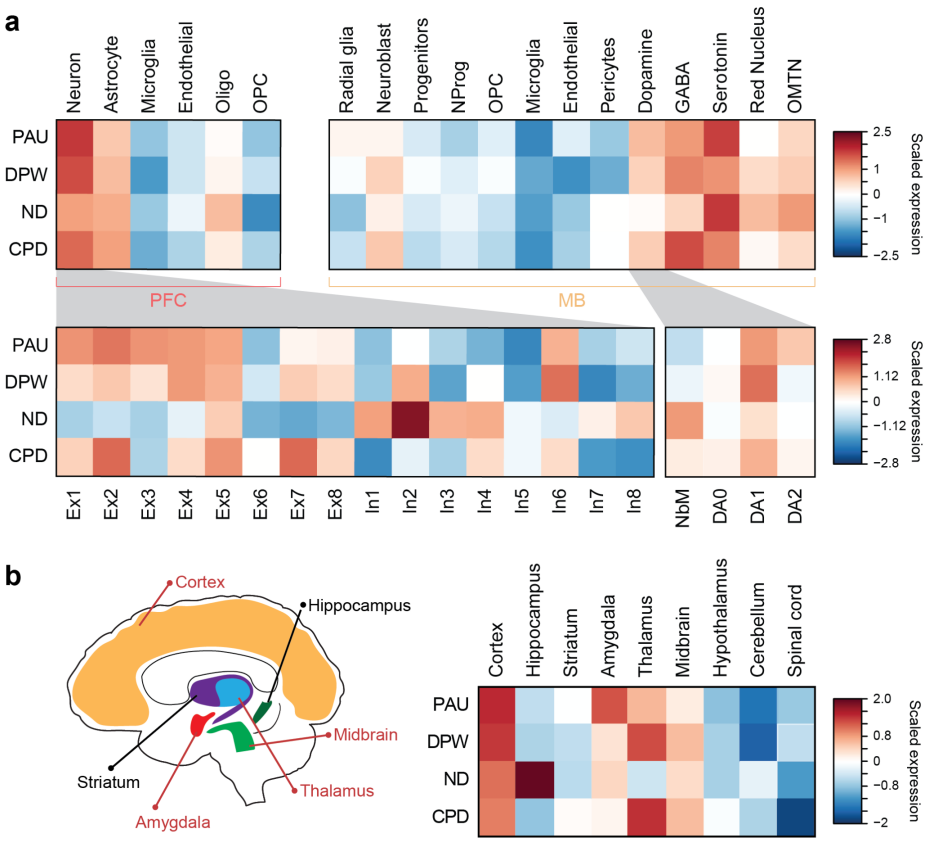


Figure 3.5. Cellular and brain regional expression profiles of cigarette smoking and alcohol use traits. (a) Top left panel represents cellular expression profiles of cigarette smoking and alcohol use risk genes identified from CN H-MAGMA using scRNA-seq data from the adult cortex[77,79]. Genetic risk factors underlying cigarette smoking and alcohol use influence genes highly expressed in neurons. Bottom left panel represents risk gene expression across neuronal subclusters. OPC, oligodendrocytes progenitor cells; Ex, excitatory neurons; In, inhibitory neurons. Top right panel, cellular expression profiles of cigarette smoking and alcohol use risk genes identified from DN H-MAGMA using scRNA-seq from the ventral midbrain of human embryo. Risk genes are highly expressed in dopaminergic, GABA-ergic, and serotonergic neurons in the midbrain. NProg, neuronal progenitors; OMNT, oculomotor and trochlear nucleus. Bottom right panel, cigarette smoking and alcohol use risk genes were enriched for DA1 across DN development in human embryonic midbrain. NbM, medial neuroblasts and precursors of DNs; DA0, immature DNs; DA1, intermediate DNs; DA2 matured DNs. **(b)** Left**,** graphic representation of brain regions with elevated expression levels of risk genes for substance use traits. Regions highlighted in red are enriched for at least three of the four traits. Right, brain regional expression profiles of cigarette smoking and alcohol use risk genes using scRNA-seq from the mouse nervous system. Risk gene expression spans multiple brain regions including the cortex, amygdala, and midbrain.

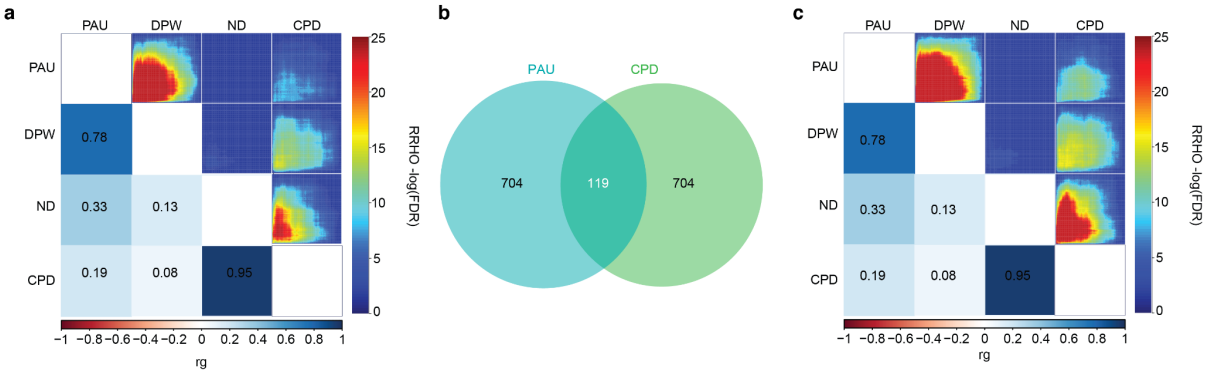


Figure 3.6. Genetic correlations and overlapping genes between cigarette smoking and alcohol use traits. (a) LDSC and RRHO were used to estimate genetic correlations and gene-level overlap between cigarette smoking and alcohol use traits, respectively. Bottom left plot represents genetic correlations (rg) while the top-right plot denotes gene-level overlap using CN H-MAGMA output files. **(b)** Overlap between PAU and CPD risk genes identified by CN H-MAGMA using RRHO. **(c)** Genetic correlations (bottom left) and gene-level overlap from DN H-MAGMA output files.

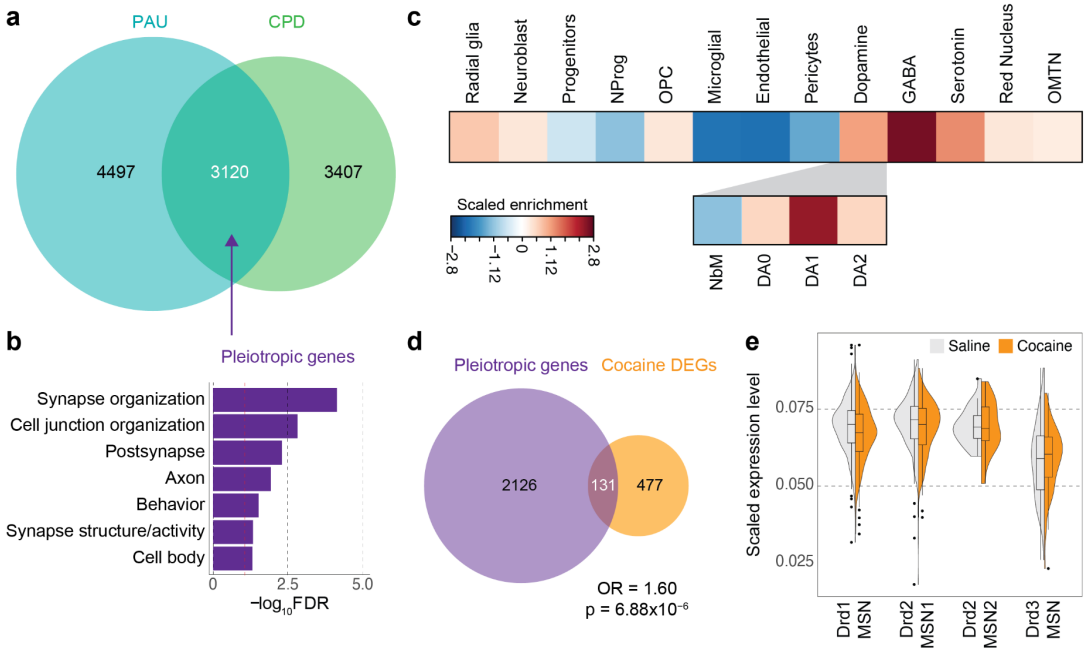


Figure 3.7. Pleiotropic genes highlight shared neurobiological bases of cigarette smoking and alcohol use. (a) Overlap between PAU and CPD risk genes identified by DN H-MAGMA using a rank rank hypergeometric overlap (RRHO) test. Overlapping genes represent pleiotropic genes. **(b)** Biological processes and molecular functions enriched for pleiotropic genes. Dotted red line denotes *FDR=0.05*. **(c)** Cellular expression profiles of pleiotropic genes in the midbrain (top plot) and dopaminergic lineage (bottom plot). Pleiotropic genes are highly expressed in GABAergic midbrain neurons and intermediate DNs (DA1). **(d)** Overlap between pleiotropic genes and differentially expressed genes (DEGs) in the rat NAc after cocaine treatment (Fisher's exact test, $p=6.88\times10^{-6}$; odds ratio [OR]=1.60; 95% confidence interval [CI]=1.34–1.96). **(e)** Cellular expression changes of pleiotropic genes in response to cocaine treatment[125]. The x-axis indicates medium spiny neuronal (MSN) clusters identified in the rat NAc while the y-axis indicates scaled expression values of the pleiotropic genes in each cluster.

