Single-cell expression quantitative trait loci in brain health and disease: a framework for drug target and biomarker discovery

Alexander Pierre Jens Haglund

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Department of Brain Sciences

Imperial College London

October 2024

Abstract

Disorders of the brain and nervous system are a leading cause of disability worldwide, spanning a wide range of phenotypes across the neurological, psychiatric, behavioural and structural spectrum. In an otherwise complicated model where phenotypic presentation is the result of complex interplay between genetics and environment, the study of their genetic architecture provides an anchor point to understand the relationships between genetic variation, phenotypic risk and the molecular mechanisms between them. In the past twenty years, genome-wide association studies (GWAS) combined with molecular quantitative trait loci (QTL) projects have identified thousands of genetic variants and the genes they regulate associated with brain-related phenotypes. Only recently has it been possible to profile molecular traits at the single-cell level, providing a new and context-specific dimension to study these relationships.

In **Chapter 3**, I present one of the largest expression quantitative trait loci (eQTL) studies to date in the human brain, profiled from single-nuclei to identify cell-type specific patterns of genetic regulation. I show that a substantial proportion of eQTLs are cell-type specific, outperforming traditional bulk-tissue eQTL studies both in terms of total discovery and biological relevance. I further demonstrate in **Chapter 4** that these eQTLs are dynamic, their association affected by both the presence of disease pathology as well as age. Brain phenome-wide genetic colocalisation **Chapter 5** shows a high degree of cell-type specificity associated with trait pre-disposition and that even colocalisation patterns are dynamic as a result of disease pathology. Finally, using samples unbiased by disease presence I show in **Chapter 6** the first results of Mendelian Randomisation (MR) for causal inference in the brain anchored in cell-type specific eQTLs, identifying putative novel drug targets and biomarkers for brain health and disease.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Statement of originality

I, Alexander Haglund, declare that the work presented in this thesis entitled "Single-cell expression quantitative trait loci in brain health and disease: a framework for drug target and biomarker discovery" is my own and has not been submitted for any other degree at any other university. Any data used in this work or content presented in this thesis that is not my own has been acknowledged accordingly.

Acknowledgements

I would like to express my sincere gratitude to everyone who has accompanied me on this journey.

First and foremost I would like to thank my supervisors. Professor Michael Johnson, whose never-ending enthusiasm and blue-skies approach to science have been a constant source of inspiration during my time at Imperial, both within and outside of my PhD. Dr. Leonardo Bottolo, for his guidance throughout my PhD and his patience in introducing me to the world of statistics for which I will be eternally grateful. Thank you both for giving me the opportunity to embark on these amazing projects, for your belief in my abilities and for shaping me into the scientist I am today. Dr. Prashant Srivastava and Dr. Verena Zuber, thank you for the helpful discussions in our lab meetings and the advice that have helped shape my PhD. Thank you also to my examiners throughout the milestones of this PhD, Dr. Ann Babtie and Professor Boris Lenhard for their constructive feedback in the early stages of this journey.

To everyone in the lab, who has supported me throughout; To Liisi, thank you for dragging me into the world of bioinformatics and bringing me into the Johnson lab - you were a fantastic MRes supervisor. Hun and Dan, thank you for being the most amazing lab scientists and for the outstanding data you have generated. The results of this thesis lie on top of your expertise and could not exist without your efforts. I want to thank Rahel and Maya, my co-PhD students with whom I have shared the highs and lows of the PhD journey (and the occasional korean barbecues). I could not have asked for better colleagues and friends to share this experience with.

A big thank you to my friends who have often provided an escape, and for listening to me complain about my PhD for the past four years; Karl and Joe for being my best friends through it all, always a source of laughter and support. Max and Jac for accompanying me on the tumultuous journey of my undergraduate degree and beyond, and our many (many many) hang out sessions as housemates. My MRes friends, Foteini, Morgan, Ella, Tom, Maddie, Natalia, thank you for the great times and for an amazing first year at Imperial.

To Myrsini, my partner who was on her own PhD journey alongside mine - I could not have done this without you. Thank you for being my unending source of inspiration, both on this scientific journey and outside of it; for being my emotional support through the highs and lows of this PhD journey and the path that led to it; and for your unreserved love and patience through it all. I am lucky to thank my loving parents, Jesper and Sandrine, without whom this thesis would not exist. They have always supported and believed in me, and have given me the privilege of these opportunities for which I am incredibly fortunate. To my younger siblings, Andreas and Emma, who I have seen grow so much in the past years and have hoped to be a model big brother for. To my grandparents, Björn, Lill and Madeleine, thank you for your understanding and patience throughout these last four years, and to my grandpa Pierre who would have loved to see me embark on this journey.

Finally and most importantly I would like to thank the people and their families who have so generously donated their brains for this research - this work would not have been possible without them.

Publications

The majority of this work has been included in the following manuscript;

Haglund, A.*, Zuber, V.*, Abouzeid, M., Yang, Y., Ko, J. H., Wiemann, L., Otero Jimenez, M., Muhammed, L., Feleke, R., Nott, A., Mills, J. D., Laaniste, L., Gveric, D. O., Clode, D., Babtie, A. C., Pagni, S., Bellampalli, R., Somani, A., McDade, K., Anink, J. J., Mesarosova, L., Fancy, N., Willumsen, N., Smith, A., Jackson, J., Alegre Abarrategui, J., Aronica, E., Matthews, P. M., Thom, M., Sisodiya, S. M., Srivastava, P. K., Malhotra, D., Bryois, J., Bottolo, L., Johnson, M. R. (2024). "Cell state-dependent allelic effects and contextual Mendelian randomisation analysis for human brain phenotypes", *Nature Genetics* (in press)

Other publications include;

Ray-Jones, H., Song, Z., Chan, L. T., **Haglund, A.**, Artemov, P., Della Rosa, M., Ruje, L., Burden, F., Kreuzhuber, R., Litovskikh, A., Weyenbergh, E., Brusselaers, Z., Tan, V. X. H., Frontini, M., Wallace, C., Malysheva, V., Bottolo, L., Vigorito, E., Spivakov, M. (2024). "Shared and distinct molecular effects of regulatory genetic variants provide insight into mechanisms of distal enhancer-promoter communication", *bioRxiv*, https://doi.org/10.1101/2023.08.04.551251 (in review at *Nature Communications*)

Zuber, V., Lewin, A., Levin, M. G., **Haglund, A.**, Ben-Aicha, S., Emanueli, C., Damrauer, S., Burgess, S., Gill, D., Bottolo, L. (2023). "Multi-response Mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes", *American Journal of Human Genetics* https://doi.org/10.1016/j.ajhg.2023.06.005

Table of contents

Preface **i**

List of abbreviations **ii**

1 Introduction **1**

1.1 Epidemiology of neurological disorders 2

1.2 Current therapeutic landscape for neurological and psychiatric disorders 4

1.3 Genome-wide Association Studies (GWAS) 6

1.4 Fine-mapping methods 9

1.4.1 Functional and bioinformatic fine-mapping 10

1.4.2 Statistical fine-mapping 11

1.5 Quantitative Trait Loci; an overview 12

1.5.1 eQTLs; recent developments 14

1.5.2 Single-cell eQTLs 16

1.5.3 Advantages and limitations of single-cell eQTL approaches 18

1.5.4 QTL study design 19

1.5.5 eQTL methodology and considerations 20

1.6 Genetic colocalisation 22

1.6.1 Recent developments in genetic colocalisation 25

1.7 Mendelian Randomisation 26

1.7.1 MR in neurology 29

1.7.2 MR methodology 31

1.8 MR and colocalisation; similar goals but contrasting outcomes 33

1.9 Aims and objectives of thesis 34

2 Methods **36**

2.1 Materials and Methods . . . 36

2.1.1 Dataset overview . . . 36

2.1.2 Sample extraction and preparation . . . 37

2.1.3 Genotyping . . . 37

2.1.4 Generation of single-nuclei RNA-seq data . . . 38

2.1.5 Single-cell analysis and cell-type identification . . . 40

2.1.6 eQTL mapping . . . 41

2.1.7 Epigenetic overlap of eQTLs . . . 42

2.1.8 Validation of eQTLs with existing bulk data . . . 42

2.1.9 Random effect and interaction modelling . . . 43

2.1.10 GWAS processing . . . 44

2.1.11 Colocalisation analysis . . . 45

2.1.12 MR analysis . . . 46

2.1.13 Assessment of MR genes at protein level . . . 46

2.1.14 Epigenetic annotation of IVs . . . 47

2.1.15 Code availability . . . 49

3 Single-nuclei eQTL discovery in the post-mortem human brain **50**

3.1 Preface . . . 50

3.1.1 Appendices . . . 50

3.1.2 Acknowledgement of contributions . . . 50

3.2 Introduction . . . 51

3.2.1 Aims and objectives: . . . 51

3.3 Results . . . 52

3.3.1 Cohort overview . . . 52

3.3.2 Dataset characteristics . . . 54

3.3.3 *cis*-eQTL discovery overview . . . 56

3.3.4 Patterns of eQTL behaviour . . . 60

3.3.5 Comparison to bulk data . . . 63

3.4 Discussion . . . 65

3.4.1 Dataset limitations . . . 65

3.4.2 Single-cell analysis . . . 66

3.4.3 eQTL discovery . . . 67
3.4.4 Epigenetic annotation . . . 68
3.4.5 Replication value . . . 69

4 Impact of disease on cell-type specific eQTL associations 70
4.1 Preface . . . 70
4.1.1 Appendices . . . 70
4.2 Introduction . . . 70
4.2.1 Aims and objectives: . . . 72
4.3 Results . . . 73
4.3.1 Overview of methodology and random effect modelling . . . 73
4.3.2 Modelling of disease interactions . . . 75
4.3.3 Controls versus full-cohort eQTLs . . . 79
4.3.4 Age-related interaction eQTLs . . . 83
4.4 Discussion . . . 86
4.4.1 Impact of disease on eQTL associations . . . 86
4.4.2 Differences between controls and full cohort . . . 87
4.4.3 Age-related interactions . . . 88

5 Colocalisation of cell-type specific eQTLs across the brain phenome 89
5.1 Preface . . . 89
5.1.1 Appendices . . . 89
5.2 Introduction . . . 89
5.2.1 Aims and objectives: . . . 91
5.3 Results . . . 91
5.3.1 Overview of GWAS datasets . . . 91
5.3.2 Colocalisation results using the full eQTL cohort . . . 93
5.3.3 Exampes of cell-type specific colocalisations . . . 95
5.3.4 Shared patterns of colocalisation . . . 97
5.3.5 Impact of disease-interacting eQTLs on colocalisations . . . 101
5.3.6 Differential colocalisations; controls versus full cohort . . . 102
5.3.7 Interaction COLOCs not present in controls . . . 105

5.4 Discussion . . . 107
5.4.1 Choice of GWAS . . . 108
5.4.2 Cell-type specificity of colocalised genes . . . 109
5.4.3 Disease-interacting colocalisations and use of control tissue . . . 110

6 Mendelian Randomisation at single-cell level in the post-mortem brain 112
6.1 Preface . . . 112
6.1.1 Appendices . . . 112
6.2 Introduction . . . 112
6.2.1 Aims and objectives . . . 113
6.3 Results . . . 113
6.3.1 Instrumental variable (IV) selection . . . 113
6.3.2 Single-cell mendelian randomisation . . . 115
6.3.3 Multi-instrument MR . . . 119
6.3.4 Protein QTLs for validation of single-cell MR . . . 121
6.3.5 Epigenetic interrogation of MR IVs . . . 124
6.4 Discussion . . . 129
6.4.1 Cell-type specific MR . . . 129
6.4.2 Protein replications . . . 130
6.4.3 Epigenetic overlaps . . . 131

7 Discussion 133
7.1 eQTLs at single-cell level; where to next . . . 134
7.2 Why using disease cases can still be valuable . . . 136
7.3 Colocalisation and MR for putative causal inference . . . 137
7.4 Beyond colocalisation and MR . . . 139
7.5 Future work . . . 141
7.6 Concluding remarks . . . 143

References 144

List of Figures

1.1 Number of GWAS publications as of 2023 . . . 7
1.2 Overview of molecular Quantitative Trait Loci (molQTLs) . . . 13
1.3 Pseudobulking for single-cell eQTLs . . . 17
1.4 Genetic colocalisation testing . . . 24
1.5 Conceptual overview of Mendelian Randomisation . . . 27
1.6 Overview of MR model. . . . 30

3.1 Cohort overview . . . 53
3.2 Single-cell analysis . . . 55
3.3 *cis*-eQTL discovery . . . 57
3.4 Examples of cell-type specific eQTLs . . . 59
3.5 Patterns of eQTL behaviour . . . 61
3.6 Single-cell eQTLs versus bulk eQTLs . . . 64

4.1 Overview of eQTL modelling methodology . . . 74
4.2 Interaction modelling results . . . 76
4.3 Examples of interaction-eQTLs . . . 78
4.4 Controls-only eQTLs . . . 80
4.5 Impact of interactions on controls-only eQTLs . . . 82
4.6 Impact of age on eQTL associations . . . 84

5.1 Overview of GWAS included . . . 92
5.2 Overview of colocalisation results . . . 94
5.3 Examples of cell-type specific colocalisations . . . 96
5.4 Shared colocalised genes . . . 98
5.5 Colocalisations with disease interaction . . . 100

5.6 Differences in colocalisation patterns between controls and disease 103
5.7 Colocalisations driven by disease 106

6.1 Instrumental variable selection 114
6.2 Overview of MR results 116
6.3 Shared MR genes 118
6.4 Multi-instrument MR using PCA-IVW 120
6.5 Validation of MR inferences using UKB-PPP 122
6.6 Validation of MR inferences using Robins et al., brain proteomics 124
6.7 Functional assessment of IVs 125
6.8 Epigenetic overlap of IVs 128

List of Tables

1.1 Examples of sequencing costs (RNA- vs scRNA-seq) 16

2.1 Overview of snRNA-seq dataset for eQTL discovery 36
2.2 Markers used in cell-type annotation of single-nuclei data 40
2.3 Impact scoring sheet for regulatory variants 47

3.1 Number of individuals retained after pseudobulking (min. 10 cells) 56

Preface

This thesis was created with Quarto. An interactive version is available at; https://haglunda.github.io/quarto_thesis/.

In addition, all related appendices which will be referenced throughout this thesis can be browsed interactively on the following shiny server; https://ah3918.shinyapps.io/shinyapp/.

Any questions related to this thesis and the analyses within, please reach out by email at a.haglund19@imperial.ac.uk.

List of abbreviations

AD	Alzheimer's disease
ADHD	Attention deficit hyperactivity disorder
ATAC-seq	Assay for transposase-accessible chromatin sequencing
CAGE	Cap analysis of gene expression
caQTL	Chromatin accessibility quantitative trait loci
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CHD	Coronary heart disease
CI	Confidence interval
CRP	C-reactive protein
CSF	Cerebrospinal fluid
DALYs	Disability-adjusted life years
dsQTL	DNaseI sensitive quantitative trait loci
eQTL	Expression quantitative trait loci
FDA	Food and Drug Administration
FDR	False discovery rate
GBD	Global burden of disease
GRO-seq	Global run-on sequencing
GWAS	Genome-wide association studies
iPS	Induced pluripotent stem cells
IV	Instrumental variable
IVW	Inverse variance weighted
LD	Linkage disequilibrium
LME	Linear mixed effects model
LRT	Likelihood ratio test

MAF	Minor allele frequency
MCMC	Markov chain monte carlo
MDD	Major depressive disorder
meQTL	Methylation quantitative trait loci
molQTL	Molecular quantitative trait loci
MR	Mendelian randomisation
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
NGS	Next-generation sequencing
NSRA	Nascent RNA sequencing analysis
OCD	Obsessive compulsive disorder
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PD	Parkinson's disease
PC	Principal component
PIP	Posterior inclusion probability
PMI	Post-mortem Interval
PP	Posterior probability
PRO-seq	Precision nuclear run-on sequencing
pQTL	Protein quantitative trait loci
qPCR	Quantitative PCR
QTL	Quantitative trait loci
RCT	Randomized controlled trial
RNA-seq	RNA sequencing
scRNA-seq	Single-cell RNA sequencing
SNP	Single-nucleotide polymorphism
snRNA-seq	Single-nuclei RNA-seq
sQTL	Splicing quantitative trait loci
TBI	Traumatic brain injury
TSS	Transcription start site
UKB	UK Biobank
UKB-PPP	UKB Pharma Proteomics Project

VEP	Variant effect predictor
WGS	Whole genome sequencing
YLDs	Years lived with disability
YLLs	Years of life lost

Chapter 1

Introduction

The majority of brain phenotypes are complex and polygenic in nature, and genetic factors as shown from extensive twin studies play a major role in susceptibility[1]. As such, understanding the underlying genetics responsible for brain phenotypes is not only crucial for the development of targeted therapy in brain disease but also the first step in understanding behavioural and structural phenotypes more intricately.

In the last 20 years, Genome-wide association studies (GWAS) have discovered thousands of statistically significant genomic loci across the human phenome, hundreds of which have been associated with the brain[2]. Given that the majority ($> 90\%$) of these genetic variants are located in non-coding genomic regions, the field of Quantitative Trait Loci (QTL) which focuses on associating the function of variants to molecular traits has in parallel grown considerably in the hope to identify therapeutically actionable targets[3].