Tables

Table 3.1. Hi-C libraries used in this chapter.

Sample information	***cis* reads**	**total reads**	***cis* proportion**	**Diagnosis**	**NeuN+ /Nurr1+ Collected for Hi-C**
Male, 35 yo African American	158,222,950	230,186,257	0.69	Control	5,336
Female, 45 yo African American	192,097,648	266,025,620	0.72	Control	5,000
Female, 46 yo African American	164,038,226	197,508,199	0.83	Bipolar	5,000
Male, 36 yo Caucasian	176,575,402	237,642,800	0.74	Bipolar	5,916
Male, 45 yo African American	243,313,838	325,461,515	0.75	Control	12,964
	934,248,064	1,256,824,391	0.74		

Table 3.2. Source of cis-regulatory elements used in this chapter.

CREs	**Cellular/Tissue source**	**Figures**	**Assay**	**Reference**
Midbrain CRE	Substantia Nigra	Figure 3.1B	chromHMM, chromatin states 1,2,6,7	Roadmap epigenomics[90]
PFC CRE	Prefrontal Cortex			
DN-CRE DN-DAR	hiPSC-derived dopaminergic neurons	Figure 3.1C,E,F, Figure 3.3A	CREs were defined from ATAC-seq peak calls. DARs were defined by differential peak calls.	Zhang et al.[92]
CN-CRE CN-DAR	hiPSC-derived cortical glutamatergic neurons			
Astro-CRE	Astrocytes sorted from the cortex	Figure 3.3B	H3K27ac, H3K4me3 merged peaks	Nott et al.[93]

CHAPTER 4: DETAILED PROTOCOL OF H-MAGMA AND EXPANDING THE TOOL TO NON-BRAIN CELL-TYPES[3]

Introduction

Genome-wide association studies (GWAS) have provided an avenue for scientists to identify common genetic variations associated with human traits and diseases. However, despite the surge in GWAS in recent years, the functional outcomes of the common variants are not well understood because the majority of them reside in poorly characterized, non-coding regions of the genome[153]. Previous studies have used Multi-Marker Analysis of GenoMic Annotation (MAGMA), a gene-based analysis tool that converts single nucleotide polymorphism (SNP)-level P-values identified from GWAS to gene-level P-values to assign variants to their target genes[11]. While revolutionary, MAGMA mainly relies on positional mapping, typically linking non-coding variants to the nearest genes. However, advancement in epigenomic profiling highlights the complexity of gene regulatory architecture. For example, it is possible for variants to interact with and regulate distal genes[12], necessitating gene regulatory architecture to be taken into account for variant-gene annotation. One way to investigate the gene regulatory architecture includes Hi-C, a genome-wide chromosome conformation capture technique[154]. Using Hi-C data, we can complement gene-based analysis tools built upon the linear genome such as MAGMA to fully capture distal regulatory relationship between variants and genes. To this end, we have developed a tool, Hi-C

[3] Reproduced with permission from Nature Springer. Sey NYA, Pratt M, Won H. Annotating genetic variants to target genes using H-MAGMA. Nat Protoc. 2022. doi: 10.1038/s41596-022-00745-z. H.W. designed the H-MAGMA framework.

Coupled MAGMA (H-MAGMA), that incorporates Hi-C based variant-gene relationships in the gene-based analytic framework of MAGMA[86].

In this protocol, we outline the H-MAGMA framework including (1) how to generate the H-MAGMA variant-gene annotation file that provides variant-gene relationships using Hi-C data from the adult human brain[34] and (2) how to run H-MAGMA using GWAS summary statistics of Parkinson's Disease[21]. While we use Hi-C data from the adult human brain and Parkinson's Disease GWAS in this protocol, the H-MAGMA framework is versatile and can be adapted to functionally annotate any GWAS by generating the H-MAGMA variant-gene annotation file from a tissue or cell type that is most enriched for the GWAS trait of interest. To highlight the versatility of the H-MAGMA framework, we also provide H-MAGMA variant-gene annotation files generated from promoter-anchored interactions of 28 tissue and cell types reported in Jung et al.[32]. In addition to the variant-gene annotation files provided in this protocol, users can also create their own annotation files from Hi-C datasets either generated by the user or acquired from publicly available resources.

Development of the protocol

We developed H-MAGMA based on the hypothesis that we can leverage chromatin architecture to extend the capacity of MAGMA to better annotate non-coding variants to their target genes. In our primary application of H-MAGMA to GWAS of psychiatric (Schizophrenia[5] [SCZ], Bipolar disorder[6] [BD], Autism Spectrum Disorder[17] [ASD], Attention Deficit Hyperactivity Disorder[18] [ADHD], Major Depressive Disorder[19] [MDD]) and neurodegenerative disorders (Parkinson's disorder[21] [PD], Alzheimer's disease[20] [AD], Multiple Sclerosis[23] [MS], and Amyotrophic Lateral Sclerosis[22] [ALS]), we identified novel

target genes and biological pathways underlying each disorder [86]. We further identified critical developmental windows and central cell types in understanding the disease biology[86]. Specifically, we identified that psychiatric disorder-associated genes exhibit prenatal enrichment compared to the postnatal enrichment of neurodegenerative disorder-associated genes. Additionally, we found that biological processes including transcriptional regulation, synaptic transmission, and neuroinflammation were involved with various brain disorders[86]. Collectively, these findings suggest that H-MAGMA can effectively broaden our understanding of disease biology by annotating common variants to their target genes with the use of functional genomic data.

Comparison with other methods

To investigate the effectiveness of H-MAGMA, we compared our results from H-MAGMA to that of conventional MAGMA. We discovered that H-MAGMA could identify disease-associated genes that were missed by conventional MAGMA by predominantly linking non-coding variants to distal genes. Variants assigned to target genes by H-MAGMA explained a significant proportion of heritability of brain disorders[86]. To further investigate the biological relevance of these findings, we compared developmental trajectories of H-MAGMA-associated genes with those of conventional MAGMA-associated genes. We noted important differences between H-MAGMA and conventional MAGMA. For instance, conventional MAGMA-associated genes for ASD exhibit postnatal enrichment which contradicts prior evidence supporting the early developmental origin of ASD[155]. On the contrary, ASD risk genes annotated by H-MAGMA showed prenatal enrichment.

We further improved H-MAGMA by implementing Imhof's algorithm which can better control for type I error rate inflation[156,157]. As a result of this update, we reexamined

our previous analysis of psychiatric disorders. While we detected a smaller number of genes for each disorder than previously reported, we recapitulate our prior findings including the prenatal enrichment of psychiatric disorder-associated genes[157]. Collectively, we demonstrate the statistical rigor of H-MAGMA and its accuracy in delineating biological processes underlying traits and diseases.

Applications of the method

Since its introduction, several studies have utilized H-MAGMA to detect risk genes associated with various human traits and diseases. For example, Matoba et al.[158] employed H-MAGMA using a variant-gene annotation file built from chromatin interactions in the fetal brain[35] to identify ASD-associated genes from risk variants. They observed that genes detected by H-MAGMA played roles in telencephalon development and regulation of synapse organization. Additionally, the authors noted that a subset of the risk genes were differentially expressed in postmortem brains of ASD compared to neurotypical individuals[158].

Similarly, Quach et al.[27] applied H-MAGMA using Hi-C datasets from fetal[35] and adult[34] brain tissues to a GWAS of nicotine dependence and detected 11 and 13 protein-coding genes, respectively. Meanwhile, Song et al.[159] ran H-MAGMA for traits including intelligence quotient (IQ), SCZ, MDD, and AD to identify implicated biological processes. They noted that H-MAGMA associated genes, particularly for AD and SCZ, supported prior evidence of disease risk.

Feleke et al.[160] used H-MAGMA built on Hi-C data from the adult brain to identify risk genes of Lewy body diseases. Here, the authors applied H-MAGMA to PD, PD with dementia, and dementia with Lewy bodies, three neurodegenerative disorders that share

several similarities. Using H-MAGMA, they were able to identify distinct cell types and biological processes that differentiate the diseases from each other[160].

Given that gene regulatory landscapes are highly tissue and cell type specific, Hu et al.[88] leveraged data from glial 3D chromatin interactions to characterize AD GWAS using H-MAGMA. They identified 181 AD-associated genes, including *BIN1*[88]. Taken together, these studies highlight that tissue and cell type specific application of H-MAGMA can predict risk genes associated with human diseases within the biologically relevant contexts.

Currently, H-MAGMA variant-gene annotation files are only available for brain cell types, limiting its application to non-brain disorders. To allow researchers from other backgrounds to use H-MAGMA for their research, we have generated variant-gene annotation files using other Hi-C datasets including lung, pancreas, and liver[32] as part of this protocol. While we have primarily used Hi-C and promoter-capture Hi-C data to identify variant-gene relationship, any epigenetic data that provides such relationship including HiChiP[161] and PLACseq[162] can be used to generate H-MAGMA variant-gene annotation files. Moreover, whereas we generated H-MAGMA variant-gene annotation files from 3D chromatin interaction data acquired from various tissue types, cell-type specific 3D chromatin interactions data are still relatively rare. We expect that cell-type specific 3D chromatin data will become more and more available, which would enable cell type specific dissection of GWAS.

Experimental design

In this protocol, we have outlined how to generate the H-MAGMA variant-gene annotation file that provides the variant-gene relationship required to run H-MAGMA as well as the steps to run H-MAGMA (Fig. 4.1). Generation of H-MAGMA variant-gene annotation

files can be broken down to two steps: First, exonic and promoter variants are directly linked to target genes based on positional mapping. This assumes that exonic and promoter variants are more likely to impact the genes in which they reside. Second, intronic and intergenic variants are mapped to their target genes based on Hi-C interactions. Prior to implementing the protocol, Hi-C interactions need to be formatted in the BEDPE file format that describes two anchor points of the chromatin interactions: chromosome1, start1, end1, chromosome2, start2, end2. Once the variant-gene annotation file is generated, the procedure to run H-MAGMA mirrors that of conventional MAGMA. All files listed in the protocol are publicly available (see Data Availability).

Level of expertise required

The protocol assumes familiarity with the R programming language for generating the variant-gene annotation file and linux commands to run H-MAGMA.

Format of H-MAGMA output files

Exon/promoter coordinates (steps 4-5, Table 4.1): The first step to creating an H-MAGMA variant-gene annotation file is to create a GenomicRange (GRange) object for exon and promoter coordinates. The corresponding GRanges objects (*promoterranges*, *exonranges*) are provided as an RData file, *exon_promoranges.rda* (see Data Availability). Here, we show the output from *exonranges*. Column names denote following:

- SEQNAMES - Chromosome in which the gene is located.
- RANGES - Genomic location of the gene.
- STRAND - The DNA strand orientation of the gene where star(*) represents an unspecified strand, plus (+) represents features from start to end, and minus (-) indicates features from end to start.

- GENE - Gene names in ENSEMBL gene IDs.

SNPs located in exons and promoters (steps 9-11, Table 4.2): In steps 9-11, we overlap SNP annotations from European Ancestry (EUR.bim) with the exon and promoter coordinates to create GRange objects that denote variant-gene relationships. These objects (*snpro*,*snpexon*) are provided as an RData file, *snp_locating_in_exon_promoter_transcript_level.rda* (see Data Availability). Here, we show the output from *snpro*.

Hi-C annotated SNPs (step 20, Table 4.3): SNPs that did not fall within an exon or gene promoter are annotated to target genes based on Hi-C interaction data. This object (*snpint*) is provided as an RData file as: *Hi-C_transcript_interacting_snp.rda.*

Variant-gene annotation file (step 28, Table 4.4): This is an H-MAGMA-compatible variant-gene annotation file that is recognizable by the MAGMA software. This file consists of a gene, its genomic location, and a list of SNPs assigned to the gene. This is provided as: *Adultbrain.transcript.annot*.

PD output file (steps 29-33, Table 4.5): Once the H-MAGMA variant-gene annotation file is created, users can run H-MAGMA for their trait of interest to obtain a gene list associated with the trait. We use Parkinson's Disease as an example in this protocol. Shown below are the first three lines from the PD output file (PD_GWAS) and the first three genes associated with GWAS summary statistics of Parkinson's Disease[21] (PD.genes.csv) using an *FDR* threshold of 0.05, respectively.

Materials

Hardware

- To run H-MAGMA, you will need a linux-based operating system with a minimum of 2.5 GB of memory.

Software

- R: This protocol requires you to install the R software (>3.6.0). It is freely available from https://www.r-project.org.
- R Libraries

Install the following libraries into R using the code below.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("GenomicRanges")
BiocManager::install("biomaRt")
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggplot2")
```

- MAGMA: Since H-MAGMA was built based on the MAGMA software, we will adopt the MAGMA framework with a modified version of the variant-gene annotation file to allow for incorporation of 3D chromatin interaction data. The most up-to-date MAGMA program can be downloaded from the following link: https://ctg.cncr.nl/software/magma. Here, we will use MAGMA v1.09.

Required data

Download all required data in a ~/work/ directory. All files required to run the protocol are also provided in Zenodo[163] at https://doi.org/10.5281/zenodo.5503876 and in our GitHub repository at https://github.com/thewonlab/H-MAGMA[164]. For the sake of this

protocol, we will use the reference genome from European ancestry which is available from the MAGMA website using the following link: https://ctg.cncr.nl/software/magma.

- Obtain gene and exon coordinates from Gencode version 26 using this link: https://www.gencodegenes.org/human/release_26lift37.html. We downloaded the GFT file for basic gene annotation and converted them to .bed files. Promoters were defined as 2kb upstream to every transcription start site (TSS). Gene, exon, and promoter coordinate files are listed as *Gencode26_gene.bed*, *Gencode26_exon.bed,* and *Gencode26_promoter.bed*, respectively, in the Zenodo repository. Columns for these files are defined as [chromosome, start, end, gene].
    - For this protocol, we will use chromatin interaction from the adult brain obtained from the PsychEncode consortium: http://resource.psychencode.org. This is listed as Promoter-anchored_chromatin_loops.bed on the PsychEncode consortium website. We also provide it in the Zenodo repository in a BEDPE file format as *adultbrain_hic.bedpe*. Columns are defined as [chrom1, start1, end1, chrom2, start2, end2]. Note that the Hi-C dataset can be changed based on the tissue or cell type of interest.
    - Download the reference genome for European ancestry from the MAGMA website using the following link: https://ctg.cncr.nl/software/magma. This file is provided in the Zenodo repository as *EUR.bim*.
    - Download summary statistics from your GWAS of interest. We will use GWAS of PD [21] as an example. This is provided as *PD.summary.stat.txt* in the Zenodo repository.

- Lastly, download the gene annotation file to convert ENSEMBL gene IDs to HGNC symbols. We have provided the annotation file as *geneAnno_allgenes.rda* in the Zenodo repository.

Procedure

CRITICAL. We have broken the protocol into two main sections. The first section walks through generating the variant-gene annotation file needed to run H-MAGMA, while the second portion walks through running H-MAGMA using GWAS summary statistics.

Prepare libraries and directory

Timing 5 min

1. Enter the following commands in R.

```
options(stringsAsFactors=F)
library (GenomicRanges)
library(biomaRt)
library(dplyr)
```

2. Navigate to the appropriate work directory with all downloaded files using the command below.

```
setwd("~/work/")
```

Create GenomicRanges (GRanges) objects for exon and promoter coordinates

Timing ~5 min

3. Read in exonic and promoter coordinate files by entering the following commands.

```
exon <- read.table("Gencode26_exon.bed")
exon$V1 <- sub ("^", "chr", exon$V1)
```

```
promoter <- read.table("Gencode26_promoter.bed")
promoter$V1 <- sub ("^", "chr", promoter$V1)
```

4. Create a GRanges object for exon and promoter definitions.

```
exonranges<- GRanges (exon [,1], IRanges(exon [,2],
exon[,3]),gene=exon[,4])
promoterranges <- GRanges (promoter [,1], IRanges(promoter
[,2], promoter[,3]),  gene=promoter[,4])
```

5. Save exon and promoter GRanges objects as an .rda file for future use.

```
save (exonranges, promoterranges, file="exon_promoranges.rda")
```

Generate a GRanges object for single nucleotide polymorphism (SNP) coordinates.

Timing ~5 min

6. Read in the SNP annotation file using the following command.

```
snps <- read.table("EUR.bim")
snps <- snps[, c(1,2,4)]
colnames(snps) <- c("chr","SNP","Position")
snps$chr <- sub("^", "chr", snps$chr)
```

7. Create a GRanges object for SNP annotations.

```
snps<-GRanges(snps$chr, IRanges(snps$Position, snps$Position),
rsid=snps$SNP)
```

8. Save the SNP GRanges object as an .rda file for future use.

```
save(snps, file="snps.rda")
```

Assign SNPs to genes by overlapping SNPs with exons and promoters.

Timing 10 - 15 min

9. Identify genes that map to SNPs residing in exons by overlapping the GRanges object for SNPs with the GRanges object for exons.

```
olap <- findOverlaps(snps,exonranges)
snpexon <- snps[queryHits(olap)]
mcols(snpexon)<-cbind(mcols(snpexon),
mcols(exonranges[subjectHits(olap)]))
snpexon <- snpexon[seqnames(snpexon)!="chrX"]
snpexon <- unique(snpexon)
```

10. Overlap the GRanges object for promoters with the GRanges object for SNPs. Similar to the code above, this command will identify genes that map to SNPs residing in promoter regions.

```
olap <- findOverlaps(snps,promoterranges)
snpro <- snps[queryHits(olap)]
mcols(snpro)<-cbind(mcols(snpro),
mcols(promoterranges[subjectHits(olap)]))
snpro <- snpro[seqnames(snpro)!="chrX"]
snpro <- unique(snpro)
```

11. Save SNPs overlapping with exons and promoters as an .rda file.

```
save(snpro,snpexon,
file="snp_locating_in_exon_promoter_transcript_level.rda")
```

Assign unmapped SNPs to genes based on Hi-C interaction data.

Timing 30 - 45 min

12. After identification of SNPs that reside in exons and/or promoters, there will be a subset of SNPs that do not overlap with either (e.g. intergenic or intronic SNPs). In the following steps, we will match those SNPs to cognate genes based on Hi-C interaction data. Identify unmapped SNPs using the code below.

```
snpranges <- snps[!(snps$rsid %in% snpexon$rsid), ]
snpranges <- snpranges[!(snpranges$rsid %in% snpro$rsid), ]
```

13. Save unmapped SNPs as an .rda file.

```
save(snpranges,file="non_exonic_promoter_snp.rda")
```

14. Read in adult brain Hi-C data. The first three columns represent genomic coordinates [chr, start, end] of loop anchor 1 and the last three columns represent genomic coordinates of loop anchor 2.

CRITICAL STEP. While we use a Hi-C dataset from the adult human brain for this protocol, Hi-C datasets from the tissue or cell-type associated with the disease/trait of interest can be used instead.

```
hic <- read.table("adultbrain_hic.bedpe", header=T)
```

15. Restructure Hi-C data to account for both anchors 1 and 2 using the command below.