The aim of this chapter will be to set the context for the ensuing work undertaken in this PhD by giving an overview of the field of GWAS and QTL in the brain, the extensive arsenal of methods developed to support them, and how the single-cell revolution has benefited its research[4].

1.1 Epidemiology of neurological disorders

Neurological disease, also referred to as neurological conditions or nervous system disorders, refer broadly to conditions detrimentally affecting the nervous system. This wide umbrella encompasses many examples of entities that affect brain growth, damaging the brain or other parts of the nervous system and impairing cognitive, behavioural, sensory and motor functions; all of which fall under wide-ranging categories of neurological disorders, including but not limited to neurodevelopmental, neurodegenerative, neuromuscular, neuro-immunological disorders, including cancers, infections and traumatic injuries. These disorders vary greatly in underlying cause (genetic predisposition, for instance), symptoms, age susceptibility and course of disease. They can in some cases be fatal or cause lifelong disability, and some are treatable and/or preventable while others are incurable, where the only option is palliative treatment.

The global burden of disease study (GBD) is the most comprehensive epidemiological study to date, employing a global approach to understand the incidence, prevalence, and mortality on an exhaustive list of diseases and injuries; the latest report in 2021 presents data for 369 diseases from 204 countries, spanning a twenty-year period (1990-2019)[5]. In estimating the impact of these diseases, the GBD study assesses two main metrics; years of life lost (YLLs), which combines the standard life expectancy for a given individual with the years of life lost upon death, and years lived with disability (YLDs) which is calculated by estimating the number of years an individual has lived with a certain disability weighted by the severity of the disability. Following adjustment for demographic and clinical characteristics, these are combined into a single metric termed disability-adjusted life years (DALYs), widely used to measure the overall impact of disease burden.

In 2021, neurological disorders became the leading contributors of DALYs worldwide. There were an estimated 3.4 billion individuals worldwide suffering from neurological conditions, an estimated 43.1% of the world population. These were collectively responsible for 11.1 million deaths and 443 million DALYs in 2021[6]. Stroke ranked first in contributions to global DALY's, closely followed by neonatal encephalopathy, migraine and dementia. This global overview however obscures region-specific epidemiological patterns which show that neurological conditions disproportionally affect people in low- to middle-income countries, in

part due to the high prevalence of developmental and neonatal disorders in these regions. For instance, disorders like neonatal encephalopathy, meningitis, neural tube defects and encephalitis consistently rank higher in low-income regions such as South Asia, Caribbean and Sub-saharan Africa. In contrast, traits like Parkinson's disease (PD) and dementia score lower in comparison to high-income regions, where there is a higher burden of neurological conditions on the aging population. As a result, there is also a disproportionate rate of mortality from neurological disorders between low- and high-income regions, likely related to limited access to high-quality healthcare services with neurological specialists and where communicable neurological disorders are much more common[7].

Across age and sex, DALY's differ drastically; generally speaking, age-standardised DALY rates are significantly higher in men than women globally (1.12 male-to-female ratio), largely contributed to by PD, traumatic brain injury (TBI) and stroke, each with male-to-female ratios of at least 1.5. On the flip side, women were more susceptible to conditions such as migraine and multiple sclerosis (MS) (both with a male-to-female ratio of ~0.7). By age, infectious neurological conditions (such as meningities, tetanus and encephalitis) mainly target children between 0-5 years old, but disproportionately so in low-income regions. DALY's from epilepsy, for instance, are highest between 5-29 years old, followed by migraine and tension-type headaches between 30-40. Disorders such as MS and motor neuron diseases become more common between the age of 40-50. At this point, stroke also becomes increasingly more common, remaining the largest contributor of global DALY's between 60 and 84 years, after which Alzheimer's disease (AD) and other dementias become the most prevalent neurological burden[6]. This burden is likely to increase in the next 25 years, with WHO estimates projecting a population of 2.1 billion over 60 by 2050, and 426 million over 80[8].

Psychiatric disorders also pose a significant health burden; a systematic review of 12 mental disorders from the GBD 2019 study identified a total of 125.3 million DALY's where depressive and anxiety disorders were observed as leading causes of burden worldwide, ranked 13th and 24th in global DALY's respectively[9]. While schizophrenia was not observed to be as prevalent, the disability weight (used in the YLD calculations) for acute psychotic episodes was the highest estimated across the entire GBD study. It's important to note that the DALY estimates for mental disorders do not include YLL's (with exception of eating disorders) and as such do not represent fatal burdens. Despite significantly impairing health in their own

regard, these disorders also increase risk of other health outcomes such as suicide[10] and commonly co-occur with cardiovascular diseases, cancer and diabetes[11]. In contrast to neurological disorders, mental disorders remained generally proportional throughout life (with the exception of conduct disorder, highly prevalent in the 10-19 years category), peaking around the 35 year mark before a steady decrease.

While epidemiological approaches have identified risk factors contributing to the prevalence of disorders across the neurological and psychiatric spectrum, we have yet to understand the root cause for the large majority of disorders. Ultimately, they present themselves as a result of complex interplay between genetics, underlying biology and environment. In fact, in addition to sharing considerable epidemiological overlaps and comorbidities, many of these disorders across both classes can share clinical manifestations, but seemingly so from different and independent biological pathways (such as psychosis which can present itself in both schizophrenia and AD). This could be attributed to largely distinct genetic architectures between traits; not only is there a significant difference between neurological and psychiatric disorders, but even within neurological disorders (where psychiatric traits correlate within themselves more often)[12]. Understanding the genetic forces that drive these complex traits, through molecular mechanisms and biological pathways, has therefore become one of the primary goals in disease understanding.

1.2 Current therapeutic landscape for neurological and psychiatric disorders

On average, the cost of bringing a therapeutic to market is $2.6 billion dollars (2013) and only 4% of drug-development programs result in licensed marketable drugs. This statistic is in large part due to two main issues; first, experimental models at the preclinical stage are inadequate in predicting therapeutic efficacy and second, evidence of target validity cannot be confirmed until randomised controlled trials in humans which occurs at the later stages of the drug development journey[13,14].

In many complex neurological disorders, there is no cure. In AD for example, the traditional forms of drugs for treatment of cognitive symptoms either work to inhibit acetylcholine

breakdown (an important neurotransmitter implicated in a wide range of cognitive and memory functions) such as in the case of Donepezil, or to maintain homeostatsis in the glutamatergic system such as with Memantine. These are also used broadly in the treatment of other dementias, such as lewy body disease dementia or PD dementia[15]. Since 2003, only three drugs have been approved for AD treatment; Aducanumab is an antibody-based therapeutic designed to target amyloid beta protofibrils[16] which received accelerated approval from the Food and Drug Administration (FDA) in 2021 but will be fully discontinued as of November 2024 due to safety concerns. Lecanemab, also an antibody-based therapy targeting amyloid beta aggregates (usually as it begins to form fibres), was approved in 2023 and has shown some success in early AD[17] as well as the recently (July 2024) FDA-approved Donanemab[18], which in contrast to Lecanemab binds to amyloid at the plaque formation stage. Similarly, there are three main types of medication for PD, all targeted around maintaining dopamine levels to address motor symptoms (such as Levodopa and dopamine agonists) as well as common psychiatric treatments for non-motor symptoms, but can vary greatly by disease subtype[19]. What is clear from the currently available treatment options is that these only help with disease management and unfortunately do not prevent or slow down disease progression.

One of the key differentiatiors between the classic definition of neurological and neuropsychiatric disorders is the presence of directly observable molecular entities, such as the presence of protein tangles or unusual electrical activity patterns. As such, treatment of psychiatric disorders is generally less targeted. These usually fall into 3 categories; anti-depressants, used for example in the treatment of clinical depression, obsessive compulsive disorder (OCD), anxiety disorder and post-traumatic stress disorder (PTSD) and generally targeting serotonin and noradrenaline pathways[20,21]; mood stabilisers as used to treat symptoms of bipolar disorder, mania as well as depression, usually lithium-based or anti-convulsants[22]; and anti-psychotics, which are generally dopamine blockers used commonly in the treatment of schizophrenia but any disorder that includes psychotic episodes[23]. Many of these currently prescribed drugs however are similar to first-generation drugs developed more than 50 years ago, and do little to address the root cause of these disorders[24].

The lack of targeted treatment options is not for lack of effort; as of 2023, there were 139 active clinical trials for the treatment of PD and 187 in AD, many of which targeting specific genes or utilizing novel cell therapies[25,26]. In MS the drug development pipeline has also been

somewhat successful, but due to a shifting understanding of the disease so too has the therapeutic strategy. For instance in progressive MS it is generally accepted that the disease is initially immune-related, which resulted in a large catalogue of immunotherapies, but progresses towards neurodegenerative mechanisms which are less understood[27]. In contrast, limited progress has been made in the development of novel medications for psychiatric disorders, partly due to lack of understanding in pathophysiological mechanisms underlying their presentation[28].

Exploring the drug development pipeline for these disorders highlights the importance of genetically supported therapeutic development, which have been shown to be more than twice as likely to be successful in clinical development as evidenced by the fact more than two thirds of the 2021 FDA-approved drugs are supported by human genetics[29,30]. Along with yielding more personalised therapeutic approaches, the impact of genetics on disease evolution also offers hope to identify targets that act before the onset of disease, with the ultimate goal to generate preventative treatment options.

1.3 Genome-wide Association Studies (GWAS)

In an otherwise complicated setting to study disease, genetic variation provides a fixed point to understand disease susceptibility which is generally environment agnostic since it is set at conception[31]. Genome-wide association studies (GWAS) have therefore provided a powerful avenue for identifying genetic variants that may contribute to the risk of developing a complex trait[2]. These studies aim to assess the statistical significance of an association between allelic frequency of a single nucleotide polymorphism (SNP) and predisposition to a phenotypic trait, typically by testing millions of SNPs across the entire genome. Over the past 18 years, GWAS have identified a plethora of robust associations across most of the human phenome. These studies not only uncover genetic predispositions; they contribute valuable insights into the underlying biology of the phenotype (and can consequently help inform therapeutic development), help estimate heritability, and in some cases enable causal inferences between risk factors and health outcomes[32].

The successful identification of genetic variants is highly dependent on adequate statistical power. Powering a GWAS requires consideration of multiple factors including sample size (or

trait prevalence), effect sizes, disease allele frequency as well as the presence and strength of linkage desiquilibrium (LD)[33]. Sample size is by far the most addressable factor and as such has prompted the creation of large cohorts, many through international consortia. Since the first GWAS in age-related macular degeneration on a cohort of 96 cases and 50 controls (2005), recent studies with the help of decreasing sequencing costs have ascended to millions of participants. As of 2023, there have been more than 45,000 published GWAS conducted on more than 5,000 traits across the human phenome[34,35] (Figure 1.1).

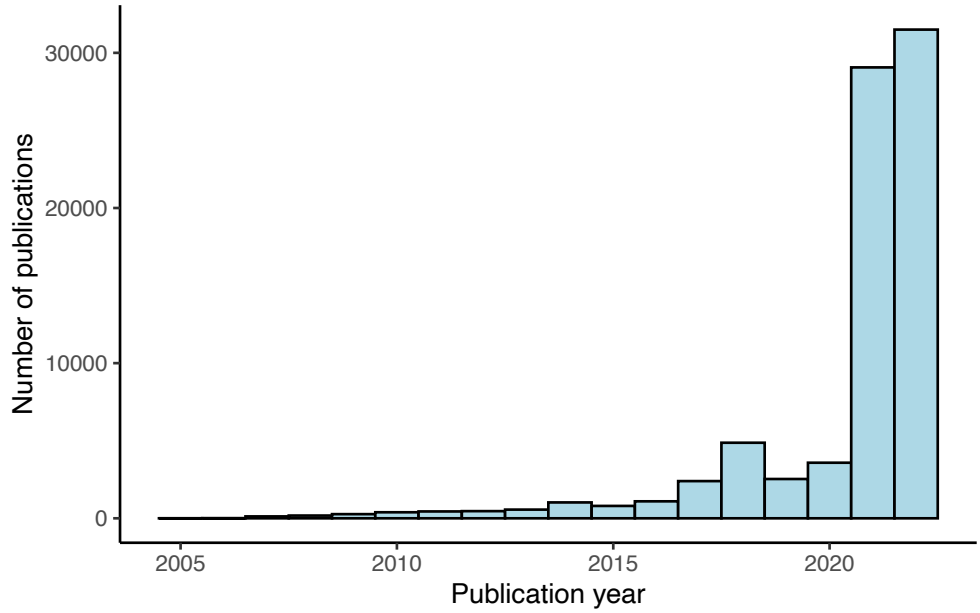


Figure 1.1: Number of GWAS publications as of 2023. The full list of GWAS was obtained from the European Bioinformatics Institute (EBI) GWAS catalog, filtered to only keep one first author per GWAS per year.

Heritability in particular is a key factor in supporting the logic to conduct these large-scale studies, as it suggests the degree to which a trait is the consequence of the underlying genetics as opposed to environmental factors. Twin studies are the gold standard for genetic heritability of disease, and have shown that many of the more prevalent brain disorders have high rates of heritability, such as in AD (~70%)[36], PD (~40%)[37] and schizophrenia (~79%)[38]. On the other hand, SNP-heritability estimates from genome-wide association studies range between 10-50%[12] and can generally reach up to 60% in cognitive and structural traits[1,39] which suggests a substantial genetic element captured by GWAS despite only capturing SNP-level associations. This is typically the result of a combined effect of common variation (i.e, the individual inheritance of multiple SNPs associated with the trait), as shown in AD, PD and schizophrenia[42].

GWAS in neurological disorders have identified hundreds of statistically significant genomic loci, which by convention (although will sometimes vary by study) have an association p-value $< 5 \times 10^{-8}$. A recent study of AD, utilising more than 110,000 clinically diagnosed AD cases and more than 670,000 controls discovered 75 significant risk loci, many of which involved amyloid/tau pathways as well as highlighting the involvement of microglia. Importantly, the increase in sample size found 42 of the 75 loci to be newly discovered in comparison to previously reported AD GWAS[40], many of which uncovered new biological mechanisms pertaining to AD pathology. In PD, Nalls et al.,[42] conducted a meta-analysis of multiple PD GWAS which included more than 37,000 cases and 1.4 million controls, identifying 78 significant genomic regions such as the well-known *LRRK2* and *SNCA* loci, as well as novel loci with *GRN* (interestingly known to be associated with frontotemporal dementia). On average, these loci explained between 16-36% of the heritability and nearly half were newly discovered.

This success is also seen in neuropsychiatric disorders; for instance, the latest schizophrenia GWAS utilized a cross-ancestry cohort including ~78,000 cases and ~243,000 controls to identify more than $>$ 260 loci, which through post-GWAS analyses revealed the implication of 106 protein-coding genes (such as *GRIN2A*, a glutamate receptor subunit, and the transcription factor *SP4*)[41]. In addition, there have been numerous GWAS spanning most if not all of the psychiatric disorder spectrum, including major depressive disorder (MDD), bipolar disorder, attention deficit hyperactivity disorder (ADHD) and eating disorders, to name a few[46].

There has also been a significant interest in recent years on the genetic role in quantitative phenotypes outside of psychiatric and neurological disorders. Cognitive abilities such as general intelligence, memory, verbal-numerical reasoning, reaction time have demonstrated high heritability and are linked to important life outcomes. For example, in 2018, Savage et al., conducted the largest GWAS on intelligence to date using a cohort of ~270,000 individuals, using a carefully combined score from multiple cognitive functioning evaluation. The authors discovered 205 associated genomic loci (190 of which were novel) which implicated more than 1,000 genes[47]. Importantly, these findings showed considerable genetic correlation with multiple other disorders, and suggested a protective effect for AD and ADHD while uncovering a potential bi-directional effect for schizophrenia.

Structural phenotypes have also been particularly relevant. These studies typically use magnetic resonance imaging (MRI) to perform measures of brain anatomy which include tissue and structure volume (such as grey matter volume or cortical volume), as well as more specific biomarkers (for example, degree of vasculature and white matter microstructure). Importantly, brain development is a dynamic process over the life-span and as such has a range of implications in cognitive, neurological and psychiatric traits[48]. One of the most comprehensive GWAS was performed by Elliott and colleagues in 2018, investigating 3,144 functional and structural brain phenotypes implicating thousands of genomic loci[49]. Half of the observed phenotypes demonstrated significant heritability, with estimates in some of these explaining up to 80% heritability.

While GWAS have been incredibly successful in identifying risk loci for neurological traits, it is important to remember that on their own they only offer genetic association estimates between SNPs and trait; rarely do they pinpoint exact causal mechanisms or specific genes, often highlighting a multitude of loci. As such, post-GWAS analysis is almost always necessary to understand the function of these variants and the pathway through which they act, as well as to prioritise risk loci[50].

1.4 Fine-mapping methods

At the core of molecular biology as outlined by Crick in 1970 is the principle by which information flows in a biological system; from the stored genetic recipe in the form of DNA, transcribed into RNA and finally translated into functional proteins that work together to dictate cellular functions and ultimately phenotypes[51]. Understanding how common genetic variation modulates this journey is a crucial aspect of post-GWAS analysis and the first step towards understanding the molecular mechanisms driving a phenotype. Given that the majority of genetic variants identified from GWAS are non-coding ($> 90\%$) and therefore not directly affecting final protein products, it is likely they are involved in one of many regulatory mechanisms governing gene expression[52].