```
hic.int1 <- hic [,1:6]
hic.int2 <- hic[, c(4:6,1:3)]
colnames(hic.int1) = colnames(hic.int2) = c("chrom1",
"start1", "end1", "chrom2", "start2", "end2")
hic.comb <- rbind(hic.int1, hic.int2)
```

16. Generate a GRanges object for the adult brain Hi-C data using the command below.

```
hicranges<-GRanges(hic.comb$chrom1,
IRanges(as.numeric(hic.comb$start1),
```

```
as.numeric(hic.comb$end1)),
int1=hic.comb$start2,int2=hic.comb$end2)
```

17. Identify promoter-anchored interactions by overlapping loop anchor 1 with promoters.

```
olap <- findOverlaps(hicranges,promoterranges)
generanges <- hicranges[queryHits(olap)]
mcols(generanges) <- cbind(mcols(hicranges[queryHits(olap)]),
mcols(promoterranges[subjectHits(olap)]))
```

18. Reverse the order of the GRanges object from Step 17. The generanges object created in Step 17 is in a format of loop anchor 1, loop anchor 2, followed by gene name. We reverse the order of loop anchors so that the resulting genebed object has a format of loop anchor 2, loop anchor 1, followed by gene name.

```
genebed<-data.frame(chr=seqnames(generanges),
snp.start=generanges$int1, snp.end=generanges$int2,
gene.start=start(generanges),
gene.end=start(generanges)+width(generanges)-1,
ensg=generanges$gene)
genebed <- unique(genebed)
```

19. Create a GRanges object from Step 18.

```
genesnpranges<-GRanges(genebed$chr, IRanges(genebed$snp.start,
genebed$snp.end), ensg=genebed$ensg)
```

20. Overlap unmapped SNPs from Step 12 with loop anchor 2 from Step 19. This step assigns SNPs (located at loop anchor 2) to the genes they interact with (located at loop anchor 1).

```
olap <- findOverlaps(snpranges,genesnpranges)
snpint <- snpranges[queryHits(olap)]
mcols(snpint)<-cbind(mcols(snpranges[queryHits(olap)]),
mcols(genesnpranges[subjectHits(olap)]))
snpint <- unique(snpint)
save(snpint, file=paste0("Hi-
C_transcript_interacting_snp",".rda"))
```

21. Integrate SNP-gene relationships derived from exons, promoters, and Hi-C interaction data.

```
load("Hi-C_transcript_interacting_snp.rda")
load("snp_locating_in_exon_promoter_transcript_level.rda")
snpdat <- data.frame(chr=seqnames(snpint), bp=start(snpint),
rsid=snpint$rsid, ensg=snpint$ensg)
snpromat<-unique(data.frame(rsid=snpro$rsid, ensg=snpro$gene))
snpexonmat<-unique(data.frame(rsid=snpexon$rsid,
ensg=snpexon$gene))
snpcomb <- unique(rbind(snpdat[,3:4], snpromat, snpexonmat))
save(snpcomb, file="SNP_to_transcript_comb",".rda")
```

Create the H-MAGMA-compatible variant-gene annotation file

Timing 15 - 20 min

22. Aggregate SNP-gene relationship to generate the variant-gene annotation file compatible with MAGMA.

```
snpagg <- aggregate(snpcomb, list(snpcomb$ensg), unique)
```

23. Read in the gene definition file.

```
genedef <- read.table("~/work/Gencode26_gene.bed")
colnames(genedef) <- c("chr", "start", "end", "ensg")
```

24. Create an index column from the gene definition file consisting of gene chromosomal location, start, and end.

```
genedef <- genedef [grep("chr", genedef$chr),]
genedef$chr<-unlist(lapply(strsplit(genedef$chr, "chr"), '[[', 2))
genedef$index<- paste(genedef$chr, genedef$start, genedef$end, sep=":")
```

25. Attach the index column from Step 24 to the variant-gene annotation file from Step 22.

```
snpagg$index<-genedef[match(snpagg$ensg, genedef$ensg),"index"]
```

26. Remove any missing values from the variant-gene annotation file.

```
snpagg <- snpagg[!is.na(snpagg$index),]
```

27. Subset gene, gene location, and SNPs from the variant-gene annotation file.

```
snpannot <- snpagg[,c("ensg", "index", "rsid")]
```

28. Save the variant-gene annotation file in an executable format.

```
writable <-  format(snpannot)
```

```
write.table(writable,file="SNP_aggregate_transcript.txt",
quote=F, row.names=F, col.names=F, sep="\t")
system(paste0("sed-e's/,/\t/g'<
SNP_aggregate_transcript",".txt
>","Adultbrain.transcript.annot"))
```

Run H-MAGMA.

Timing 30 - 50 min

29. We can now run H-MAGMA using the variant-gene annotation file generated in the previous steps. As an example for this protocol, we will use GWAS summary statistics of PD [21]. Verify that all necessary files to run H-MAGMA (v1.09), including the program files that best fit your operating system, have been downloaded from the MAGMA website using the following link: https://ctg.cncr.nl/software/magma. Download all necessary files into a subfolder under the main working directory. This should appear as the directory below:

    `~/work/magma1.9/magma`

    The MAGMA subfolder should contain the following items.

    - A CHANGELOG that describes the MAGMA version. We will run MAGMA version 1.09.
    - Executable magma program
    - A magma.log file that details the executed code including date and time lapsed.
    - A manual describing the software
    - A README file

30. Run H-MAGMA in linux using the command below. H-MAGMA requires reference genome data which is denoted here with `--bfile ~/work/magma1.9/g1000_eur/g1000_eur` to account for linkage disequilibrium between SNPs. Here, we use the European reference genome, but readers may replace it with a different ancestry. The GWAS summary statistics file is placed at the `--pval flag` with the `use and ncol` parameters modified according to the column names of the GWAS summary statistics. The H-MAGMA variant-gene annotation file generated through the Steps 1-28 is placed at the `--gene-annot` flag. Lastly, the output file name to write gene-level association statistics is placed after the `--out` flag.

```
~/work/magma1.9/magma
--bfile ~/work/magma1.9/g1000_eur/g1000_eur
--pval ~/work/PD.summary.stat.txt  use=rs,p ncol=N
--gene-annot ~/work/Adultbrain.transcript.annot
--out ~/work/PD_GWAS
```

CRITICAL STEP. MAGMA outputs a .log file after each run session. We advise taking a look at the log file to ensure the program runs in the manner it is expected to.

31. Retrieve the number of genes associated with PD at different thresholds after multiple corrections by running the following commands. Genes can be stratified into either protein-coding or non-coding genes.

```
options(stringsAsFactors=F)
setwd ("~/work/")
```

```
load("geneAnno_allgenes.rda")
backgroundset<-
unique(geneAnno1[geneAnno1$gene_biotype=="protein_coding",
"ensembl_gene_id"])
diseasename <- c("PD")
fdrdisease <- c()
diseasemat <- read.table("PD_GWAS.genes.out", header=T)
diseasemat <- diseasemat[diseasemat$GENE %in% backgroundset, ]
diseasemat$FDR <- p.adjust(diseasemat$P, "BH")
backgroundensg <- diseasemat$GENE
queryensg0 <- diseasemat[diseasemat$FDR<0.1, "GENE"]
queryensg1 <- diseasemat[diseasemat$FDR<0.05, "GENE"]
queryensg2 <- diseasemat[diseasemat$FDR<0.01, "GENE"]
fdrgene<-c(diseasename,length(queryensg0),length(queryensg1),
length(queryensg2))
fdrdisease <- rbind(fdrdisease, fdrgene)
colnames(fdrdisease)<-c("disease", "FDR<0.1", "FDR<0.05",
"FDR<0.01")
fdrdisease <- data.frame(fdrdisease)
write.csv(fdrdisease,file="PD.genes.csv",col.names=T,
row.names=F, sep="\t", quote=F)
```

The commands above provide the number of protein-coding genes at different thresholds. To retrieve the number of all genes at the same threshold, simply comment out the command line as displayed below.

```
#diseasemat<-diseasemat[diseasemat$GENE%in%
backgroundset, ]. This will treat the command line
as a comment and prevent it from being included
from the rest of the analysis thereby generating a
list of all genes rather than just protein-coding
genes.
```

32. Retrieve a list of PD risk genes at FDR<0.05 in HGNC symbols by entering the command below

```
pd. genes<-unique (geneAnno1[match(queryensg1,
geneAnno1$ensembl_gene_id), "hgnc_symbol"])
```

33. Generate a bar plot to compare the number of all genes versus protein-coding genes at different thresholds using the commands below.

```
library(tidyr)
library(ggplot2)
df <- read.csv ("/work/PD.genes.csv", header = T)
df$Threshold <- row.names(df)
lab <- c("FDR<0.1","FDR<0.05","FDR<0.01")
df_long <- gather (df, key = var, value = value,
All_genes,PCG)
ggplot(df_long, aes(x = Disorder, y = value, fill = var)) +
```

```
geom_bar(stat = 'identity', position = 'dodge') +
scale_x_discrete(labels = lab)+
geom_text(aes(label=value), vjust=1.6, color="black") +
scale_fill_manual(values = c("#00AFBB", "#CC79A7"))
```

To generate the bar plot, you must first create a .csv file containing the number of all genes and protein-coding genes from Step 31. The file should be structured to have the numbers of all genes and protein-coding genes associated with PD as column names and the three FDR thresholds as row names as shown in (Fig. 4.2)**.**

Troubleshooting

Table 4.6 displays solutions to some common errors that might occur while implementing the protocol.

Timing

The total time to follow the protocol is about 2 hours. However, you should allow additional time to configure all necessary files needed to successfully implement the protocol. The step most likely to require additional time is running H-MAGMA using GWAS summary statistics due to the varying sizes of GWAS summary statistics. Overall, generation of the H-MAGMA variant-gene annotation file should take about 90 minutes. An additional 30 minutes are required to run H-MAGMA for the GWAS of interest.

Anticipated results

This protocol yields a variant-gene annotation file to run H-MAGMA that assigns genetic variants to cognate genes based on chromatin architecture. Application of H-

MAGMA to GWAS summary statistics will result in four output files (*.log.suppl, .genes.out, .genes.raw, and a .log*). The *.genes.raw* file can be subsequently used for pathway analysis. Of particular interest to this protocol is the *.genes.out* file which contains the gene-level summary statistics for the trait of interest. Within the *.genes.out* file, users will find features including gene identifications in the ENSEMBL format and gene-level P-values which can be corrected using multiple testing corrections to derive significant risk genes for a given trait. Once risk genes are identified, users can further characterize risk genes by leveraging resources such as gene ontologies to characterize the biological processes associated with the trait and transcriptomic datasets to investigate either developmental or cellular expression profiles to pinpoint important developmental periods or cell types associated with the trait[86]. It is important to note that the threshold used and sample size of the GWAS might impact the number of identifiable risk genes (Fig. 4.2). In particular, a smaller number of risk genes may be identified for less powered GWAS compared to well powered GWAS5. We recommend users adjust the threshold according to the sample size of the GWAS.

Data availability

All required data to run the protocol are publicly available. Downloadable versions of files are also provided in the Zenodo repository at https://doi.org/10.5281/zenodo.5503876[28] .In addition to the variant-gene annotation file generated from adult brain Hi-C data[34], we have uploaded variant-gene annotation files generated from 28 different cell and tissue types using promoter-capture Hi-C data from Jung et al. 2019[32]. All 28 variant-gene annotation files including commands used to generate these files are also available in the Zenodo repository. All files including source code and documentation to run MAGMA are available on the MAGMA website using this link: https://ctg.cncr.nl/software/magma

Figures

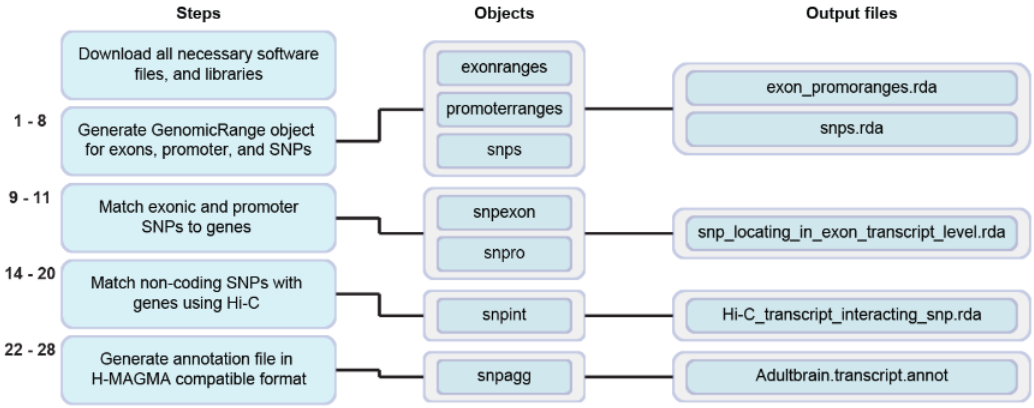


Figure 4.1. Schematic of the protocol. The “Steps" column represents the main steps of the H-MAGMA protocol. For each step, the corresponding step number in the protocol is described on the left. The “Objects" and “Output files" columns represent R objects and output files created from the corresponding step, respectively. For example, Steps 1-8 generate GRange objects for exons, promoters, and SNPs, under the names of *exonranges*, *promoterranges*, and *snps*, respectively. Objects *exonranges* and *promoterranges* are saved as an RData (.rda) file named *exon_promoranges.rda*, while the object *snps* is saved as an RData file named *snps.rda*.

a

Threshold	All genes	Protein-coding genes
FDR <0.1	324	228
FDR <0.05	235	159
FDR <0.01	145	99

b

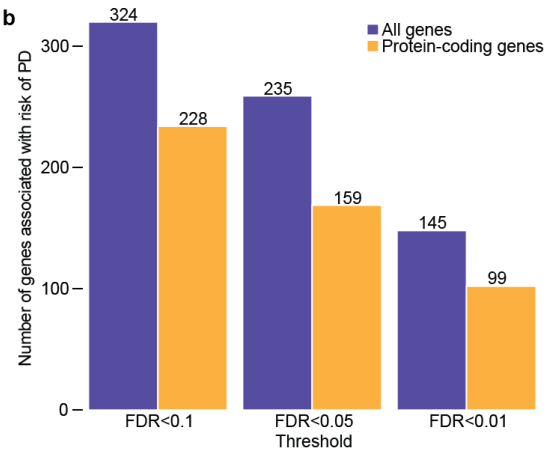


Figure 4.2. Number of PD risk genes at different FDR thresholds. (a) A table display of the number of PD risk genes at varying FDR thresholds. **(b)** Barplots of the number of risk genes associated with PD at different FDR thresholds.

Tables

Table 4.1. Exonic and promoter coordinates corresponding to steps 4 and 5 of Chapter 4

	SEQNAMES <Rle>	**RANGES <IRanges>**	**STRAND <Rle>**	**GENE <Characterer>**
[1]	chrX	99883667-99884983	*	ENSG00000000003
[2]	chrX	99885756-99885863	*	ENSG00000000003
[3]	chrX	99887482-99887565	*	ENSG00000000003

Table 4.2. Exonic and promoter SNPs corresponding to steps 9 to 11 of Chapter 4.

	SEQNAMES <Rle>	RANGES <IRanges>	STRAND <Rle>	RSID <Character>	GENE <Character>
[1]	chr1	10177	*	rs367896724	ENSG00000223972
[2]	chr1	10352	*	rs555500075	ENSG00000223972
[3]	chr1	10616	*	rs376342519	ENSG00000223972

Table 4.3. Hi-C annotated SNPs corresponding to step 20 of Chapter 4.

	SEQNAMES <Rle>	RANGES <IRanges>	STRAND <Rle>	RSID <Character>	GENE <Characterer>
[1]	chr1	570638	*	rs368120791	ENSG00000197530
[2]	chr1	574151	*	rs74047006	ENSG00000197530
[3]	chr1	583483	*	rs577801075	ENSG00000197530

Table 4.4. Variant-gene annotation file corresponding to step 28 of Chapter 4.

GENE ID	GENE LOCATION	VARIANT ID
ENSG00000000419	20:49551404:49575092	rs6126129, rs7274624, rs6096200, …
ENSG00000000457	1:169818772:169863100	rs11577641, rs10800486, rs3766155, …
ENSG00000000460	1:169763871:169823221	rs112141016, rs140700407, rs112042073, …

Table 4.5. Parkinson's disorder output file corresponding to steps 29 to 33 of Chapter 4.

PD_GWAS

GENE	CHR	START	STOP	NPARAM	N	ZSTAT	P
ENSG00000238009	1	89295	133723	1	2137	1.3031	0.09627
ENSG00000239945	1	89551	91105	1	2137	1.3031	0.09627
ENSG00000228327	1	661265	714006	1	12776	-1.1272	0.87018

PD.genes.csv

PD Genes
HAX1
ADAM15
EFNA4

Table 4.6. Troubleshooting steps associated with Chapter 4.

Step	Problem	Possible reason	Solution
1	No such file or directory	Missing the directory with downloaded files	Change working directory to the path with downloaded files
9, 10, 14	The two combined objects have no sequence levels in common	Chromosome annotations for SNPs, exon/promoter, and Hi-C files do not match. For instance, chromosome 1 can be listed as "1" in one file, while "chr1" in another file	Make sure chromosome annotations are identical in every GRanges object
30	version `GLIBCXX_3.4.20'; `CXXABI_1.3.8'; `GLIBCXX_3.4.21' not found (required by /work/magma1.9/magma)	Missing GNU Compiler Collection (GCC).	Follow the instructions on https://gcc.gnu.org/install/ to install gcc/6.3.0
	Variable “x” not found	H-MAGMA is unable to find required columns in the summary statistics	Verify that column names from the GWAS summary statistics are correctly listed

CHAPTER 5: GENERAL DISCUSSION

In this work we presented H-MAGMA, a refined framework for gene pathway analysis, that aggregates SNP-level summary statistics into gene-level association statistics. Compared with cMAGMA, H-MAGMA links noncoding SNPs to their target genes based on functional genomic evidence and adds relevant cellular context to gene mapping by using chromatin interaction data from disease-relevant tissue and cell types. While the basic concept of mapping SNPs to genes using functional genomic resources is similar to FUMA[38], H-MAGMA leverages the MAGMA framework to obtain gene-level association statistics in a genome-wide fashion, while FUMA maps a selected set of genomic loci to target genes. Therefore, H-MAGMA can provide an attractive framework to identify genes and biological pathways for low-powered GWAS. It also allows the comparison of different GWAS to elucidate shared biological pathways.

H-MAGMA can be expanded into many different forms. For example, we decided to use MAGMA among many other tools available because it is most widely used; however, this framework is applicable to any other tools that convert SNP-level *P* values into gene-level association statistics[46]. Moreover, H-MAGMA can be built on Hi-C datasets from multiple tissue and cell types to distill biological mechanisms of any GWAS (for example, Hi-C datasets from immune cells for rheumatoid arthritis GWAS). Finally, while we primarily used Hi-C datasets to link SNPs to target genes, other functional genomics tools such as chromatin accessibility correlations and machine-learning-based enhancer–promoter predictions can be used to generate SNP–gene pairs. In fact, a similar approach using eQTL

(eMAGMA) has been recently reported to detect more risk genes underlying psychiatric disorders compared to other e-QTL gene mapping tools[165].

In Chapter 2, we demonstrated the effectiveness of H-MAGMA in comparison to MAGMA by applying H-MAGMA built from fetal and adult brain Hi-C profiles to GWAS summary statistics of five psychiatric disorders (SCZ, ASD, ADHD, BD, MDD) and four neurodegenerative disorders (AD, PD, MS, ALS), respectively. Our finding confirmed that H-MAGMA can detect more disease-associated risk genes underlying psychiatric and neurodegenerative disorders by primarily linking non-coding variants to their target genes based on functional genomic evidence. Importantly, we also noted that risk genes identified by H-MAGMA explained a significant proportion of the heritability underlying brain disorders, suggesting the significance of the noncoding genome in explaining disease etiology[7,153,166]. After identifying brain disorder-associated risk genes, we next investigated their functional impact by exploring developmental windows, biological processes and cellular expression profiles. Using these approaches, we identified that psychiatric disorder-associated genes exhibit a prenatal enrichment compared to the postnatal enrichment for neurodegenerative disorder-associated genes which supports the early onset of diagnosis for psychiatric disorders compared to neurodegenerative disorders[167,168]. Additionally, biological processes underlying brain disorders included transcriptional regulators, synaptic transmission, and immune processes. Lastly, cellular expression profile of the risk genes hinted to the role of excitatory neurons as the primary cell type in understanding psychiatric disorders which was in contrast to the non-neuronal enrichment for neurodegenerative disorders.