This begs the question on whether a genetic variant (or variants) may itself be causal for a phenotypic presentation through its regulatory funcions. A large part of the success of GWAS in the last decades can be attributed to the use of cost-efficient genotype microarrays, which

rely on "tag SNPs" that are highly correlated through LD to neighbouring SNPs[53]. These tag SNPs serve as substitutes for large genomic regions containing a majority of unmeasured SNPs, which must therefore routinely be imputed[54], in turn making it difficult to identify the "true" causal variant. To address this question, many methods have been developed to pinpoint causal variants and their mechanism of action (collectively grouped under the term "Fine-mapping") through the application statistical, bioinformatic or functional approaches[55].

1.4.1 Functional and bioinformatic fine-mapping

Having identified independent genome-wide significant regions, a relatively straightforward approach to attach functional significance to identified variants is to overlap these with known regulatory elements such as promoters and enhancers. Published data from studies interrogating DNA accessibility using methods such as assay for transposase-accessible chromatin sequencing (ATAC-seq)[56], transcription factor binding sites with Chromatin immunoprecipitation followed by sequencing (ChIP-seq)[57], histone methylation and acetylation interrogation to identify promoter and enhancer sites[58] can all help understand the potential function of variants.

More intricate methods also measure the transcriptional activity of promoter and enhancer regions using approaches under the umbrella term "Nascent RNA sequencing analysis" (NSRA), utilizing methods such as precision nuclear run-on sequencing (PRO-seq)[59] and global run-on sequencing (GRO-seq)[60]. These two methods differ slightly in that GRO-seq labels nascent RNA transcripts and therefore transcriptional dynamics involving regulatory regions, whereas PRO-seq labels the RNA polymerases directly to identify interactions and positioning relative to regulatory regions.

Other methods such as Cap Analysis of Gene Expression (CAGE) combine RNA expression quantification with identification of transcription start sites (TSS) which may also provide mechanistic insight when overlapping with genetic variants[61]. Finally, methods such as Hi-C aim to capture regions of chromatin contact and can explain distal variant effects which act through structural configurations[62], with recent methods enabling chromatin conformation capture at base-pair resolution[63–65].

Importantly, there have been large efforts to create and maintain global databases of these epigenomic data which are publicly accessible, such as ENCODE[66] and FANTOM[67] as well as bioinformatic tools such as *ANNOVAR*[68] and the Ensembl Variant Effect Predictor (*VEP*) which are simple to use and enable researchers to quickly query a set of SNPs for potential functions[69].

1.4.2 Statistical fine-mapping

Given a region of genome-wide significance, it is possible to perform some simple heuristic approaches in variant prioritisation. For instance, filtering variants according to their correlation with the lead SNP (which has the lowest association p-value) can quickly reduce the number of variants to investigate. Alternatively, clustering SNPs based on pairwise r^2 may reveal interesting LD structural patterns but depend on a user-defined r^2 threshold. However, these approaches are in part subjective (due to choice of thresholds) and do not consider joint effects of SNPs, nor do they provide a measure of confidence relating to the probability of a SNP being causal. There have therefore been numerous statistical approaches developed to rigorously assess these relationships.

Penalized regression models were traditionally used to this end, which simultaneously estimate variant effect sizes and often shrinking of small effect sizes towards zero, where traditional regression models would suffer due to the high number of SNPs and the high pairwise correlations. These include methods applied broadly in a variety of fields such as lasso and elastic net[70], and generally use cross-validation to minimize prediction error. By contrast, Bayesian approaches have become the most popular approaches and instead focus on the probability of a specific hypothesis which allows for a more interpretable result, and have shown to outperform penalized regression models[71]. In essence, Bayesian probability theory aims to calculate a probability distribution of parameters of interest (the posterior probability), conditional on observed data and in combination with prior probability, which is assigned to the parameters of interest to represent some form of prior knowledge. In a fine-mapping context, the prior probability is assigned based on the assumption that there is a causal SNP in the region, with the aim to assign a posterior probability to each SNP, which in turn represents the confidence with which a SNP may be causal. Several methods have been developed and employ a variety of different approaches in calculating these posterior

probabilities; some of the earlier methods assume there is only one single causal variant (such as *BIM-BAM*)[72] while others allow user input. The underlying methodology in some tools employ Markov chain Monte Carlo (MCMC), while others employ an exhaustive search method (whereby every possible combination of variants is interrogated)[55]. Generally speaking, modern methods such as *CAVIARBF*[71] and *FINEMAP*[73] output a posterior inclusion probability (PIP), which denotes at SNP-level the probability of it being causal (essentially by summing the probabilities of it being causal in all models tested). Importantly, Bayesian fine-mapping allows direct comparison of PIPs between SNPs unlike GWAS *p*-values, and because the models are based on the joint SNP effects allow better control of effect sizes in the identification of causal SNPs.

1.5 Quantitative Trait Loci; an overview

Fine-mapping approaches are useful in attempting to determine which SNP is causal, and through genomic overlaps may give some insight into their potential function. Ultimately, determining the causal path from genetic variant to trait requires quantitative knowledge of molecular phenotypes. Quantitative trait loci (QTL) is an overarching term for genomic loci with a statistical association to a quantitative trait (generally molecular); these not only include QTLs for quantified gene expression (eQTLs) and protein levels (pQTLs), but also other modalities such as methylation and splicing patterns (meQTLs and sQTLs respectively) and more recently chromatin accessibility (caQTLs) as a successor to DNaseI sensitive QTLs (dsQTLs)[74–79] (Figure 1.2). Similar to a GWAS, these associations measure the level of change in the observed phenotype based on the presence of a mutant allele in contrast to the wild-type configuration, often referred to as alternative and reference alleles.

These associations are divided into *cis*- (in close proximity to the gene/feature/trait, typically in a window of 100kb – 2Mb depending on study design) and *trans*- associations (anything outside the defined window > 2Mb and even in other chromosomes). The biological implication is that *cis*-regulatory associations are presumed to regulate the moiety directly, while *trans*- effects are speculated to act through intermediary entities in a regulatory network. In practice, *trans*-effects are usually weaker and therefore more difficult to identify compared to *cis*- effects partly due to a much larger multiple testing burden, requiring

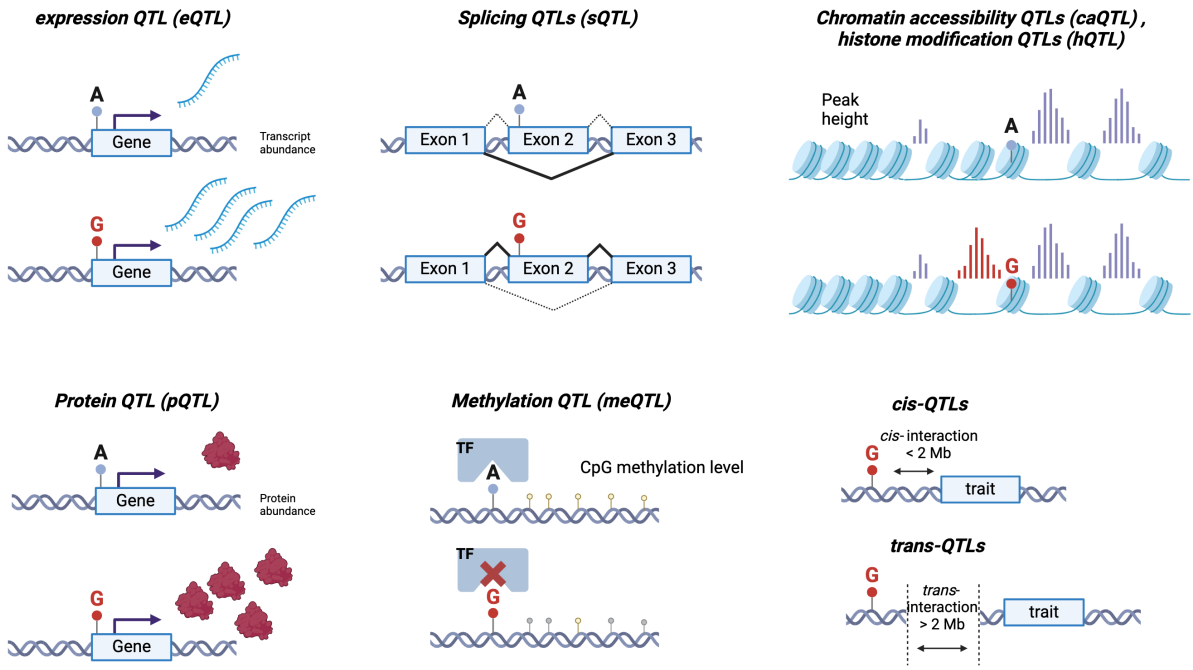


Figure 1.2: Overview of molecular Quantitative Trait Loci (molQTLs). Molecular Quantitative Trait Loci (molQTLs) can in principle be identified with any quantitative molecular phenotype if genotype information is available on the same sample. These include (but are not limited to) through measured gene expression (eQTL), protein abundance (pQTL), abundance of transcript variants as generated through splicing (sQTL), methylation patterns (mQTL) and peak data such as chromatin accessibility (caQTL) or histone modification measurements as measured through ChIP-seq (hisQTL). These can be separated into *cis*- and *trans*- associations depending on the distance of the SNP to the trait (generally <2Mb for *cis*- associations, and >2Mb for *trans*-associations).

enormous sample sizes. for instance, one of the largest eQTL studies to date on whole blood from 31,684 individuals discovered 16,987 genes with associated *cis*-regulation but only 6,298 genes with *trans*-regulation[4].

Importantly, the uncovered regulatory relationships are most often tissue as well as cell-type specific[80]. The Genotype-Tissue Expression (GTEx) consortium published in 2020 an extensive map of eQTLs (both *cis*- and *trans*-) across 49 human tissues from 838 human donors, showing a large proportion of these associations were tissue specific. Generally, *cis*-eQTLs were more likely to be shared across tissues than *trans*-eQTLs[81]. As expected, this observation is not unique to gene expression; it has been shown in a comparative study of pQTLs in cerebrospinal fluid (CSF), brain and plasma[82], in caQTLs[83], sQTLs[77] and more recently for mQTLs[84].

1.5.1 eQTLs; recent developments

The vast majority of QTL studies have been conducted using quantified gene expression; it is easily measurable and has well-established experimental methodology. While the concept of genetically driven variation in gene expression is at least 60 years old[85], the first genome-wide eQTL study was conducted in 2002, when Brem and colleagues through the use of microarrays were able to quantify differences in gene expression between two strains of budding yeast[86]. This was shortly followed by Schadt and colleagues the following year, expanding the field to mammals and widely recognised as the first paper to coin the term[87].

In the last twenty years, the eQTL field has grown considerably and in many directions, prompting analyses in multiple organisms and the creation of international consortia to perform large-scale studies in humans, such as the GTEx, PsychEncode and eQTLGen consortia[4,66,74]. Transcriptomic profiling by RNA-sequencing (succeeding traditional microarray techniques and first introduced 15 years ago) thanks to the development of next-generation sequencing (NGS) allows quantification of tens of thousands unique transcripts genome-wide through the use of cDNA synthesis followed by deep sequencing in a high-throughput fashion[89]. More recently and specifically in the brain, the Metabrain initiative generated joint expression and genotype data to perform large-scale eQTL discovery on more than 6,500 individuals from multiple ancestries[90].

The latest development in the eQTL field comes largely as a by-product of the single-cell revolution[91]. Single-cell RNA-sequencing (scRNA-seq) is not a new concept; the history of single-cell transcriptomics can be traced back to experiments using quantitative PCR (qPCR) in the early 90s[92], where Eberwine and colleagues in 1992 isolated live neurons in which they compared mRNA abundance of neuronal genes. It was not until 17 years later that the landmark study by Tang and colleagues in 2009 became one of the first to comprehensively sequence the whole transcriptome of a single cell, investigating gene expression by RNA-seq in isolated germ cells during embryonic mouse development[93]. These early experiments, albeit 17 years apart, show that the primary motivations for single-cell approaches were to understand the behaviour of specific cells or cell types.

Thanks to the rapid development of novel microfluidic devices and the evolution of next-generation sequencing, following studies were motivated to perform more broad experiments targeting larger amounts of cells in order to understand the systems as a whole. Guo et al. in 2010 was one of the first to initiate the pivot to larger scale studies, with their analysis of more than 500 embryonic mouse single cells[94]. Following this approach, the Linnarsson group from the Karolinska Institute started developing methods for multiplexed RNA-seq to allow parallel characterization of multiple cells at the same time, eventually leading to their landmark study characterizing the mouse brain by sequencing of more than half a million cells in 2018[95,96].

Since then, the field has seen an exponential increase in publications, most with an equally exponential increase of number of cells sequenced. It has become more commonplace for the generation of large single-cell "atlases", comprising several millions of sequenced cells. Perhaps the best example of the scale to which the single-cell revolution has grown is the Human Cell Atlas consortium, a global collaboration spearheaded by Aviv Regev and Sarah Teichmann and established in 2017 with the aim to map cell-types in all human tissues[97]. The consortium, which involves more than 1,700 institutes in 99 countries, has compiled single-cell data for 58.4 million cells in the majority of human tissues. As part of this collection, the Linnarsson group and collaborators published the human brain atlas which comprises more than 3.3 million cells in 100 specific brain locations, split into 31 cell-type superclusters and more than 3,300 subclusters[98].

1.5.2 Single-cell eQTLs

Because we know eQTLs to be tissue-specific the question becomes whether genetic regulation of genes is also cell-type specific, and whether bulk approaches suffer from obscuring cell-type specific signals. As of January 2024, there have been 16 published studies utilizing single-cell approaches for eQTL mapping[99]. Wills et al., in 2013 marked the first exploration of eQTLs at cell-type resolution, studying the genetic variation of 15 individuals and their effect on the expression of 92 genes from 1,440 single cells[100]. The first truly single-cell eQTL project came in 2018 following the commercialization of single-cell sequencing by Van der Wijst et al., performing a genome-wide *cis*-eQTL discovery in 25,000 peripheral blood mononuclear cells (PBMCs) from 45 donors[101]. The single-cell data was first analyzed to classify cells into the major immune cell-types, after which the expression matrices were "pseudobulked", a process by which the gene expression values are summed or averaged per cell-type and per donor to mimic a bulk dataset. This method also has the advantage of bypassing hurdles inherent to single-cell data, in which data sparsity is likely the most common as well as the design of appropriate statistical models[102]. In addition, this allows the use of traditional bulk expression tools for eQTL mapping (such as MatrixEQTL)[103] (Figure 1.3).

While cost of sequencing has been reduced, the extension to single-cell resolution has added a new dimension, especially in the field of QTL mapping. Where bulk eQTL studies were routinely using hundreds of individuals, if not thousands for larger projects, generating data for the same scale with scRNA-seq was (and still is) difficult due to cost (Table 1.1).

Table 1.1: Examples of sequencing costs (RNA- vs scRNA-seq)

Library type	Method	Cost per sample
(Bulk) PolyA RNA	Universal Plus mRNA	$202
(Single-cell) 3'/5' Gene expression	10x Genomics	$3,095
Example of sequencing costs at optimal sequencing depths at the UCSF core genomics facility (https://functionalgenomicscore.ucsf.edu/genomic-technologies		

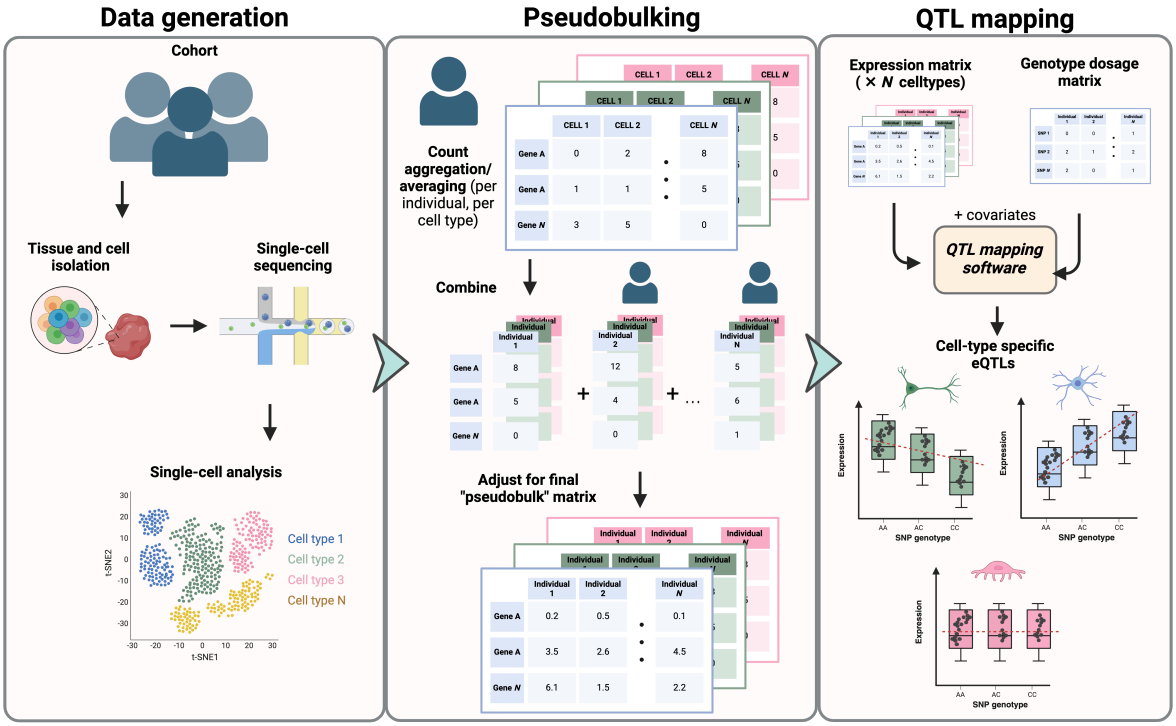


Figure 1.3: Pseudobulking for single-cell eQTLs. Following data generation in a cohort of suitable statistical power, single-cell analysis is undertaken to identify cell-type populations in a tissue. Pseudobulking involves aggregation of counts across all cells in a given cell-type population for a single individual, producing a single vector of gene counts. Individual vectors from each individual are then combined into a single expression matrix of size E × N, where E is the number of genes and N is the number of individuals for each cell-type. These combined matrices can then be corrected using traditional bulk RNA-seq approaches, such as normalization and scaling, and with the addition covariates and a genotype dosage matrix (of size G × N, where G is the number of SNPs) containing allele dosages which can be used in QTL mapping software for cell-type specific eQTL discovery.