Driven by the findings in Chapter 2, we further expanded H-MAGMA beyond homogenate tissue to specific cell types in the brain in Chapter 3. Here, we developed H-MAGMA from cortical and dopaminergic neurons, two neuronal subtypes critical to understanding substance use vulnerability. We hypothesized that using H-MAGMA built from chromatin interaction profiles from cortical and dopaminergic neurons will allow us to detect risk genes underlying cigarette smoking and alcohol use traits in a relevant biological context. We then characterized risk genes by identifying their biological and cellular functions using gene ontology analysis and specific cell types enriched for each trait. We denoted that risk genes underlying cigarette smoking and alcohol use traits are involved with stress response, learning or memory, and protein folding. Interestingly, we identified biological functions relating to other drugs of abuse not explored as part of this chapter, such as response to morphine and cocaine. Prior research among substance use traits have elucidated a strong comorbidity among multiple substance use traits, instigating a shared genetic signal among drug use[24]. However, not all substance use traits have a well-powered GWAS for downstream analyses. For context, the most current GWAS of cocaine use disorder reported genetic variants in a sample of 9965 individuals[169] . Therefore, prompted by these findings, we characterized the shared genetic architecture among substance use traits by generating a list of shared genes between cigarette smoking and alcohol use traits. We found the shared genes to play a role in synaptic functioning including synapse organization and structure. Additionally, they were enriched for GABAergic neurons and overlapped with cocaine DEGs, suggesting that shared genes derived from cigarette smoking and alcohol use may provide additional insight into the neurobiological mechanisms associated with multiple substance use traits.

Lastly, to contribute further to the field, we provide a detailed protocol to users on how to generate the H-MAGMA variant-gene annotation file by generating their own Hi-C libraries or using publicly available datasets. We also provide additional H-MAGMA variant-gene annotation files beyond brain cell types using data from Jung et al. to generate annotation files for additional 28 cell types[32].

Limitations of H-MAGMA

While H-MAGMA wildly improves on MAGMA, it is important to draw attention to lingering limitations of both tools. For example, an important limitation of H-MAGMA that should be taken into consideration is that, while H-MAGMA detects risk genes associated with a trait, it cannot determine the directionality of the effects of risk genes, such that it cannot detect whether risk genes may be upregulated or downregulated in the diseased state. This limitation can be remedied by incorporating gene expression datasets such as expression Quantitative Trait Loci (eQTL). For instance, when an eQTL is detected for an H-MAGMA associated gene in a matching tissue, the eQTL can determine whether the risk allele of the variant is associated with upregulation or downregulation of the corresponding gene.

In addition, due to the confounding effects of linkage disequilibrium (LD) in GWAS findings, not all risk genes identified by H-MAGMA are necessarily implicated with the trait. Thus, it is important to follow-up with functional validation experiments to prioritize risk genes. For instance, high-throughput techniques such as Massively Parallel Reporter Assays (MPRA) may be used to functionally validate the regulatory effects of risk variants in the relevant tissue or cell type[170].

Moreover, given that the sample size of a GWAS is closely associated with its statistical power, users should note that the number of risk genes identified by H-MAGMA

may be impacted by the power of the specific GWAS. For instance, the number of risk genes identified for Autism Spectrum Disorder in Chapter 2 and Nicotine Dependence in Chapter 3 were smaller compared to the other psychiatric disorders and substance use phenotypes discussed in both chapters, respectively, due to the sample size of the Autism and Nicotine Dependence GWAS. Therefore, it is important to factor in the possible effect of sample sizes when selecting parameters such as FDR thresholds for downstream analysis of risk genes. Additionally, the quality of the Hi-C dataset may affect the robustness of variant-gene annotation. Thus, we recommend users pay particular attention to quality control measures of Hi-C data such as the read depth, cis-to-trans ratio, and its relationship with other functional genomic data (e.g. whether enhancer-promoter interactions closely align with gene expression).

Lastly, due to the lack of diversity in genetic studies and overrepresentation of European ancestry in genetic studies, we applied H-MAGMA to GWAS from European ancestry in Chapters 2 through 4 which may limit its relevance to other populations. Given that representation in genetic studies may bolster our understanding of disease etiology, it is important to expand GWAS beyond European ancestry in order to fully benefit from the technique[171,172]. Consequently, H-MAGMA can be modified to identify risk genes from GWAS from other ancestries by incorporating a definition of sub-populations when running using the tool. However, it is important to also note that the number of identifiable genes for other ancestries may be smaller in comparison to European ancestry due to their smaller sample size.

Future directions

While our analyses provide insights into the genes and potential biological mechanisms underlying brain disorders, part of our findings warrants further investigation. For instance, while pleiotropic genes between psychiatric disorders discussed in Chapter 2 provides insights into the shared etiology among psychiatric disorders, the rate of misdiagnosis of psychiatric disorders, especially between Schizophrenia and Bipolar disorder remains high due to overlap in their symptoms[173,174]. Therefore, it would be equally informative to parse out biological characteristics that differentiates the disorders from each other to better understand their unique biomarkers to improve diagnosis. Additionally, overrepresentation of substance use among individuals with psychiatric disorders has been widely reported. Previous research has identified high levels of alcohol use among individuals with anxiety and mood disorders[175,176] while others have found high prevalence of nicotine use among Schizophrenic patients[177,178]. There is future work to be pursued into the comorbidity between substance use and psychiatric disorders which could include generating a list of shared genes between psychiatric disorders and substance use phenotypes to further probe their biological characteristics and critical cell types.

Lastly, while our analysis in Chapter 3 prioritizes the role of neuronal cells in substance use, non-neuronal cells including astrocytes and microglia have been shown to modulate substance use. For instance, alcohol has been shown to differentially affect astrocyte activity in various brain regions[179]. Indeed, our heritability enrichment in Chapter 3 showed enrichment for astrocytes for all traits. Given these findings and the clear role of non-neuronal cells in substance use, it would be informative to link substance use variants to risk genes based on H-MAGMA derived from glial cells to further investigate their role in

substance use vulnerability. Collectively, the work presented in this dissertation underscores the biologically relevant information that can be obtained from GWAS results after identifying trait-associated genes via H-MAGMA.

REFERENCES

1. Substance Abuse and Mental Health Services Administration. (2019). *Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health* (HHS Publication No. PEP19-5068, NSDUH Series H-54). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.

2. Lake, J. & Turner, M. S. Urgent Need for Improved Mental Health Care and a More Collaborative Model of Care. *Perm. J.* **21**, 17–024 (2017).

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

4. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).

5. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).

6. Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).

7. Hrdlickova, B., de Almeida, R. C., Borek, Z. & Withoff, S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta* **1842**, 1910–1922 (2014).

8. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).

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

10. Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).

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

12. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).

13. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2020).

14. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).

15. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

16. Brainstorm Consortium *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**, (2018).

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

18. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).

19. 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).

20. 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).

21. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

22. Van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).

23. Andlauer, T. F. M. *et al.* Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Sci Adv* **2**, e1501678 (2016).

24. Meyers, J. L. & Dick, D. M. Genetic and environmental risk factors for adolescent-onset substance use disorders. *Child Adolesc. Psychiatr. Clin. N. Am.* **19**, 465–477 (2010).

25. Kranzler, H. R. *et al.* Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* **10**, 1499 (2019).

26. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).

27. Quach, B. C. *et al.* Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. *Nat. Commun.* **11**, 1–13 (2020).

28. Zhou, H. *et al.* Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat. Neurosci.*

23, 809–818 (2020).

29. Koob, G. F. & Volkow, N. D. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry* **3**, 760–773 (2016).

30. Zhang, W.-H. *et al.* Role of prefrontal cortex in the extinction of drug memories. *Psychopharmacology (Berl.)* **236**, 463–477 (2019).

31. Chen, W. *et al.* Role of dopamine signaling in drug addiction. *Curr. Top. Med. Chem.* **17**, 2440–2455 (2017).

32. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature Genetics* vol. 51 1442–1449 (2019).

33. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).

34. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).

35. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

36. Roadmap Epigenomics, C. *et al.* Heravi-428 Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human 429 epigenomes. *Nature* **518**, 317–330 (2015).

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

38. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

39. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).

40. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

41. Gusev, A. *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (2018).

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

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

(2016).

44. Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address: plee0@mgh.harvard.edu & Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469–1482.e11 (2019).

45. Schork, A. J. *et al.* A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).

46. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat. Neurosci.* **18**, 199–209 (2015).

47. Marín, O. Developmental timing and critical windows for the treatment of psychiatric disorders. *Nat. Med.* **22**, 1229–1238 (2016).

48. Wegiel, J. *et al.* The neuropathology of autism: defects of neurogenesis and neuronal migration, and dysplastic changes. *Acta Neuropathol.* **119**, 755–770 (2010).

49. Zoghbi, H. Y. & Bear, M. F. Synaptic dysfunction in neurodevelopmental disorders associated with autism and intellectual disabilities. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).

50. Deutsch, S. I., Urbano, M. R., Burket, J. A., Herndon, A. L. & Winebarger, E. E. Pharmacotherapeutic Implications of the Association Between Genomic Instability at Chromosome 15q13.3 and Autism Spectrum Disorders. *Clinical Neuropharmacology* vol. 34 203–205 (2011).

51. Muller, C. L., Anacker, A. M. J. & Veenstra-VanderWeele, J. The serotonin system in autism spectrum disorder: From biomarker to animal models. *Neuroscience* **321**, 24–41 (2016).

52. Berman, J. A., Talmage, D. A. & Role, L. W. Cholinergic circuits and signaling in the pathophysiology of schizophrenia. *Int. Rev. Neurobiol.* **78**, 193–223 (2007).

53. Melnikov, M., Pashenkov, M. & Boyko, A. Dopaminergic Receptor Targeting in Multiple Sclerosis: Is There Therapeutic Potential? *Int. J. Mol. Sci.* **22**, (2021).

54. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

55. Iqbal, K., Liu, F. & Gong, C.-X. Tau and neurodegenerative disease: the story so far. *Nat. Rev. Neurol.* **12**, 15–27 (2016).

56. Haass, C. & Selkoe, D. J. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. *Nat. Rev. Mol. Cell Biol.* **8**, 101–112 (2007).

57. Mulligan, K. A. & Cheyette, B. N. R. Neurodevelopmental Perspectives on Wnt Signaling in Psychiatry. *Mol Neuropsychiatry* **2**, 219–246 (2017).

58. Edmunds, S. R., Kover, S. T. & Stone, W. L. The relation between parent verbal responsiveness and child communication in young children with or at risk for autism spectrum disorder: A systematic review and meta-analysis. *Autism Research* vol. 12 715–731 (2019).

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

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

61. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *J. Cell Biol.* **217**, 459–472 (2018).

62. Clement, A. M. *et al.* Wild-type nonneuronal cells extend survival of SOD1 mutant motor neurons in ALS mice. *Science* **302**, 113–117 (2003).

63. Halliday, G. M. & Stevens, C. H. Glia: initiators and progressors of pathology in Parkinson's disease. *Mov. Disord.* **26**, 6–17 (2011).

64. Ortiz, G. G. *et al.* Role of the Blood–Brain Barrier in Multiple Sclerosis. *Arch. Med. Res.* **45**, 687–697 (2014).

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

66. Rajarajan, P. *et al.* Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* **362**, (2018).

67. Clarke, L. E. *et al.* Normal aging induces A1-like astrocyte reactivity. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1896–E1905 (2018).

68. Brambilla, R. The contribution of astrocytes to the neuroinflammatory response in multiple sclerosis and experimental autoimmune encephalomyelitis. *Acta Neuropathol.* **137**, 757–783 (2019).

69. Plaisier, S. B., Taschereau, R., Wong, J. A. & Graeber, T. G. Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res.* **38**, e169–e169 (2010).

70. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

71. 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).

72. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).

73. Mohammadi, S., Davila-Velderrain, J. & Kellis, M. Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution. *Cell Syst* **9**, 559–568.e4 (2019).

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

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

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

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

78. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).

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

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

81. Abuse. … Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health (HHS Publication No …. *Substance Abuse and Mental Health Services*.

82. Peacock, A. *et al.* Global statistics on alcohol, tobacco and illicit drug use: 2017 status report. *Addiction* **113**, 1905–1926 (2018).

83. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. (Centers for Disease Control and Prevention (US), 2014).

84. Dekker, J. Gene regulation in the third dimension. *Science* **319**, 1793–1794 (2008).

85. Mah, W. & Won, H. The three-dimensional landscape of the genome in human brain tissue unveils regulatory mechanisms leading to schizophrenia risk. *Schizophr. Res.* **217**, 17–25 (2020).

86. Sey, N. Y. A. *et al.* A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat.*

Neurosci. (2020) doi:10.1038/s41593-020-0603-0.

87. Lammel, S., Lim, B. K. & Malenka, R. C. Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology* **76 Pt B**, 351–359 (2014).

88. Hu, B. *et al.* Neuronal and glial 3D chromatin architecture informs the cellular etiology of brain disorders. *Nat. Commun.* **12**, 3968 (2021).

89. Espeso-Gil, S. *et al.* A chromosomal connectome for psychiatric and metabolic risk variants in adult dopaminergic neurons. *Genome Med.* **12**, 19 (2020).

90. Consortium, Roadmap Epigenomics *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

91. Berke, J. D. & Hyman, S. E. Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* **25**, 515–532 (2000).

92. Zhang, S. *et al.* Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science* vol. 369 561–565 (2020).

93. Nott, A. *et al.* Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).

94. Stark, R., Brown, G. & Others. DiffBind: differential binding analysis of ChIP-Seq peak data. *R package version* **100**, 4–3 (2011).

95. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999–1014.e22 (2018).

96. Metzakopian, E. *et al.* Genome-wide characterization of Foxa2 targets reveals upregulation of floor plate genes and repression of ventrolateral genes in midbrain dopaminergic progenitors. *Development* **139**, 2625–2634 (2012).

97. Lee, H.-S. *et al.* Foxa2 and Nurr1 synergistically yield A9 nigral dopamine neurons exhibiting improved differentiation, function, and cell survival. *Stem Cells* **28**, 501–512 (2010).

98. Saucedo-Cardenas, O. *et al.* Nurr1 is essential for the induction of the dopaminergic phenotype and the survival of ventral mesencephalic late dopaminergic precursor neurons. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4013–4018 (1998).

99. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

100. Simon, H. H., Saueressig, H., Wurst, W., Goulding, M. D. & O'Leary, D. D. Fate of midbrain dopaminergic neurons controlled by the engrailed genes. *J. Neurosci.* **21**, 3126–3134 (2001).

101. Palmer, A. A., Low, M. J., Grandy, D. K. & Phillips, T. J. Effects of a Drd2 deletion mutation on ethanol-induced locomotor stimulation and sensitization suggest a role for epistasis. *Behav. Genet.* **33**, 311–324 (2003).

102. Herman, A. I., DeVito, E. E., Jensen, K. P. & Sofuoglu, M. Pharmacogenetics of nicotine addiction: role of dopamine. *Pharmacogenomics* **15**, 221–234 (2014).

103. June, H. L. *et al.* Dopamine and Benzodiazepine-Dependent Mechanisms Regulate the EtOH-Enhanced Locomotor Stimulation in the GABAA α1 Subunit Null Mutant Mice. *Neuropsychopharmacology* vol. 32 137–152 (2007).

104. Jeanblanc, J. *et al.* Endogenous BDNF in the dorsolateral striatum gates alcohol drinking. *J. Neurosci.* **29**, 13494–13502 (2009).

105. Zhou, Z., Enoch, M.-A. & Goldman, D. Gene expression in the addicted brain. *Int. Rev. Neurobiol.* **116**, 251–273 (2014).

106. Semick, S. A. *et al.* Developmental effects of maternal smoking during pregnancy on the human frontal cortex transcriptome. *Mol. Psychiatry* **25**, 3267–3277 (2018).

107. Jensen, K. P. *et al.* Alcohol-responsive genes identified in human iPSC-derived neural cultures. *Transl. Psychiatry* **9**, 96 (2019).

108. Skorput, A. G. J., Gupta, V. P., Yeh, P. W. L. & Yeh, H. H. Persistent Interneuronopathy in the Prefrontal Cortex of Young Adult Offspring Exposed to Ethanol In Utero. *J. Neurosci.* **35**, 10977–10988 (2015).

109. Kazemi, T. *et al.* Investigating the influence of perinatal nicotine and alcohol exposure on the genetic profiles of dopaminergic neurons in the VTA using miRNA–mRNA analysis. *Scientific Reports* vol. 10 (2020).