Given that publicly available scRNA-seq datasets became increasingly more common, there were many efforts to develop computational deconvolution methods in an attempt to integrate bulk RNA-seq and scRNA-seq to identify cell-type specific *cis*-eQTLs such as with *CIBERSORT* (and later *CIBERSORTx*)[104,105] and *TIMER*[106]. However, these methods were generally limited to the analysis of known cell-type as identified from reference datasets, and disallows the *de novo* discovery of cell-type populations and consequently rare cell-type eQTLs (although less of a problem with now publicly available large-scale atlases with hundreds of reference cell-types).

The largest eQTL study using scRNA-seq to date was published by Yazar and colleagues in 2022, analyzing 1.27 million cells in whole blood from 982 individuals (the OneK1K cohort)[107]. They discovered a total of 26.6k *cis*-eQTLs and 990 *trans*-eQTLs across 14 major immune cell-types, the majority of which were only present in specific cell-type populations. In the brain, the largest study to date was published by Bryois et al., performing a genome-wide eQTL study in post-mortem brain tissue from 192 individuals and discovering 7,607 genes with significant genetic regulation by one or more SNPs (eGenes)[108] in 8 major brain cell-types. Importantly, both studies showed that using scRNA-seq for eQTL discovery outperformed deconvolution methods and by proxy showed a significant gain in statistical power when compared to bulk.

1.5.3 Advantages and limitations of single-cell eQTL approaches

Some of the main advantages of using single-cell approaches have already been listed, such as the gain in statistical power and the ability to detect genetic regulation for rare cell-types which are diluted in bulk expression data. Another key advantage is the ability to identify eQTLs regulating genes with low expression levels; thanks to expression values across individual cells, it is possible to measure the variability around the expression of a gene and perform more accurate gene expression estimation[101].

Perhaps one of the strongest advantages single-cell eQTL approaches offer aside from cell-type specific eQTLs is the possibility to map out dynamic eQTLs that vary across cell *state*. Cells are inherently dynamic in nature and routinely transition between populations, where it is not unreasonable to assume changes in regulatory landscape also follow. This has been

demonstrated in multiple studies including the aforementioned work by Yazar and colleagues[107] but also specifically in a study of unstimulated T-cells in 259 Peruvian individuals[109], as well as in vitro activated CD4+ T-cells in 119 individuals across multiple time-points[110].

Naturally, single-cell eQTL approaches inherit many of the pitfalls present in single-cell RNA sequencing, such as experimental noise, high drop-out rates and high costs[102]. In many ways, pseudobulking offers a straightforward approach to circumvent much of the noise, aided by the fact traditional statistical frameworks can then be applied such as differential expression and as discussed extensively, eQTL mapping[111]. It is important to note however that pseudobulking is not devoid from bias as it relies on accurate labelling of cell-type populations or clusters that are then used in pseudobulking. In a way, although to a much lesser extent, these pseudobulked clusters suffer from a similar problem that bulk expression does in that expression signals from rarer sub-populations of cell-types may be obscured. In addition, it is currently unclear whether the total number of cells within a cell-type population used in pseudobulking play an important role; Bryois et al., used as little as 10 cells as a minimum per individual to generate combined gene expression values, whereas Yazar et al., do not mention any threshold[107].

1.5.4 QTL study design

The procedure for performing a QTL study generally starts with identifying a study cohort in which both the quantified molecular phenotype and genotype will jointly be generated for each sample or individual. In most studies, material is collected from tissue samples (such as whole-blood or post-mortem brain) although also possible for cell-lines, which opens up opportunities for carefully controlled environments and introduction of specific stimuli such as in the case of induced pluripotent stem (iPS) cells[112].

Choosing the target tissue should align with the research question and may raise practical concerns; for instance, RNA sequencing (RNA-seq) data can be reliably generated on a variety of tissues with relatively low amounts starting material, whereas more elaborate methods such as assay for transposase-accessible chromatin sequencing (ATAC-seq) generally requires more starting material and high sequencing depth which can be costly. In addition,

the use of archived tissue where RNA degradation is a concern (Formalin-fixed paraffin-embedded tissue, for example) requires careful planning and is still an area of active method development in the case of RNA-seq[56,113].

Following this, a key practical concern is the availability of sample. Cohorts used in QTL studies are generally quite large, with pilots ranging from 50-100 samples while larger studies can comprise thousands of individuals. Similar to GWAS, QTL discovery is highly dependent on the distribution of allele frequencies and as such need to be adequately sized to achieve acceptable statistical power[114]. Studies performed on whole-blood are often large, as it is routinely collected and minimally invasive in contrast to brain, which are most often carried out with post-mortem tissue that is more difficult to obtain[4,90].

Genotype data is typically generated genome-wide to maximize the potential discovery of QTLs. While it is possible to perform genotype calling directly on the reads obtained from the quantitative phenotype (such as RNA-seq), the variability in the reads make SNP detection and genotype calling challenging. Instead, the most common approach is to use methods such as SNP genotype arrays and whole-genome sequencing. In the case of SNP genotype arrays, coverage is generally low and usually requires imputation using a reference population; however, it is relatively inexpensive and is often the default method for genotyping in GWAS as well as in QTL studies[115–117]. Whole-genome sequencing (WGS) is comparably more expensive but offers the possibility to investigate genetic variation at a higher resolution and consequently identify QTL effects from more complex variants such as tandem repeats or structural variants[118].

1.5.5 eQTL methodology and considerations

In many ways, QTL studies can be directly compared to GWAS study design. At the crux of eQTL analysis is how the association between genotype and gene expression is made, and how the effect is measured. There have been several statistical methods for this purpose, including linear regression and ANOVA models, as well as generalized mixed models and Bayesian approaches[72]. However, non-linear methods can be computationally intensive and relatively slow[119] and as such linear models have largely been the most popular tool for eQTL discovery or “mapping”.

It is common for the input data to undergo significant quality control prior to eQTL mapping. In the case of RNA-seq, sequencing read depth can vary significantly between samples which is directly reflected in the resulting transcript counts. This is typically adjusted for using normalisation methods such as cpm (counts per million) which scale the raw read counts of each gene in a sample-specific approach, by using a scaling factor of one million. The normalised counts are commonly transformed to a log-scale, which not only is important for the expression to conform to regression model assumptions of homoscedasticity (meaning almost constant variance) but also helps with interpretability, especially when comparing across genes[120].

Despite this correction there are multiple other factors to consider to avoid confounding effects. The presence of unobserved factors that influence the SNP-gene association may introduce false positives, as well as obscure potential relationships. Technical factors, such as sequencing platform or batch effects may adversely impact the measured gene expression and must be considered when performing the association[121]. Biological and clinical factors also play an important part; whether the tissue is from disease or healthy samples, the age and sex of the donors, and in the case of brain the post-mortem interval all have a significant effect on the resulting gene expression[81,122,123].

In most linear models, a single SNP-transcript pair is tested, where the genotype dosage (for example AA, AB or BB) is the predictor variable and the transcript abundance is the outcome variable in an additive model[124]. In its most basic form, the linear model looks as such (Equation 1.1);

$$y = a + bx + \epsilon \tag{1.1}$$

Where y indicates the resulting gene expression in a gene the dependent variable, x is the additive genetic effect (0,1,2 of the study allele) and ϵ is the residual error and the goal is to test the significance of b, which represents the effect size of the genotype on expression levels (also known as slope). However, it is important to note that most models also incorporate other factors such as clinical covariates or principal components (PCs) of expression to account for structural artefacts in the observed expression[103,125].

This can quickly become computationally intensive; in a study consisting of measured gene expression for 20,000 genes and genotype data on 5 million SNPs, there are approximately $\sim 1 \times 10^8$ *cis-* associations (assuming around 5,000 SNPs per *cis-* window) and $\sim 1 \times 10^{11}$ *trans-*associations to be calculated. Despite the individual models being relatively quick to compute, performing these sequentially would severely hinder the study. To circumvent this issue several packages have been developed that are highly optimized for the scale of these computations; the MatrixEQTL package for instance uses matrix multiplication to compute associations in blocks[103], and recently the TensorQTL package implements linear algebra operations in a deep learning setting, which increases computation speeds more than 200-fold but requires the use of graphics processing units[126].

1.6 Genetic colocalisation

The key question that drives QTL studies is whether variants in non-coding regions have a regulatory function on the phenotype in question; that is, whether the presence of one or more alternate alleles significantly alter the a measured molecular feature in a certain region (be it through open chromatin, splicing, methylation or gene expression). Having identified a robust set of QTLs, the natural next question that a researcher may ask is whether these variants also predispose individuals to a certain measurable trait as identified in GWAS. This approach assumes that GWAS variants must act through an intermediary moiety to confer risk, such as gene expression or epigenetic modifications. For instance, an early study by Moffatt et al., (2007) showed that SNPs with a significant association to childhood asthma also affected *ORMLD3* expression in lymphoblastoid cell lines[127]. A similar correlation was observed by Dubois et al., (2010) in celiac disease, identifying several genes involved in immune functions that were also associated to celiac risk variants[128].

However, identifying a causal variant from these results is difficult for many reasons. First, the majority of genotype data has been generated to this day through genotype arrays, especially in GWAS where costs of WGS would severely hinder the feasibility. Genotype arrays typically exploit the concept of LD to achieve genome-wide coverage, by only using a set of "tagging" SNPs that would allow for imputation of adjacent. These are commonly referred to as "typed" versus "untyped" SNPs[72]. The consequence of this is that the results of

genetic studies (QTL or GWAS) are inherently ambiguous and may not properly highlight causal SNPs. Second, it is also possible for genomic loci to contain more than one causal SNP that collaboratively contribute to the association[129].

Identifying shared association signals across genetic studies (such as eQTL x GWAS or GWAS x GWAS) is therefore a demanding statistical problem. A few methods were initially proposed; regulatory trait concordance (RTC) by Nica et al., suggests ranking SNPs with associations to gene expression and a GWAS trait by measuring the residual association of SNPs to the gene conditioned on their association to the GWAS trait[130]. This method however does not make any assumptions as to the number of causal variants, and underestimates the effect of "Winner's curse" (due to random fluctuations the causal variant may not always have the strongest association)[131].

Alternatively, He et al., (2013) developed a method called Sherlock which does not explicitly test statistical relationships shared between SNP, gene expression and GWAS. This method instead focuses on the convergence of signals rather than the identification of common causal SNPs, applying a "finger-printing" approach based on a Bayesian framework that utilizes binary indicators from association p-values to determine whether two signals converge[132]. In addition, the method disproportionately prioritizes the association of SNPs to gene expression in determining the test statistic; it is possible that there are SNPs associated to both traits in the same region but in a divergent manner (i.e one set of SNPs associated to gene expression but not the GWAS, and vice-versa) which would discourage a shared mechanism for both traits.

The term colocalisation in the context of genetic association studies was therefore coined to ask whether two independent association signals are supported by a shared causal variant. In 2012, Wallace and colleagues performed one of the first colocalization studies, performing a GWAS on type 1 diabetes in combination with an eQTL study from monocyte expression. They discovered 21 colocalized genes in 14 significant loci of type 1 diabetes, many of which supported pancreatic autoimmune pathways[133]. However, this study had access to full individual level-data (meaning genotype and transcriptomic information for each participant) which is often inaccessible in published GWAS and therefore limiting applicability. This led to the development of COLOC, a bayesian framework for genetic colocalisation testing that works on summary-level data[134]. Since then, COLOC has been used extensively for multiple

types of studies, such as in immune diseases with immune cell-type eQTLs[135] or across a range of phenotypes using GTEx data[136] as well as in in PD using brain specific eQTLs further integrated with methylation data[137].

The COLOC method works as follows. Given a set of variants in a genomic locus, the method explicitly tests 5 hypotheses each delineating a specific causal configuration (Figure 1.4);

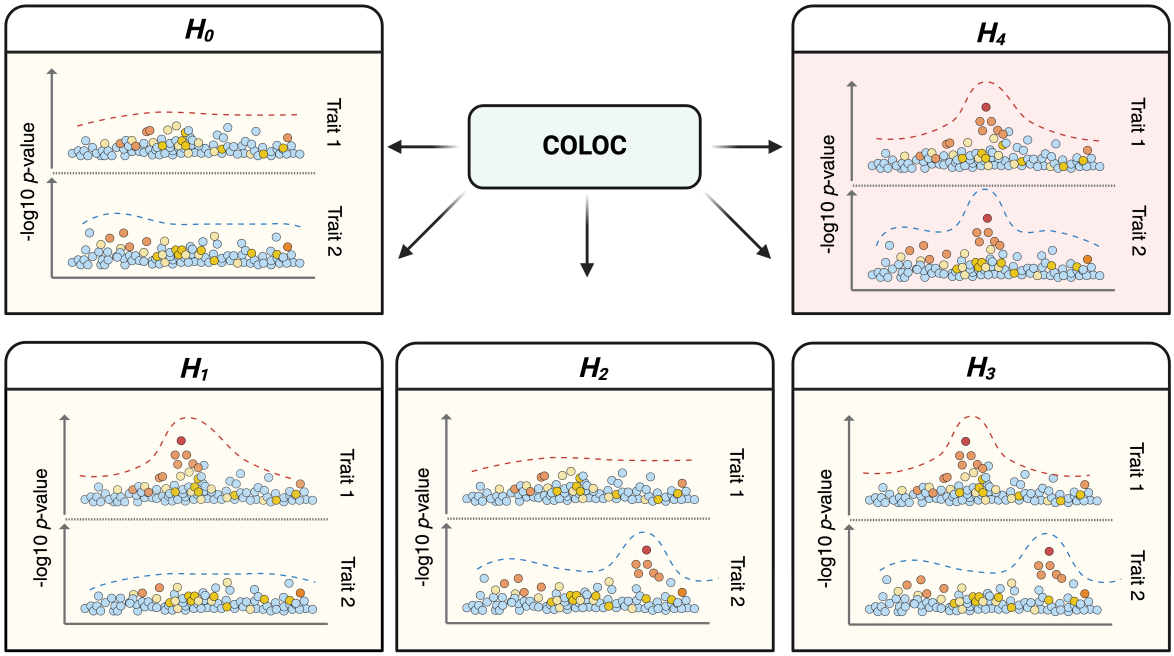


Figure 1.4: Genetic colocalisation testing. Genetic colocalisation, popularised by the tool *COLOC* for the assessment of two phenotypic traits, generally tests 5 hypotheses in a Bayesian probability framework (described below). The null hypothesis (H_0) represents no association with either trait, meaning a weak or no signal in both of the traits. COLOC, combining prior probablities with the association data from both traits, calculates and reports the probability for each hypothesis in a given region, where a high H_4 posterior probability indicates strong support for a shared causal variant.

- H_0: There is no association with either trait
- H_1: There is an association with trait 1, but not trait 2
- H_2: There is an association with trait 2, but not trait 1
- H_3: There is an association with both traits, however there are two independent signals
- H_4: There is an association with both traits, supported by a shared causal variant

The level of support for a colocalised signal is then represented by the posterior probability for the H_4 hypothesis (abbreviated as PP.H4), where a large posterior probability supports a shared causal variant between the two traits. Importantly, and distinctively from typical

fine-mapping approaches, the COLOC method is not formulated to ask which variant is likely to be causal, but whether there is a statistical probability of a shared causal variant (although this method also report a SNP-specific PP.H4, which may be interpreted if the overall PP.H4 is strong).

1.6.1 Recent developments in genetic colocalisation

Over the last few years, several methods have also built on the initial COLOC framework to expand the capabilities of genetic colocalisation. In 2018, the authors of COLOC developed MOLOC (multiple-trait-coloc), which expands on the existing Bayesian framework to include multiple traits by increasing the number of hypotheses tested. In their example using SCZ eQTL, mQTL data and the schizophrenia GWAS, their method tested a total of 15 hypotheses each denoting a specific causal configuration (and consequently resulting 15 posterior probabilities). This approach discovered 52 genes which were colocalized with schizophrenia risk, both through gene expression and changes in methylation pattern, marking an important step in prioritizing potentially causal genes as well as suggesting a regulatory mechanism[138]. While increasing the number of potential causal genes that can be discovered, the approach is computationally prohibitive due to the exponential increase of hypotheses to test for each added trait.