110. Fox, H. C., Milivojevic, V., Angarita, G. A., Stowe, R. & Sinha, R. Peripheral immune system suppression in early abstinent alcohol-dependent individuals: Links to stress and cue-related craving. *J. Psychopharmacol.* **31**, 883–892 (2017).

111. Pasala, S., Barr, T. & Messaoudi, I. Impact of alcohol abuse on the adaptive immune system. *Alcohol Res.* **37**, 185 (2015).

112. Díaz-Villanueva, J. F., Díaz-Molina, R. & García-González, V. Protein Folding and Mechanisms of Proteostasis. *Int. J. Mol. Sci.* **16**, 17193–17230 (2015).

113. Elman, I. & Borsook, D. Common Brain Mechanisms of Chronic Pain and Addiction. *Neuron* **89**, 11–36 (2016).

114. Goodman, J. & Packard, M. G. Memory Systems and the Addicted Brain. *Front. Psychiatry* **7**, 24 (2016).

115. Morin, J.-F. G. *et al.* A Population-Based Analysis of the Relationship Between

Substance Use and Adolescent Cognitive Development. *Am. J. Psychiatry* **176**, 98–106 (2019).

116. Elmenhorst, E.-M. *et al.* Cognitive impairments by alcohol and sleep deprivation indicate trait characteristics and a potential role for adenosine A1 receptors. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 8009–8014 (2018).

117. Xu, Z. *et al.* Cancer mortality attributable to cigarette smoking in 2005, 2010 and 2015 in Qingdao, China. *PLoS One* **13**, e0204221 (2018).

118. Matzeu, A. & Martin-Fardon, R. Drug Seeking and Relapse: New Evidence of a Role for Orexin and Dynorphin Co-transmission in the Paraventricular Nucleus of the Thalamus. *Front. Neurol.* **9**, 720 (2018).

119. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566–580.e19 (2016).

120. Morello, F. & Partanen, J. Diversity and development of local inhibitory and excitatory neurons associated with dopaminergic nuclei. *FEBS Lett.* **589**, 3693–3701 (2015).

121. Kirby, L. G., Zeeb, F. D. & Winstanley, C. A. Contributions of serotonin in addiction vulnerability. *Neuropharmacology* **61**, 421–432 (2011).

122. Gould, T. J. Nicotine and hippocampus-dependent learning. *Mol. Neurobiol.* **34**, 93–107 (2006).

123. Zhu, Y., Wienecke, C. F. R., Nachtrab, G. & Chen, X. A thalamic input to the nucleus accumbens mediates opiate dependence. *Nature* **530**, 219–222 (2016).

124. Abuse, S. Mental Health Services Administration.(2018). Key substance use and mental health indicators in the United States: Results from the 2017 National Survey on Drug Use and Health (HHS Publication No. SMA 18-5068, NSDUH Series H-53). Rockville, MD: Center for Behavioral Health Statistics and Quality. *Substance Abuse and Mental Health Services Administration. Retrieved from https://www. samhsa. gov/data* (2019).

125. Savell, K. E. *et al.* A dopamine-induced gene expression signature regulates neuronal function and cocaine response. *Sci Adv* **6**, eaba4221 (2020).

126. Hendershot, C. S., Wardell, J. D., Samokhvalov, A. V. & Rehm, J. Effects of naltrexone on alcohol self-administration and craving: meta-analysis of human laboratory studies. *Addiction Biology* vol. 22 1515–1527 (2017).

127. Stead, L. F. *et al.* Nicotine replacement therapy for smoking cessation. *Cochrane Database Syst. Rev.* **11**, CD000146 (2012).

128. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–7 (2016).

129. Castillo-Carniglia, A., Keyes, K. M., Hasin, D. S. & Cerdá, M. Psychiatric comorbidities in alcohol use disorder. *Lancet Psychiatry* **6**, 1068–1080 (2019).

130. Murthy, P., Mahadevan, J. & Chand, P. K. Treatment of substance use disorders with co-occurring severe mental health disorders. *Curr. Opin. Psychiatry* **32**, 293–299 (2019).

131. Hartz, S. M. *et al.* Association Between Substance Use Disorder and Polygenic Liability to Schizophrenia. *Biol. Psychiatry* **82**, 709–715 (2017).

132. Chang, L.-H. *et al.* Associations between polygenic risk for tobacco and alcohol use and liability to tobacco and alcohol use, and psychiatric disorders in an independent sample of 13,999 Australian adults. *Drug and Alcohol Dependence* vol. 205 107704 (2019).

133. Hoffman, J. L. *et al.* Alcohol drinking exacerbates neural and behavioral pathology in the 3xTg-AD mouse model of Alzheimer's disease. *Int. Rev. Neurobiol.* **148**, 169–230 (2019).

134. Nicholatos, J. W. *et al.* Nicotine promotes neuron survival and partially protects from Parkinson's disease by suppressing SIRT6. *Acta Neuropathologica Communications* vol. 6 (2018).

135. Piao, W.-H. *et al.* Nicotine and inflammatory neurological disorders. *Acta Pharmacologica Sinica* vol. 30 715–722 (2009).

136. Bush, T., Lovejoy, J. C., Deprey, M. & Carpenter, K. M. The effect of tobacco cessation on weight gain, obesity, and diabetes risk. *Obesity* vol. 24 1834–1841 (2016).

137. Germeroth, L. J. & Levine, M. D. Postcessation weight gain concern as a barrier to smoking cessation: Assessment considerations and future directions. *Addict. Behav.* **76**, 250–257 (2018).

138. McCrory, E. J. & Mayes, L. Understanding Addiction as a Developmental Disorder: An Argument for a Developmentally Informed Multilevel Approach. *Current Addiction Reports* vol. 2 326–330 (2015).

139. Powell, S. K. *et al.* Induction of dopaminergic neurons for neuronal subtype-specific modeling of psychiatric disease risk. *Mol. Psychiatry* (2021) doi:10.1038/s41380-021-01273-0.

140. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).

141. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

142. Kaul, A., Bhattacharyya, S. & Ay, F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat. Protoc.* **15**, 991–1012 (2020).

143. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2016).

144. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. (2010).

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

146. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

147. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

148. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).

149. Kramer, N. E. *et al.* Plotgardener: Cultivating precise multi-panel figures in R. *bioRxiv* 2021.09.08.459338 (2021) doi:10.1101/2021.09.08.459338.

150. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–200 (2007).

151. Tran, M. N. *et al.* Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. *Neuron* **109**, 3088–3103.e5 (2021).

152. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222 (2019).

153. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–10 (2015).

154. Pratt, B. M. & Won, H. Advances in profiling chromatin architecture shed light on the regulatory dynamics underlying brain disorders. *Semin. Cell Dev. Biol.* **121**, 153–160 (2022).

155. Constantino, J. N. & Marrus, N. The Early Origins of Autism. *Child Adolesc. Psychiatr. Clin. N. Am.* **26**, 555–570 (2017).

156. Yurko, R., Roeder, K., Devlin, B. & G'Sell, M. H-MAGMA, inheriting a shaky statistical foundation, yields excess false positives. *bioRxiv* (2020) doi:10.1101/2020.08.20.260224.

157. de Leeuw, C., Sey, N. Y. A., Posthuma, D. & Won, H. A response to Yurko et al: H-MAGMA, inheriting a shaky statistical foundation, yields excess false positives. 2020.09.25.310722 (2020) doi:10.1101/2020.09.25.310722.

158. Matoba, N. *et al.* Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. *Transl. Psychiatry* **10**, 265 (2020).

159. Song, M. *et al.* Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* **587**, 644–649 (2020).

160. Feleke, R. *et al.* Cross-platform transcriptional profiling identifies common and distinct molecular pathologies in Lewy body diseases. *Acta Neuropathol.* (2021) doi:10.1007/s00401-021-02343-x.

161. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).

162. Fang, R. *et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).

163. Sey, N., Pratt, B. & Won, H. *H-MAGMA Protocol*. (2021). doi:10.5281/zenodo.5503877.

164. *H-MAGMA*. (Github).

165. Gerring, Z. F., Gamazon, E. R., Derks, E. M. & Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. *PLoS Genet.* **15**, e1008245 (2019).

166. Gloss, B. S. & Dinger, M. E. Realizing the significance of noncoding functionality in clinical genomics. *Exp. Mol. Med.* **50**, 1–8 (2018).

167. Solmi, M. *et al.* Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol. Psychiatry* **27**, 281–295 (2022).

168. Hou, Y. *et al.* Ageing as a risk factor for neurodegenerative disease. *Nat. Rev. Neurol.* **15**, 565–581 (2019).

169. Sun, J., Kranzler, H. R., Gelernter, J. & Bi, J. A genome-wide association study of cocaine use disorder accounting for phenotypic heterogeneity and gene–environment interaction. *J. Psychiatry Neurosci.* **45**, 34–44 (2020).

170. Mulvey, B., Lagunas, T., Jr & Dougherty, J. D. Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants Across Biological Contexts. *Biol. Psychiatry* **89**, 76–89 (2021).

171. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse

Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).

172. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080 (2019).

173. Woolley, J. D., Khan, B. K., Murthy, N. K., Miller, B. L. & Rankin, K. P. The diagnostic challenge of psychiatric symptoms in neurodegenerative disease: rates of and risk factors for prior psychiatric diagnosis in patients with early neurodegenerative disease. *J. Clin. Psychiatry* **72**, 126–133 (2011).

174. Singh, T. & Rajput, M. Misdiagnosis of bipolar disorder. *Psychiatry* **3**, 57–63 (2006).

175. Ummels, S. A. *et al.* The bidirectional relationship between anxiety disorders and alcohol use disorders in adults: Findings from a longitudinal population-based study. *J. Affect. Disord.* **314**, 126–132 (2022).

176. Lai, H. M. X., Cleary, M., Sitharthan, T. & Hunt, G. E. Prevalence of comorbid substance use, anxiety and mood disorders in epidemiological surveys, 1990–2014: A systematic review and meta-analysis. *Drug Alcohol Depend.* **154**, 1–13 (2015).

177. Chen, J. *et al.* Genetic Relationship between Schizophrenia and Nicotine Dependence. *Sci. Rep.* **6**, 25671 (2016).

178. Isuru, A. & Rajasuriya, M. Tobacco smoking and schizophrenia: re-examining the evidence. *BJPsych Advances* **25**, 363–372 (2019).

179. Lacagnina, M. J., Rivera, P. D. & Bilbo, S. D. Glial and Neuroimmune Mechanisms as Critical Modulators of Drug Use and Abuse. *Neuropsychopharmacology* **42**, 156–177 (2017).