To address this issue, a recent method termed HyPrColoc was developed, where a major development was the possibility to approximate the probability of colocalization to only a small number of causal configurations. In essence, HyPrColoc partitions the posterior calculation into two steps; first, by assessing the regional association probability (i.e there is evidence of a shared signal in all, a subset or none of the traits tested) and second, by assessing the "alignment level" (the level to which the same variants influence the set of traits) in line with a single causal variant assumption. Importantly, this approach when compared to MOLOC is extremely computationally efficient; in their example, the authors simulated a set of 100 traits each with 1000 SNPs which managed computation in less than 1 second compared to MOLOC's 1 hour[139]. This becomes especially practical when considering traits which may have shared underlying genetic architecture in need of a formal statistical test, as has been applied recently to uncover a shared aetiology of dementia using AD with stroke, diabetes, alcohol consumption and cholesterol level GWAS[140]. In addition, using QTL

datasets from multiple tissues (such as cross-brain region QTLs) that may all play a role in disease pathogenesis or trait predisposition within the HyPrColoc framework may elucidate novel targets and better explain the underlying biology.

1.7 Mendelian Randomisation

When establishing a link between a molecular phenotype and an observable phenotypic trait it is important to remember that correlation is not causation. There may be several reasons an association between two traits (such as an eQTL and the risk of disease) is observed, under confounding effects or the possibility of reverse causation. In a clinical scenario, the optimal way to deduce whether intervention has a causal effect on disease is through the use of randomized controlled trials (RCT), in which the cohort is randomly divided into the groups such that any unobserved counfounders are equally distributed. Perhaps the best recent examples are the numerous trials for treatment of coronavirus disease 2019 (COVID-19) as a result of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, such as with Remdevisir[141] or Nirmatrelvir[142]. In both cases and as is standard in a RCT, patients were randomly and blindly selected into two groups, one receiving a placebo and the other receiving the drug in question. Given that both groups are treated equally, any observation can plausibly be attributed to a causal effect[143].

In the field of epidemiology it was not uncommon for indications towards a mechanism of action as obtained from observational data to fail clinical translation; for instance, increased intake of β-carotene was suggested to reduce risk of smoking-related answers but was strongly shut down following the results of a controlled RCT[144,145], and as observed in another example in the association between Vitamin E intake and coronary heart disease[146]. This is likely the cause of unobserved confounding, where the effect between a risk factor and an outcome is mediated by other pathways. Studies utilising dietary intake in particular may be strongly influenced by environmental factors, such as socioeconomic positions or behavioural factors that are multi-faceted.

To circumvent this issue, the medical field began the application of what later came to be known as “Mendelian Randomisation” (MR), a method which exploits the random allocation of alleles in accordance with Mendel’s law of independent assortment[147]. This random

allocation, in line with a RCT, allows equal distribution of confounders among the two groups (reference versus effect allele) and can infer a causal relationship between a risk factor and an outcome. An early study by Timpson et al.,[148] applied this concept in the context of circulating c-reactive protein (CRP) and metabolic syndrome. The distinction between causal and correlative observations was importantly exemplified by their results which showed that while an association was well-observed between CRP and metabolic syndrome, this is not the case in a mendelian randomisation framework suggesting no causal association between the CRP levels and metabolic syndrome phenotypes (Figure 1.5).

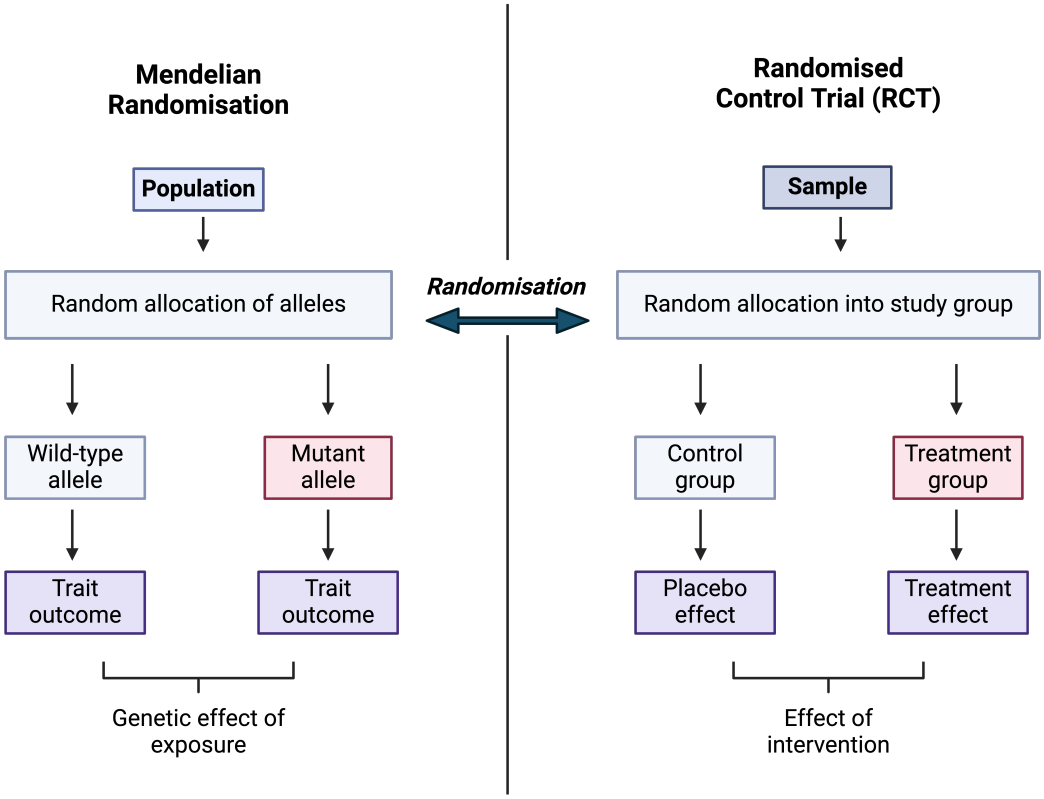


Figure 1.5: Conceptual overview of Mendelian Randomisation. Mendelian randomisation (MR) exploits the natural random allocation of alleles, which are fixed at conception, to mimic a randomised controlled trial (RCT). In this framework, alleles are separated into effect (similar to the intervention group in an RCT) and reference allele groups (the control group in an RCT). This randomisation allows us to assume that any confounding effect is therefore equally spread between the two groups, and allow measuring of the effect of an allele on a trait outcome (in MR), similar to the effect of a drug intervention in an RCT.

The aim of this method is to obtain a valid causal inference between the risk factor (such as CRP) and the outcome (such as metabolic syndrome) as well as effect size, as instrumented by genetic variants[149]. There are several criteria which must be fulfilled to validate an

observed causal inference; (1) There must be a an association between the level of effect alleles and level of measured risk, so that there is a clear difference in the levels of exposure upon separating the population into the two groups (the relevance assumption) and (2) this association is not the result of confounding pathways (exchangeability assumption). Finally, (3) it is assumed that the variant does not directly influence the outcome but rather through the measured exposure (exclusion restriction assumption)(Figure 1.6).

In practice however, only the first assumption (relevance) is directly measurable. The exchangeability assumption is generally assumed to be satisfied by controlling for genetic ancestry, as the majority of GWAS or QTL studies are undertaken in a single major population. However, this assumption may be violated by unobserved genetic structure in the population, such as the result of assortive mating which includes behavioural and socioeconomic influences[150]. The exclusion criterion assumption is by far the strongest but also the most likely to be violated; it is plausible that genetic variants have pleiotropic effects, not only acting on multiple biological pathways other than the risk factor that may affect the outcome (also known as horizontal pleiotropy) but also affecting multiple traits through the same risk factor (vertical pleiotropy). In general, horizontal pleiotropy is considered much more problematic than vertical pleiotropy because it may directly bias causal estimates if there are many parallel exposures[31]. This becomes more complicated when considering the recent proposition of an "omnigenic" model for complex diseases, which suggests that given how disease-associated genes are spread across most of the genome, it is possible that all genes expressed within disease-relevant cell-types have an impact on heritability through the presence of gene regulatory networks[151]. In an MR framework, the implication of an omnigenic model is therefore that no inference made is free from pleiotropic effects, and that genetic effects of variants on genes are all affected to some degree by the expression of other genes in the network. It remains difficult however to measure strength of these pleiotropic effects, be it through a polygenic pathways or omnigenic networks, and the impact they have on the significant signals although several models have been proposed to try and capture these[152,153].

In the context of MR, genetic variants are used as "instrumental variables" (IVs), a term which has been extensively used in the econometric field before making its way to epidemiological research[154,155]. IVs at their core are used to isolate the causal relationship between a dependent variable and an independent variable when there might be confounding

factors. Genetic variants are therefore adequate IVs for several reasons; first, randomly allocated and inherited at birth, an individual's genotype is unaffected by environmental factors and remains constant throughout life; second, this random allocation implies that variants should not typically be associated with competing risk factors (and if present, is more suggestive of environmental factors or other confounders); finally, genetic variants may have biological function through the regulation of genes which may solidify an association to a risk factor[3]. It is important to note that genetic effects are typically very small, and as such MR estimates of association have large confidence intervals (CIs) in comparison to typical epidemiological effect size estimations. As Burgess et al., point out, the rationale behind using MR for causal estimation is that "an unbiased, imprecise estimate is preferable to a precise, biased estimate of causal association"[156].

Similar to colocalisation, MR has seen a sharp increase in popularity owing to the explosion of GWAS over the last 15 years. In addition, while the original method required the analysis to be conducted on the same sample set (with measured data on genotype, exposure and outcome for all individuals), later versions of the method saw two key developments; first, the reformulation of the method to allow the use of summary level data[157] and second, the extension to use two datasets in a "two-sample MR" setting in which one trait acts as the exposure and the second as the outcome. These two innovations have allowed any individual with access to publicly available GWAS summary statistics to infer putative causal relationships between traits[158].

1.7.1 MR in neurology

Perhaps the best way to exemplify the level of popularity MR has enjoyed with the above developments is through examples in neurology. MR has been applied extensively to investigate clinical hypotheses across the neurological spectrum over the last few years. In AD alone, using search terms "Mendelian randomization" and "Alzheimer's disease" in the PubMed database revealed 170 publications in 2023 (499 total since 2006), the majority of which apply available summary statistics for causal inference in MR. A few examples include the causal effect of COVID-19 on AD[159], educational attainment and structural brain reserve on AD[160], alcohol use and AD[161] and even MR applied to aspirin use and AD[162].

While these inferences are useful in understanding co-factors driving pathogenesis, few offer a path to clinical intervention or therapeutic development. In contrast, studies utilizing measured protein or transcript levels in combination with GWAS summary statistics may identify actionable targets, but has been historically difficult to perform at large scale due to availability of post-mortem brain tissue. Most of these applications of MR are termed *cis*-MR because they utilize genetic variants within a single locus. As the majority of studies are underpowered for *trans*-QTL associations, they can rely on *cis*-QTL associations to adhere to the relevance assumption in which the molecular trait of interest is treated as the risk factor. A recent study by Wingo et al., (2021) performed a large-scale pQTL study in post-mortem dorsolateral prefrontal cortex samples of 376 individuals, quantifying the levels of 8,356 proteins. Through the application of two-sample MR, they identified 9 genes with a putatively causal role in AD[163]. A similar study was also performed combining brain, cerebrospinal fluid (CSF) and plasma protein levels from particiants, identifying 34 proteins with a putatively causal role in PD as well as stroke[82].

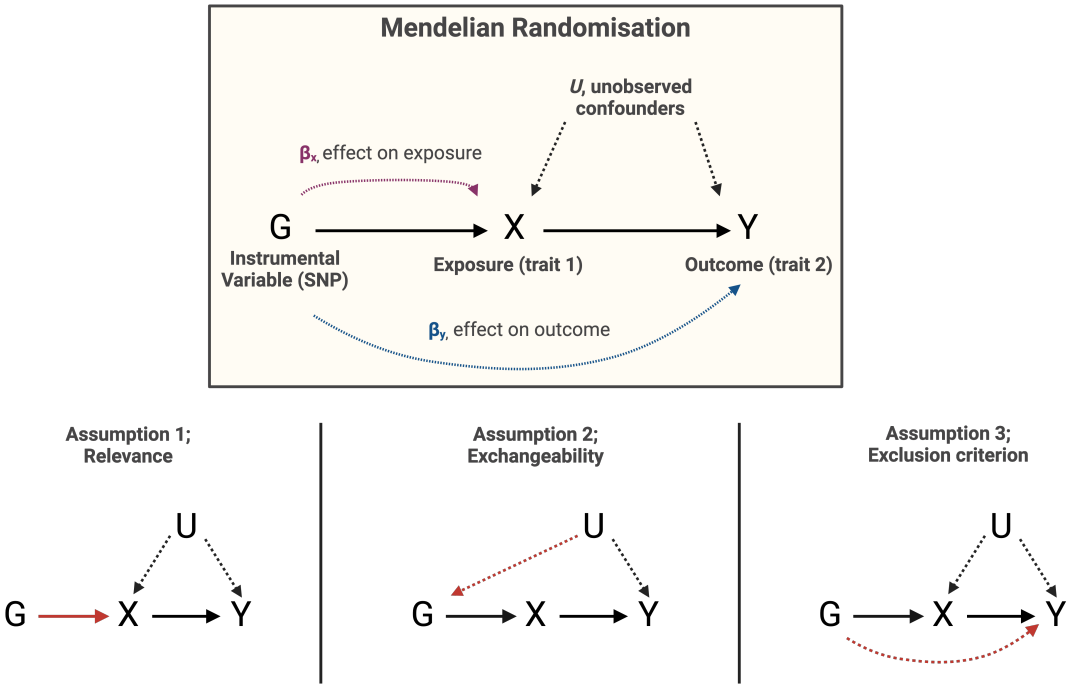


Figure 1.6: Overview of MR model. In a Mendelian Randomisation framework, the goal is to assess the causal effect of the exposure (X) on the outcome (Y), using a genetic variant (G) as an instrumental variable. For genetic variants to be valid IVs, they must satisfy three core assumptions; the relevance assumption, stating that the variant G is associated with the exposure of interest X; the exchangeability (or independence) assumption, which states that there are no confounders of the association between the IVs and the outcome Y; and finally the exclusion criterion assumption, which states that the effect of the genetic variant G on the outcome Y is only through its effect on the exposure X.

1.7.2 MR methodology

Conceptually, MR generally requires a single genetic variant as the instrument for a measured risk factor; as an example, an early study of genetic variants in the *PCSK9* gene through MR showed a strong association to blood concentrations of LDL cholesterol and coronary heart disease (CHD), further supported by later phase 3 clinical trials[165]. The simplest way of calculating the causal effect of the exposure (X) on the outcome (Y, as shown in Figure 1.6), when using a single IV (Z) is known as the Wald ratio method (Equation 1.2):

$$Wald\ Ratio = \frac{\hat{\beta}_{y|z}}{\hat{\beta}_{x|z}} \tag{1.2}$$

In this equation, $\hat{\beta}_{y|z}$ represents the beta coefficient of the IV in the regression of the exposure, and $\hat{\beta}_{x|z}$ represents the beta coefficient of the IV in the regression of the outcome. We can then interpret the ratio method output as the change in outcome resulting from a unit increase in the exposure is equal to the change in exposure for a unit increase in the IV, scaled by the change in exposure for a unit increase in the IV. Put simply, the ratio method measures the change in outcome (trait 2), as a result of changes in exposure (trait 1), which itself is the result of changes in allele dosage (the genetic variant as the IV). Note that this simplified method excludes standard error calculations, which are routinely provided to estimate confidence intervals. As per the previous example of LDL-C and CHD;

$$Wald\ Ratio = \frac{\hat{\beta}\text{CHD}|z}{\hat{\beta}\text{LDL-C}|z}$$

Where:

$\hat{\beta}_{\text{CHD}|z} = 0.095$ (log-odds increase in CHD risk per allele), and $\hat{\beta}_{\text{LDL-C}|z} = 0.2$ (mmol/L increase in LDL-C per allele)

Thus:

$$Wald\ Ratio = \frac{0.095}{0.2} = 0.475$$

Which represents a 0.475 log-odds increase in CHD risk per mmol/L increase in LDL-C.

Importantly, the first equation exemplifies how it is possible to perform two-sample MR using only summary statistics, since only the beta coefficients for each trait (and their standard errors) are necessary and one of the reasons MR has become so popular as of late. There are also multiple other methods for single instrument causal estimation, including the original (albeit somewhat outdated) two-stage least squares (2SLS) method, maximum likelihood approaches and non or semi-parametric methods[166].

However, the use of single genetic variants as IVs may be problematic if the sample size is small; low statistical power leads to weaker associations between the genetic variant(s) and trait in question, where a single IV may only explain a small proportion of the effect. Although IVs due to their nature and satisfaction of MR assumptions are unbiased by confounding, the lack of certainty around the effect size as a result of finite sample sizes can lead to what the field has termed "weak instrument bias", where bias is the difference in observed effect size and the "true" value.

As a result, a multitude of methods have been developed to allow incorporation of multiple IVs. By 2022, there were 47 published MR methods[167], the majority of which were developed for the use of public GWAS summary statistics in a two-sample setting. One of the earliest and simplest methods which is still used routinely is the inverse variance weighted (IVW) MR, which combines the Wald ratio in fixed effects meta-analysis (where each IV is treated as a separate "study" as in traditional meta-analysis), where the weight of each ratio is the inverse of the variance in the IV-outcome association. The Wald ratios are then meta-analysed under a fixed effects model to produce a causal estimate[157]. In the last 5 years, among the most popular approaches to emerge are MR-Egger[168], MR-PRESSO[152], Bayesian MRMix[169], and more recently weighted MR (BWMR)[170], to name a few. Each of these employ different approaches when it comes to IV selection, statistical inference and importantly modelling and controlling for horizontal pleiotropy.

1.8 MR and colocalisation; similar goals but contrasting outcomes

In many ways, MR and genetic colocalisation have a conceptual overlap; both attempt, through the use of genetic variants, to establish a link between two traits but encompass important differences. In MR, the absolute aim is to identify a causal effect and as such adheres to strict statistical assumptions. For instance, the choice of IVs is dependent on their association to the exposure (relevance assumption) but does not necessitate an association to the outcome. This contrasts the approach of colocalisation, which usually requires a signal from both traits in the region of interest and in most cases, only loci of genome-wide significance in the GWAS are selected for colocalization analysis[134]. In addition, it is common to only retain independent variants predictive or associated to the exposure through LD thresholding, as including them may bias the effect size estimate and violate confounder assumptions[171]. Colocalisation on the other hand assumes there is a causal variant shared by the two traits within the region, and as such expects a large number of variants present in the region for the analysis.

Another clear difference is the asymmetric nature of MR in comparison to colocalisation. In a colocalisation test, both traits can be interchanged and still output the same result, whereas in MR a directionality is implied (exposure causes outcome). If one were to flip the test in MR, the question asked by the analysis would be drastically different; for example, "Is gene X expression causal for PD?" seeks a different answer than "Is PD causal for gene X expression?", whereas colocalisation asks "Is gene X expression and PD genetically linked?". This specific example also highlights the potential for reverse causation; it is entirely possible that physiological changes such as gene expression are only observed in the presence of disease and not necessarily causal for the disease itself[172,173]. While the assumptions of MR strictly necessitate measured exposures free from unobserved confounders (such as the presence of disease), colocalisation does not. It would therefore be possible to observe QTL associations that only arise by disease presence and consequently result in a positive colocalisation signal with a target GWAS.

These methods have however often be used jointly to strengthen the observed inferences[149]. For instance, Gaziano et al., performed a combined colocalisation and MR analysis in an

attempt to find causally associated proteins to COVID-19[174]. They first performed genome-wide MR analyses using GTEx *cis*-eQTLs as instruments (restricted to eQTLs regulating genes within the druggable genome as the exposure and COVID-19 hospitalisation as the outcome trait) and found 6 genes to be statistically significant after multiple testing correction. Further, colocalisation analyses showed that three genes (*IL10RB*, *IFNAR2* and *ACE2*) had strong evidence (high PP.H4) of colocalisation with COVID-19 hospitalisation, supporting a shared causal variant. The *ACE2* gene was already a well-known target for therapeutic development against COVID-19, coding for the primary cell-surface receptor of SARS-CoV-2 which has been the subject of multiple clinical trials[175]. However, the *IL10RB* and *IFNAR* genes shared *cis*-eQTLs used as IVs in the MR analysis and begged the question of which gene was the primary driver for the association with COVID-19 hospitalisation, as well as being a good example of horizontal pleiotropy. Downstream pQTL and pathway enrichment analyses, as well as the growing body of evidence led the authors to conclude that *IFNAR2* was more likely to be the causal gene[177].

In another contrasting example, Zuber and colleagues demonstrate a strong causal association between LDL cholesterol and CHD risk (which as previously mentioned has been demonstrated in the *PCSK9* gene region), as well LDL cholesterol and AD risk driven by variants in the *APOE* region. While there was also strong colocalisation support between LDL cholesterol and CHD risk in the *PCSK9* region, this region did not have any strong associations to AD. When performing colocalisation between LDL-cholesterol and AD risk around the *APOE* region, where there were variants driving the MR association, a divergent association pattern was revealed suggesting two distinct causal variants for each trait and suggesting violation of MR IV assumptions due to LD[149].

1.9 Aims and objectives of thesis

The work conducted in this thesis overarches four main themes;

1. The cell-type specific eQTL landscape of the brain. In this first chapter, I aim to assess the genetic regulatory landscape of cell-type specific gene expression in the brain, utilising snRNA-seq generated on post-mortem brain from 391 individuals with jointly assessed genotype.

2. State-specific eQTLs, and differences between brain and control cohorts. The second chapter aims to dissect the dynamic nature of cell-type specific eQTLs in the post-mortem brain, by interrogating whether the presence of disease samples impacts association signals.

3. The genetic colocalisation landscape across the brain phenome, and the impact of state-specific eQTLs on colocalisations. Following generation of genome-wide eQTLs, I aim in this third chapter to integrate eQTL signals with associations to traits across the brain phenome, utilising publicly available GWAS summary statistics.

4. Mendelian Randomisation and target prioritization. In this final chapter, I use the colocalisation results to prioritize genetic variants to use as instrumental variables in the first ever MR study done using snRNA-seq derived *cis*-eQTLs in the brain.

Chapter 2

Methods

2.1 Materials and Methods

This section contains an overview of the methodology prior to cell-type specific eQTL mapping.

2.1.1 Dataset overview

The snRNA-seq data generated for this chapter encompasses both in-house generated snRNA-seq and processed snRNA-seq data obtained from collaborators (listed in the preface of this thesis). In total, there are samples from 409 individuals which are split as follows (Table 2.1);

Table 2.1: Overview of snRNA-seq dataset for eQTL discovery

Dataset	Description	Number of individuals
MRC_60	Prefrontal cortex and hippocampus samples generated in-house	60 (controls-only)
ROCHE_92	Temporal cortex samples from a PD cohort selected in-house and sequenced by Roche	92 (70 PD cases, 22 controls)

Dataset	Description	Number of individuals
MATTHEWS	Cortex (entorhinal and somatosensory) samples on an AD cohort (with controls), some of which have been published[178,179]	65 (38 AD cases, 27 controls)
BRYOIS_192	Mixture of temporal and prefrontal cortex samples (detailed as per publication), obtained from our collaborators and publicly available[108]	192 (105 MS and AD cases, 87 controls)

2.1.2 Sample extraction and preparation

Snap-frozen human brain tissue samples were all obtained from tissue banks following ethical approval and material transfer agreements set out by the institutions. All statutory and ethical regulations for use of human post-mortem tissue were approved by the Imperial College Research Ethics Committee. Samples at each brain bank were collected post-mortem and snap-frozen in liquid nitrogen for 20 minutes prior to long-term storage at -80C. Sample blocks were assessed using immunohistochemistry, specifically looking at beta-amyloid, tau, TDP43, alpha synuclein and p62, to confirm clinical diagnoses. Control samples were then defined based on their clinical history, and/or the presence of disease upon neuropathological examination.

2.1.3 Genotyping

For the *MRC_60* dataset, donor DNA from each individual was processed in-house and used for genotyping using the Illumina Infinium Global Screen array v2.0. Both the *BRYOIS_192* and *ROCHE_92* datasets were genotyped as described previously[108]. For the *MATTHEWS* dataset, DNA was extracted using DNeasy kit following manufacturer's instructions, and genotyped with the UK Biobank Axiom Array.

All genotypes from each dataset were imputed on the Michigan Imputation Server (version 1.6.3) using Minimac4[180] using a European ancestry reference panel from the Haplotype

Reference Consortium (r1.1), pre-phased with Eagle (version 2.4)[54,181]. Following retrieval of imputed genotypes, the files from each dataset were merged into a single vcf before quality control using PLINK (version 2.0)[182]. First, SNPs with an imputation score < 0.4, genotype missingness $> 5\%$ and deviating from Hardy-Weinberg equilibrium (p-value $< 1 \times 10^{-6}$) were removed, as well as multi-allelic SNPs (more than 2 alleles). In addition, individuals with more than 2% of missing genotypes were removed. Before all vcf files were combined, each quality controlled vcf underwent a check using bcftools[183] to ensure allele strand orientation matched the latest ENSEMBL hg38 genome build[184], flipping alleles when mismatches were found. In the merged genotype file, I also performed a kinship analysis to identify duplicated samples (samples in different datasets that by chance came from the same individual) and excluded individuals with kinship > 0.2 to retain independent individuals. Finally, only SNPs with a minor allele frequency (MAF) $> 5\%$ were kept. To confirm that all samples were of European Ancestry, I further merged the genotype with genotypes from the 1000 genomes project[185] and conducted a genotype PCA analysis as described in Chapter 3.

2.1.4 Generation of single-nuclei RNA-seq data

For the *MRC_60* dataset, snRNA-seq was generated in-house at Imperial College. From 60 individuals, both prefrontal cortex and hippocampus samples were collected at the tissue banks including Imperial College London, University of Oxford, University of Edinburgh or Amsterdam Medical Centre brain banks. Nuclei were isolated as described in a previous publication from our group[186]; briefly, brain tissue sections were suspended in homogenisation buffer in preparation for tissue homogenisation, which was then applied to a cell strainer cap for nuclei isolation. Samples were then subject to snRNA-seq generation using the 10X Single Cell Next GEM Chip aiming to recover 5,000 minimum nuclei per sample. Sequencing libraries were then prepared using the Chromium Single Cell 3' Library and Gel Bead v3 kit following manufacturer's instruction. The subsequent cDNA libraries were then sequenced with the Illumina NovaSeq 6000 targeting a minimum sequencing depth of 30,000 paired-end reads per nucleus.

For the *ROCHE_92* dataset, nuclei were isolated from 10mg fresh-frozen tissue using the Nuclei Pure Prep Isolation kit (Sigma Aldrich) with an optimized protocol. The tissue was lysed in Nuclei Pure Lysis Solution with 0.1% Triton X, 1mM DTT 0.4U/ l SUPERase-In$^{\text{tm}}$

RNase Inhibitory (ThermoFisher Scientific). The mixture was then homogenized in two steps, first with a 23G and then a 29G syringe. Further, 1.8M Sucrose Cushion Solution was added before filtering of nuclei through a 30 m strainer. The lysates were then carefully layered on top of 1.8M Sucrose Cushion Solution prepared in new eppendorf tubes, after which they were centrifuged for 45 minutes at 16,000xg (4'C). Nuclei pellets were then re-suspended in nuclei storage buffer, including RNAse Inhibitor, and transferred to new Eppendorf tubes before an additional two rounds of centrifugation for 5 minutes at 500xg (4'C). Purified nuclei were then again re-suspended in nuclei storage buffer with RNAse inhibitor, before staining with trypan blue and counted using Countess II (Life technology). Nuclei were then permeabilized as per the ATAC + Gene Expression sequencing kit from 10X genomics for the transposition step, targeting approximately 12,000 nuclei per sample. This was then loaded onto the Chromium Next GEM Single Cell Chip J. ATAC and expression libraries were then constructed according to the manufacters instructions. Libraries were then sequenced using Illumina NovaSeq 6000 aiming at a minimum sequencing depth of 30,000 paired-end reads per nucleus.

For the *MATTHEWS* dataset, fresh frozen tissue blocks were cryosectioned at 8 m and 200mg of grey matter was placed in RNAse-free Eppendorf tube. The tissue was lysed as described in our collaborators previous publication[178], and then homogenised in buffer containing 1% Triton X-100, 0.4 U/ l RNAseIn + 0.2U/ l SUPERaseIn and 1 l of 1mg/ml DAPI staining using a 2ml glass douncer. The homogenate was then centrifuged at 500xg for 8 minutes (4'C), and the pellet was resuspended in the homegenisation buffer before filtering through a 70 m filter. The filtered nuclei were then further centrifuged at 13,000xg for 40 minutes, and the pellet was washed and filtered in PBS buffer (PBS + 0.5mg/ml BSA + 0.4 U/ l RNAseIn + 0.2U/ l SUPERaseIn). Nuclei were then once again pelleted, washed with the PBS buffer (twice) before resuspension in 1ml PBS buffer. A small sample of 100μl was taken for snRNA-seq processing. First, nuclei were stained with Acridine Orange dye and counted using the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems, L20001). Around 7,000 nuclei were used per sample with the 10X Genomics Chromium Single Cell 3' kit as per manufacturers instructions. Libraries were then sequenced on an Illumina HiSeq 4000.

2.1.5 Single-cell analysis and cell-type identification

For the *MRC_60* dataset, raw sequencing reads were mapped to the GRCh38 genome where read counts were quantified as UMIs using the *CellRanger count* function (version 5.0.1), including reads that mapped to intros using the *–include-introns* flag, which has been shown to result in more genes detected per nucleus and improved cell type classification[187,188]. In the *MRC_60* dataset, samples that were pooled were demultiplexed using *Demuxlet* with default parameters, using the genotype data to assign cells to each individual[189]. For the *ROCHE_92* dataset, raw sequencing files were mapped using *CellRanger-ARC count* (version 2.0.2), due to being 10X multiome data and also containing ATAC-seq reads. After mapping, only RNA counts were retained. Both the *MATTHEWS* and *BRYOIS_192* dataset were mapped as previously described[108,178], and the *BRYOIS_192* data was obtained as a processed Seurat object with annotated cell-types.

On the mapped single-cell data, per individual in the *MATTHEWS*, *MRC_60* and *ROCHE_PD* datasets I also assessed the proportion of empty droplets using the *EmptyDrops* package, which returns a *p-value* and *FDR* for droplets deemed to contain nuclei[190]. Nuclei with FDR < 0.01 were retained, and then Seurat objects were constructed using the *Seurat* package (version 4.0)[191], retaining nuclei with at least 500 UMIs, 300 genes and less than 5% mitochondrial content, and removed doublets using *DoubletFinder*[192]. Samples were then integrated, within each dataset, into a single seurat object using reciprocal PCA (rPCA) and excluding samples with less than 500 cells. This merged Seurat object from each dataset was then used to assign cell-types using canonical expression markers as below (Table 2.2). The markers chosen here are an abbreviated version of the brain cell-type specific markers published by Wang et al., (2018)[193], and follow closely the list of markers used within a similar study for cell-type specific eQTL analysis in the brain by Bryois et al., (2022)[108].

Table 2.2: Markers used in cell-type annotation of single-nuclei data

Cell-type	Gene markers
Excitatory Neurons	*SLC17A7, SATB2*
Inhibitory Neurons	*GAD1, GAD2*
Astrocytes	*AQP4, FGFR3*
Oligodendrocytes	*MAG, MOG*

Cell-type	Gene markers
Oligodendrocyte Precursor Cells (OPCs)	*PDGFRA, VCAN*
Microglia	*C1QB, CSF1R*
Endothelial cells	*CLDN5*
Pericytes	*RGS5*

As a final quality control step, I performed ambient RNA removal using *DecontX*[194] on all datasets (including *BRYOIS_192*), which uses given cell-type labels to estimate the proportion of ambient RNA counts in each population and returns a corrected count matrix.

2.1.6 eQTL mapping

For each cell-type in each dataset, ambient RNA corrected count matrices were extracted and summed within each individual to obtain a single count value ("pseudobulking"), and only individuals with at least 10 cells in that cell-type were retained. Aggregated count matrices were then normalized using counts per million (*cpm*) from the *edgeR* package[195], and then log-transformed and scaled. Further, I only retained genes expressed in at least 5% of individuals.

For eQTL discovery, a two-step process was followed. First, I fitted a linear model for the expression of each gene against all individual covariates (both clinical and technical) including age, sex, post-mortem interval (PMI), sample source and diagnosis as fixed effects. The residuals were then obtained as "corrected" gene expression values. Then, I calculated the principal components (PCs) of expression on the full matrix in each cell-type to be used as fixed effects in the eQTL discovery. In addition, genotype PCs were also calculated to be used as covariates in the modelling. Using *MatrixEQTL*, *cis*-eQTL discovery was performed using a *cis*-distance of 1Mb (meaning, 1Mb from each end of the gene) with the residual expression for each cell-type, in the following model (Equation 2.1);

$$E \sim G + \text{expression PCs}(1-N) + \text{genotype PCs}(1-5) \tag{2.1}$$

Where E represents the residual gene expression (after correcting expression through linear modelling against covariates), G is the genotype dosage of the reference allele (0, 1 or 2) for a specific SNP, expression PCs (1-N) is dependent on the optimal number of PCs for each cell-type (tested prior to the final MatrixEQTL model and determined by the number of PCs yielding the highest number of discovered eQTLs). The final set of eQTL results were then filtered by False Discovery Rate (FDR), retaining associations below five percent[196].

2.1.7 Epigenetic overlap of eQTLs

To interrogate the discovered eQTLs and their epigenetic overlaps, I utilised the *annotatR* package[197]. This package collates genomic information from a number of databases including UCSC for gene and transcript information, GENCODE for lncRNA features, FANTOM5 for enhancer features and chromMM for chromatin state features, and combines them into a readable format for annotation of genomic positions. To this end, I annotated the positions of each SNP in the eQTL outputs (hg38) and intersected these positions with the *annotate_regions()* function.

2.1.8 Validation of eQTLs with existing bulk data

Replication of eQTLs was assessed using a published eQTL study conducted on bulk RNA-seq assessed on prefrontal cortex samples from 6,523 individuals[90]. Genome-wide summary statistics for the *cis*-eQTL associations were downloaded and reformatted from the metabrain website (https://metabrain.nl/). First, only SNPs assessed both in metabrain and in my *cis*-eQTL discovery were retained. The percentage replication for each cell-type was then calculated as the number of SNP-gene pairs significant in both datasets (FDR $< 5\%$) divided by the total number of eQTLs found in the cell-type specific eQTLs. In addition, allelic information was checked for each SNP to ensure the same effect allele was assessed in both studies, flipping the beta sign if alleles were mismatched.

2.1.9 Random effect and interaction modelling

For the modelling of random effects and interaction with disease status on eQTL associations, I utilised the eQTLs generated on the full dataset, including both disease and control samples (n=391). First, *cis*-eQTL associations were filtered to only keep significant associations (FDR < 5%), and only the top SNP for each gene was retained. Then, each association between the top SNP and each gene, per cell-type, were re-assesed using a series of linear mixed effects models (LME) using the *lmerTest* package[198].

First, each SNP-gene association was re-assessed in baseline model called the "null model" (M_0), where the gene expression from each individual (normalised aggregated counts) in each cell-type was modelled against all covariates (Equation 2.2);

$$E \sim 1 + \text{Diagnosis} + \text{Age} + \text{PMI} + \text{Sex} + (1 + \text{Diagnosis}|\text{Sample Source}) + \text{expression PCs}(1 - N) + \text{genotype PCs}(1 - 5) \tag{2.2}$$

Where in addition to testing each clinical covariate as a fixed effect, including diagnosis, there is an additional random effect on Sample Source (the originating brain bank), conditioned by diagnosis. This allows for a random slope for diagnosis within each level of sample source and can be interpreted as accounting for unknown bias in neuropathology performed across brain banks by different clinicians. This null model M_0 assesses whether any genes are associated to any of the covariates, prior to the addition of a genotype. In the replication model (M_1), the model now includes the addition of genotype dosage, with the aim to replicate the two-step approach for the MatrixEQTL associations (Equation 2.3);

$$E \sim 1 + \mathbf{G} + \text{Diagnosis} + \text{Age} + \text{PMI} + \text{Sex} + (1 + \text{Diagnosis}|\text{Sample Source}) + \text{expression PCs}(1 - N) + \text{genotype PCs}(1 - 5) \tag{2.3}$$

And finally, the interaction model (M_2), where an interaction term is included between diagnosis and genotype (Equation 2.4);

$$E \sim 1 + \mathbf{G} * \text{Diagnosis} + \text{Age} + \text{PMI} + \text{Sex} + (1 + \text{Diagnosis}|\text{Sample Source}) + \text{expression PCs}(1 - N) + \text{genotype PCs}(1 - 5) \tag{2.4}$$

Where the interpetation is that significant interaction indicates these two variables are not independent, meaning the effect of genotype on gene expression varies depending on diagnosis. To then assess the reliability of these models, I used a likelihood ratio test (LRT) for model selection[199]. In each case, the baseline model was tested against the more complex model (M_0 versus M_1, and if significant, M_1 versus M_2), where a significant p-value (< 0.05) suggests the complex model is statistically more appropriate for the association.

For the age-related interaction term used on *cis*-eQTL associations discovered in the control-only dataset (n=183), the following model was implemented (Equation 2.5):

$$E \sim 1 + \mathbf{G} * \text{Age} + \text{PMI} + \text{Sex} + (1|\text{Sample Source}) + \text{expression PCs}(1 - N) + \text{genotype PCs}(1 - 5) \tag{2.5}$$

Finally and as implemented by Nathan et al.,[109], LRT p-values were further corrected for multiple-testing using the *qvalue* package[200] where LRT q-value < 0.05 was considered to represent a significant interaction for an association.

2.1.10 GWAS processing

All GWAS studies were curated using the following approach. First, traits were queried on the EBI GWAS catalog (https://www.ebi.ac.uk/gwas/), where the GWAS summary statistics were downloaded, if available. Otherwise, summary statistics were downloaded as indicated in the original publications. Where possible, the latest GWAS was used and a full list of traits analysed including original publication are included in (Appendix 5). All GWAS summary statistics were uniformly processed using the *format_sumstats()* function from *MungeSumstats* package, which queries column names against a comprehensive list of naming conventions[201]. In cases where either the effect sizes (betas) and standard errors were missing, I implemented the *impute_se=TRUE* and *impute_beta=TRUE* to enforce imputation using adjoining values

(such as Z-score or Odds Ratio). In addition, GWAS summary statistics where SNP positions were on hg19 were lifted over to hg38 using the internal *convert_ref_genome* parameter.

2.1.11 Colocalisation analysis

For the colocalisation analysis, each processed summary statistics file from each GWAS first underwent clumping using the *ld_clump()* function from the *ieugwasr* package[202] to avoid overlapping regions (for instance, a GWAS may have 5,000 genome-wide significant SNPs, which are in 75 loci). First, each GWAS was scanned to find regions of genome-wide significance (p-value $< 5 \times 10^{-8}$), and clumping was performed to retain the top variant in a 1Mb region. Then, each window was re-populated with the full set of SNPs to retain complete loci to test for colocalisation analysis. In addition, regions whith less than 100 SNPs were excluded from colocalisation analysis to avoid unstable results.

For each GWAS, the outputs from MatrixEQTL (both the full dataset, n=391, and the control dataset, n=183) were first intersected with each clumped GWAS. Each GWAS region was then assigned genes within the 1Mb window and the cell-type eQTL associations. Colocalisation was then carried out using the *coloc.abf* function from the *COLOC* package[134], each time testing the associations from one cell-type / gene combination versus the GWAS associations. For instance, in a GWAS locus containing 20 genes, colocalisation would be performed through 20 (number of genes) $\times$8 (number of cell-types) = 160 colocalisation tests. The *COLOC* package requires specifying the "type" of trait measured; for all case/control GWAS studies, *type="cc"* was specified, and *type="quant"* for quantitative/continuous trait measurements, including the eQTL associations. Finally, traits were considered colocalised where a posterior probability (indicated by PP.H4) was above 0.8. In addition, genes within the *MAPT* and *HLA* loci were aggregated into single regions (*"MAPT-region"* and *"HLA-region"*, respectively) because of inaccuracy of transcript quantification in these regions[203,204], but are reported individually in the appendices.

2.1.12 MR analysis

Because MR analysis requires unbiased estimates to infer putative causal associations, MR was only performed using eQTLs from the control-only cohort (n=183). Control-only eQTLs underwent colocalisation as described above, and regions with colocalisations (cell-type/gene/GWAS) with a posterior probablity (PP.H4) > 0.8 were retained for MR analysis. For each colocalised region, the lead colocalised SNP (as proposed by *COLOC*) was retained. First, all SNPs in the region were assessed for MR analysis. To ensure the MR analysis was centered around the lead COLOC SNP, the eQTL FDR was set to 0 and all other eQTLs with an FDR $> 5\%$ were removed. In addition, only independent SNPs were retained to avoid inflating the MR estimates, removing all SNPs correlated with the lead SNP using the *ld_clump* function ($r^2 > 0.01$). Then, alleles were harmonized, flipping effect sizes when the GWAS and eQTL effect alleles were mismatched. Finally, MR was carried out in a two-sample setting with the eQTL effect sizes as the exposure and GWAS effect sizes as the outcome using the *mr_ivw()* function from the *MendelianRandomization* package[205]. Because more than 90% of tests only retained one IV (genetic variants), only the IVW MR method was considered. Finally, given the strict selection of associations to test in MR through colocalisation, a nominal IVW p-value < 0.05 was interpreted as a significant putatively causal association.

2.1.13 Assessment of MR genes at protein level

To interrogate whether the inferred MR associations would also be valid at protein level, I obtained summary statistics for pQTL analyses conducted in brain on a cohort of 330 individuals[206], as well as pQTL analyses conducted on circulating plasma protein as reported by the UK Biobank (UKB) Pharma Proteomics Project (UKB-PPP)[207]. For each summary statistic file, I retained associations where the protein assessed was a also a significant MR gene in my single-cell eQTL analyses. For the UKB-PPP associations, multiple testing correction was first done on the associations (which only reported nominal p-value associations) using the base *R p.adjust()* function, specifying "*fdr*" to conduct the Benjamini-Hochberg multiple testing correction[196]. Similarly to the single-cell eQTL MR analyses, 1Mb regions were defined around GWAS summary statistics as previously.

Associations were filtered to < 5% FDR, and independent genetic variants were kept as IVs using the *ld_clump()* function from the *ieugwasr* package, removing all correlated SNPs ($r^2 > 0.01$). Effect sizes were harmonized depending on reported effect alleles, flipping the direction if mismatched. Again, once a set of IVs were defined, two-sample MR using the *mr_ivw()* function from the *MendelianRandomization* package[205] was conducted with the association to protein abundance as the exposure, and the trait as the outcome, finally retaining MR associations with IVW p-value < 0.05. In addition, colocalisation was performed as described above for the UKB-PPP pQTL associations, with the small modification of setting the prior probability $p_{12} = 0.01$ given the previous evidence of cell-type eQTL colocalisations.

2.1.14 Epigenetic annotation of IVs

Because IVs are not necessarily the causal and functional SNP, IVs were intersected with a range of epigenetic data. For each MR hit, the IV(s) used in the MR analysis were queried using the *ensemblQueryLDwithSNPwindow()* function from the *ensemblQueryR* package (https://github.com/ainefairbrother/ensemblQueryR), a custom package to allow submission of user queries to the Ensembl API via R. Setting an LD r^2 threshold > 0.9, all correlated genetic variants were obtained for each MR hit, and used in overlaps with epigenetic data.

Table 2.3: Impact scoring sheet for regulatory variants

Variant consequence	Impact
• transcript ablation or amplification	HIGH
• splice acceptor/donor	
• stop or start gain/loss	
• frameshift	
• feature elongation/truncation	
• Inframe deletion/insertion	MEDIUM
• missense variant	
• protein altering	

Variant consequence	Impact
• Inframe deletion/insertion	LOW
• missense variant	
• protein altering	
• Splice "region" variant	
5' or 3' UTR variants, non-coding exon or intronic variants, proximity to gene, regulatory region modification, TFBS modification, intergenic variant	MODIFIER

The "Variant Effect Predictor" (VEP)[69] command line tool was first used to query each IV(s) and correlated SNPs, by iteratively submitting queries to the Ensembl database and compiling the results into a single table. Variant consequences were defined as per Ensembl (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences), summarized in Table 2.3.

For SNP2TFBs[208], data was downloaded directly from the SNP2TFBs portal (https://epd.expasy.org/snp2tfbs/). To assess whether an IV and/or its correlated genetic variants disrupted transcription factor binding, a straight intersection was applied with the SNP2TFBs table. In deciding the most "impacted" transcription factor, I prioritized first transcription factors that came up the most for all the correlated variants, and in case of ties prioritised the transcription factors disrupted by the genetic variant with the highest correlation (r^2) to the IV.

To intersect with cell-type specific epigenetic data by Nott et al.,[209], I downloaded filtered peak files as per their publication available here https://github.com/nottalexi/brain-cell-type-peak-files. For each IV, I obtained all correlated genetic variants ($r^2 > 0.9$) and overlapped these with the coordinates from the epigenetic peaks. I considered an overlap from any of the correlated variants for each IV with an epigenetic peak to be a successful intersection.

2.1.15 Code availability

The majority of the code used in these analyses can be found here https://github.com/johnsonlab-ic/singlecell-MR, excluding figure-specific scripts which I can provide upon request.

Chapter 3

Single-nuclei eQTL discovery in the post-mortem human brain

3.1 Preface

3.1.1 Appendices

The appendices referenced in this chapter are "Appendix 1", which contains the top eQTL associations for each cell-type, restricted to the top SNP for each significant eGene.

3.1.2 Acknowledgement of contributions

The work conducted in this chapter is on the analyses of snRNA-seq generated in post-mortem brain from various data sources; the in-house data (MRC_60) was generated in the lab by Dan Clode and Jeong Hun Ko. The data from BRYOIS_192 was generously shared by Julien Bryois and Dheeraj Malhotra, as well as the data ROCHE_92 dataset which was generated by their team but processed by myself. Finally, I would also like to acknowledge Paul Matthews and his team (Johanna Jackson, Nurun Fancy, Amy Smith) who supplied me with the the MATTHEWS dataset.

3.2 Introduction

The majority of genetic variants discovered to predispose genetic risk are non-coding and are therefore assumed to constitute a regulatory function in relation to molecular expression[2,74]. Discovery of QTLs has therefore been a primary motivation for the past twenty years, as evidenced in multiple tissues at large-scale[4,90] and a range of modalities[77,83,206], as well at the single-cell level[101,107,108].

In the brain, eQTL studies making use of whole tissue RNA-seq have reached thousands of individuals[90], uncovering an almost complete map of regulatory SNP-gene associations. Extending this to a cell-type specific context, Bryois et al., generated what has been termed as the first single-cell eQTL study in the post-mortem brain, with jointly generated snRNA-seq and genotype information in 192 individuals[108]. However, compared to its bulk counterparts it is of modest statistical power and it is likely many eQTL associations remain to be discovered with cell-type specific context.

At time of generation, the dataset presented here is the largest single-cell dataset in the post-mortem brain with jointly generated genotype information. In this chapter, I aimed to map the regulatory landscape of gene expression in 8 major cell-types (excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, oligodendrocyte precursors, microglia, endothelial cells and pericytes) in the brain, using snRNA-seq generated on post-mortem brain tissue from 391 individuals with jointly generated genotype information.

3.2.1 Aims and objectives:

The broad aims of this chapter are to:

- Assess the genetic *cis*- regulatory landscape of cell-type specific gene expression in the human brain, using joint genotype information and snRNA-seq generated on 409 individuals (391 after quality control).
- Evaluate patterns of regulation across cell-types, including;
  - The identification of shared and distinct eQTLs across cell-types

  - Investigating the epigenetic functions of SNPs that may drive associations

- Assess the level of replication of cell-type specific eQTLs in published bulk eQTL studies of the brain

3.3 Results

3.3.1 Cohort overview

Following quality control the cohort encompassed 391 individuals, split into 183 individuals with no clinical diagnosis of brain disease (which will later be referred to as the control-only cohort), and 208 subjects diagnosed with either MS (n = 48), AD (n = 92) or PD (n = 68) (Figure 3.1 A). This dataset was roughly split equally between males and females (188 females, 48% and 203 males, 52%) (Figure 3.1 B). All brain tissue was collected post-mortem where the most common cause of death was old age; as such, the majority of samples were above 65, with a mean age of 74 (median 77) and ranging between 31 and 104 years old (Figure 3.1 B). This was similar across diagnosis categories as well; Controls (31-104, mean 72, median 74); AD (43-101, mean 81.7, median 85); MS (38-88, mean 59.3, median 61) and PD (63-93, mean 78.9, median 78). Controls were generally well distributed across all ages, whereas PD and AD samples were generally more commonly around the age of 80, with the exception of a small proportion of early onset AD cases. Younger individuals were more common as expected in the MS samples, which routinely presents itself between the ages of 20 - 40[210].

Using the full genotype data per patient, I confirmed that these samples were of European ancestry (as indicated by clinical information) by merging with cross-ancestry genotype data from the 1000 Genomes Project[185]. Plotting the first two eigenvectors of a genotype PCA shows that the samples clustered clearly within the European ancestry population (Figure 3.1 E).

Following genotype quality control and a minor allele frequency (MAF) threshold of >0.05, 5.22M SNPs were retained (Figure 3.1 H) across the 22 autosomal chromosomes. The distribution of MAFs showed an enrichment of SNPs with lower allele frequencies (0.05-0.25) compared to the uniform distribution at higher MAFs (Figure 3.1 F). The most common

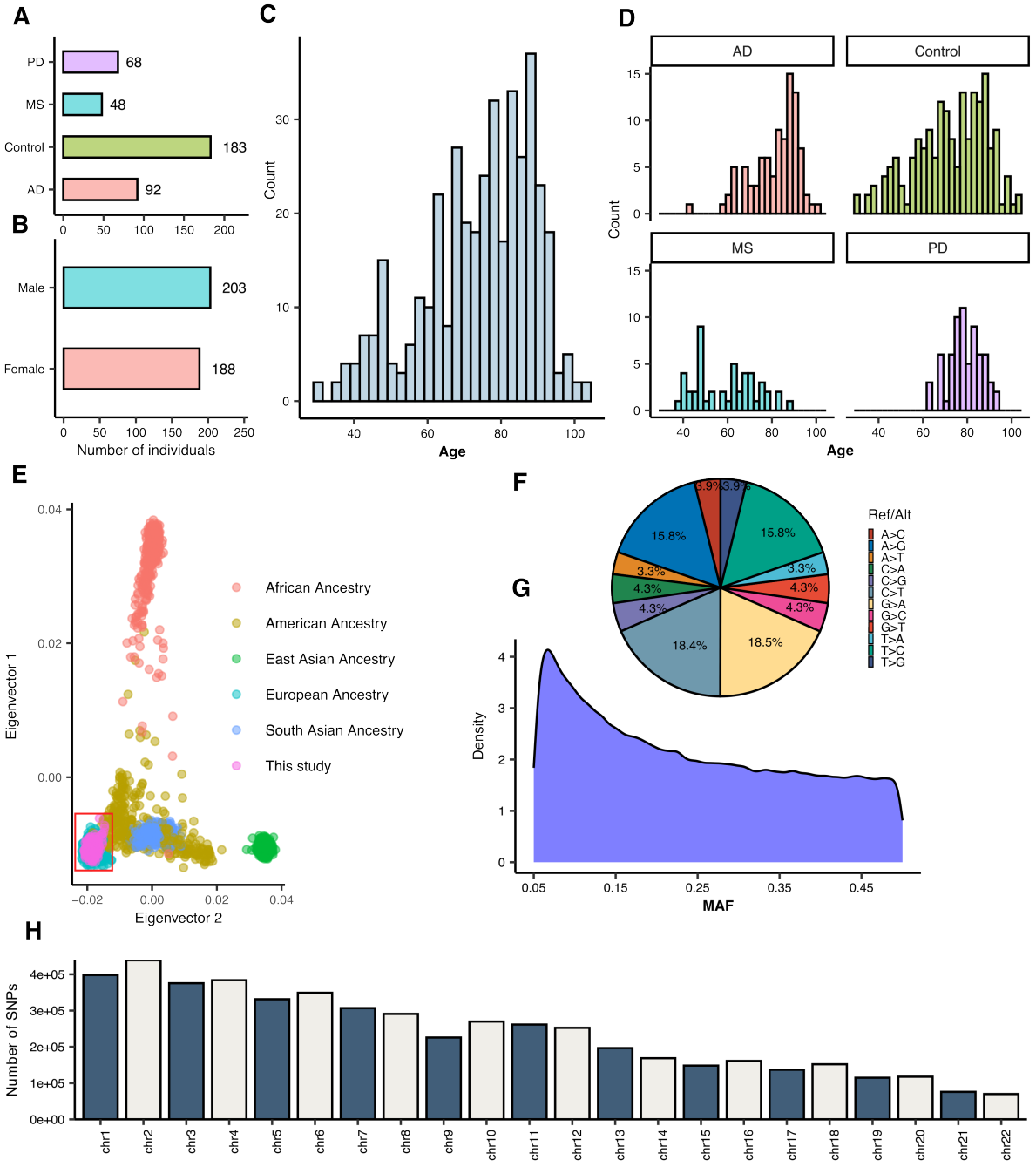


Figure 3.1: Cohort overview. ***A***, Distribution of individual diagnoses in the final QC'd cohort of 391 individuals. The x-axis represents the number of individuals in each category (PD, MS, Control or AD). ***B***, Male-to-female ratio in the cohort. ***C***, Age distribution of cohort. ***D***, Age distribution of cohort, categorised by diagnosis groups. ***E***, Genotype PCA plot of the first two eigenvectors, using genotype from this dataset ("This study") versus genotype data from the 1,000 genomes consortium across the 5 superpopulations. ***F***, distribution of reference versus alternate allele substitutions in final set of SNPs used in the analysis. ***G***, Minor allele frequency (MAF) distribution for final set of SNPs in a density plot (x-axis represents MAF, y-axis represents the smoothed estimate of the probability density function). ***H***, Number of SNPs assessed per chromosome.

reference/alternate allele combination was G-to-A (18.5%), followed by C-to-T (18.4%) and T-to-C / A-to-G (15.8%), whereas the two rarest were T>A and A>T (both 3.3%).

3.3.2 Dataset characteristics

Following snRNA-seq quality control, 2,348,438 single nuclei were retained over 8 major brain cell-types (excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, microglia, oligodendrocyte precursor cells (OPCs), endothelial cells and pericytes), based on clustering (Figure 3.2 A) and canonical marker gene expression (Figure 3.2 B). Of these, there were 795,655 Oligodendrocytes, 728,385 excitatory neurons, 280,938 astrocytes, 220,391 inhibitory neurons, 141,996 microglia, 109,436 OPCs, 47,585 endothelial cells and 24,052 pericytes based on these annotations. Overall, there was a mean of 6,006 nuclei per individiual (median 4,391) (Figure 3.2 C).

Cell-type proportions were mostly similar when comparing across disease cases (Figure 3.2 D); astrocytes, for instance ranged between 11.0 - 13.1% across all diagnoses, and OPCs between 4.0 - 5.6%. Interestingly, PD had a much smaller average proportion of excitatory neurons, with an average 18.4% excitatory neurons compared to 30.1 - 34.2% in all other samples, consistent with the well-known loss in dopaminergic neurons in PD[211]. This in turn resulted in a higher than average oligodendrocyte proportion (45.8% compared to 31.4 - 33.4 %). Microglia on the other hand had a wider range ranging from 4.4 -8.6% with the lowest proportion observed in the AD cases. The rarer cell-types, endothelial cells (1.0-2.5%) and pericytes (0.8 - 1.2 %) were generally much more difficult to capture, and highlights the potential for deeper sequencing at cell-type level to capture these cell-types.

I then pseudobulked each cell-type requiring a minimum of 10 cells per individual in that cell-type, retaining between 323 - 391 individuals in each pseudobulk expression matrix (Table 3.1). Although annotation was performed separately for each dataset (BRYOIS_192, MATTHEWS, ROCHE_92 and MRC_60), marker gene expression was consistent with cell-type when looking at the combined pseudobulk matrices (Figure 3.2 D).

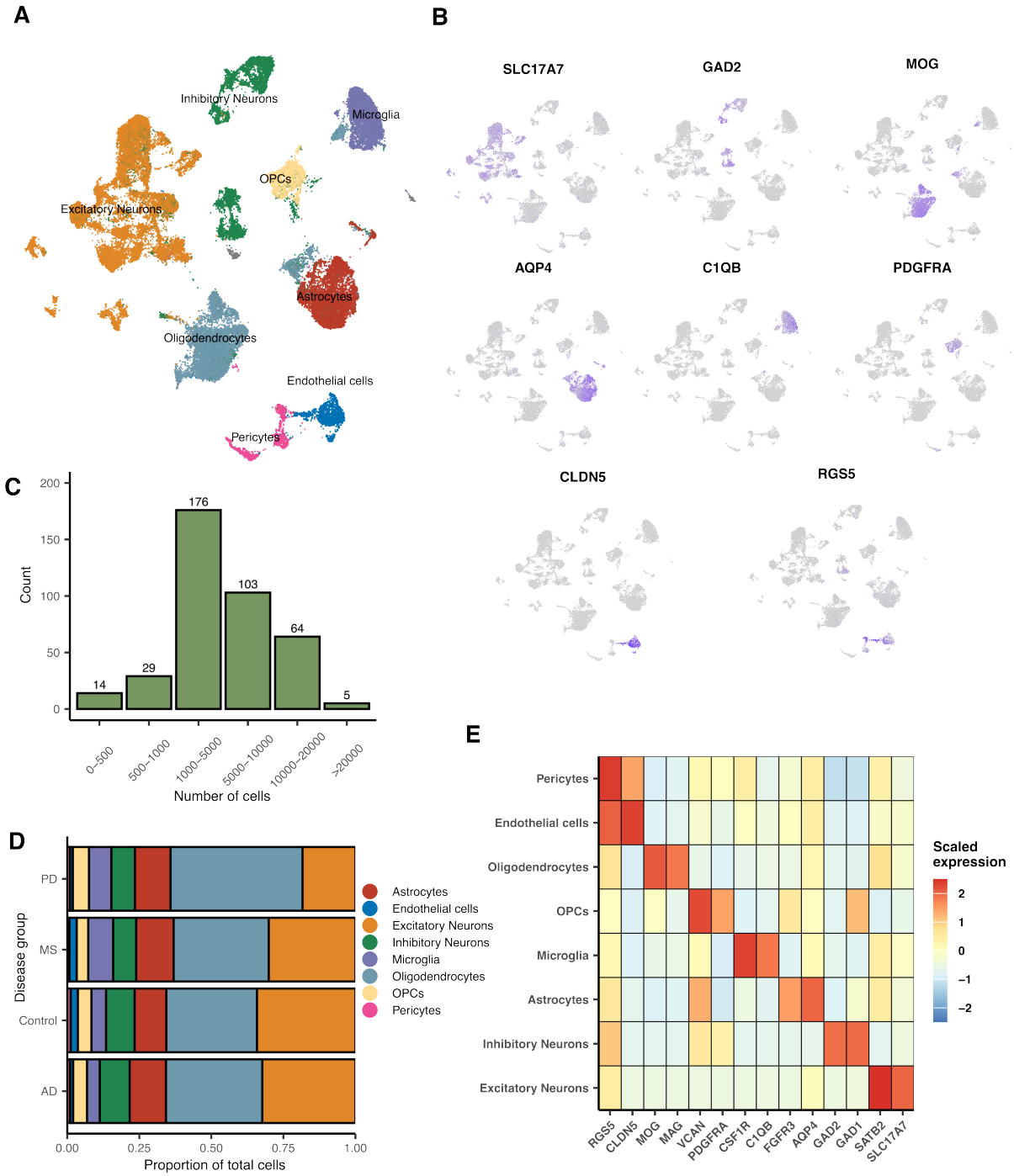


Figure 3.2: Single-cell analysis. ***A***, UMAP of single-cell clusters from subset of the the *ROCHE_92* single-cell dataset. ***B***, Marker gene expression in each cell-type cluster; *SLC17A7* for excitatory neurons, *GAD2* for inhibitory neurons, *MOG* for oligodendrocytes, *AQP4* for astrocytes, *C1QB* for microglia, *PDGFRA* for oligodendrocyte precursor cells (OPCs), *CLDN5* for endothelial cells and *RGS5* for pericytes. ***C***, Histogram of cell number ranges across dataset. ***D***, Cell-type proportions across dataset, grouped by diagnosis. The x-axis represents the total proportion of cells (0 - 1 signifying 0 - 100%). ***E*** Marker gene expression following pseudobulking of single-cell data.

Table 3.1: Number of individuals retained after pseudobulking (min. 10 cells)

Cell-type	Number of individuals retained
Excitatory Neurons	386
Inhibitory Neurons	377
Astrocytes	389
Oligodendrocytes	391
Oligodendrocyte Precursor Cells (OPCs)	385
Microglia	382
Endothelial cells	330
Pericytes	323

3.3.3 *cis*-eQTL discovery overview

To perform *cis*-eQTL discovery, I developed a computational pipeline to automatically handle single-cell data with genotype information to create the necessary input files for *MatrixEQTL*. Briefly, this pipeline generates pseudobulk matrices from single-cell data as well as a genotype dosage matrix (where rows represent SNPs, columns represent individuals and containing either 0, 1 or 2 to indicate dosage of the reference allele) from *vcf* data. As described in the methods section, this pipeline includes a principal component (PC) optimisation step, which iteratively performs MatrixEQTL runs each time with a set of PCs from the residual expression of each cell-type as covariates, where the number of PCs yielding the highest number of discovered eGenes (at $< 5\%$ false discovery rate, FDR) are retained for the final eQTL discovery.

In total, I tested between 16,383 - 18,030 genes per cell-type against 5.22 million SNPs in a *cis*-window of 1Mb from each end of the gene. Following FDR correction, 13,939 unique genes (eGenes) were regulated by one or more SNPs, in one or more cell-types (FDR $< 5\%$) (Figure 3.3 A) (Appendix 1). The number of eGenes discovered per cell-type had a significant correlation (Pearson r=0.92, p-value $= 1.14 \times 10^{-3}$) with the number of cells annotated (Figure 3.3 B); I found 8,706 eGenes in excitatory neurons (of which 2,228 unique to excitatory neurons), 6,355 (1,239 unique) in oligodendrocytes, 4,437 (660 unique) in

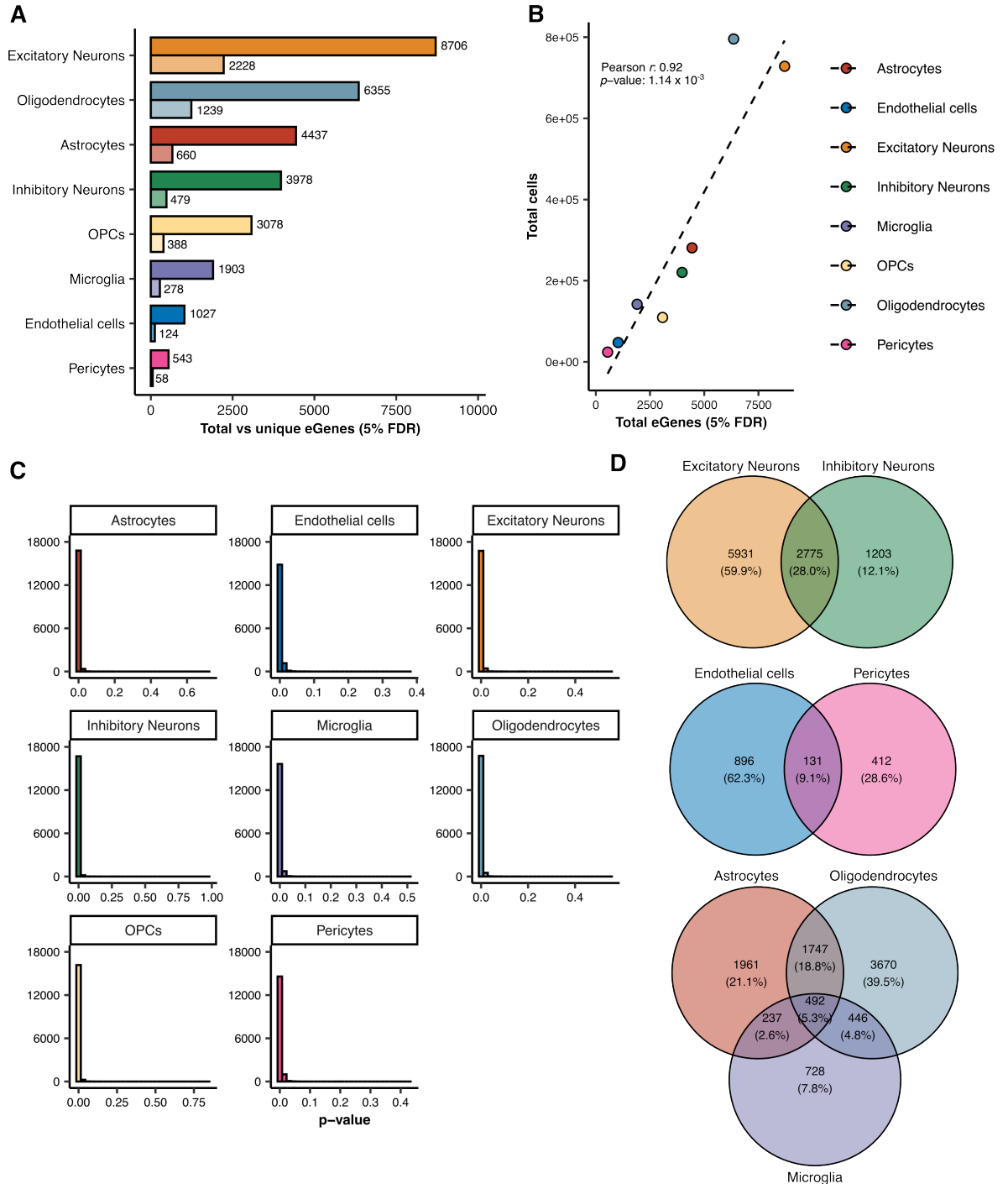


Figure 3.3: *cis*-eQTL discovery. ***A***, Number of significant eGenes discovered per cell-type (< 5% FDR). Each set of bars for each cell-type (y-axis) represents the total number of eGenes discovered (darker, top bar) versus the number of eGenes discovered unique to that cell-type (lighter, bottom bar). For example, 8,706 significant eGenes were discovered in excitatory neurons, of which 2,228 were unique to excitatory neurons and not present in any other cell-type. ***B***, Number of cells assayed per cell-type (y-axis) versus number of eGenes discovered (x-axis). ***C***, *p*-value distributions, using the associations from the top SNP for each gene in each cell-type. The x-axis represents the *p*-value, and the y-axis represents the frequency (bin size = 30). ***D***, Proportion of shared eGenes between select cell-type groups. For example, 2,775 eGenes were shared between excitatory neurons and inhibitory neurons, representing 28.0% of the total number of eGenes between the two (9,909).

astrocytes, 3,978 (479 unique) in inhibitory neurons, 3,078 (388 unique) in OPCs, 1,903 (278 unique) in microglia, 1,027 (124 unique) in endothelial cells and 543 (58 unique) in pericytes, for a total of 30,027 eGenes in aggregate. Of the 13,939 unique eGenes across all cell-types, a 5,454 (39.1%) were only found in a single cell-type, demonstrating the high level of cell-type specificity. In addition, the association p-values were well distributed above the expected uniform null distribution, showing an enrichment of associations (Figure 3.3 C). A high proportion of eGenes were shared between cell-types, most of all between the two neuronal classes (28.0% shared eGenes) (Figure 3.3 D) when compared to endothelial cells and pericytes (9.1%) or glial cell types (18.8% between astrocytes and oligodendrocytes) (Figure 3.3 D).

A substantial proportion of eGenes were unique to a single cell-type (39.1%). For instance, the *PXDNL* gene was only found to have significant associations within inhibitory neurons, exemplified by the association of *rs28619907* and *PXDNL* (p-value $= 1.50 \times 10^{-11}$, FDR $= 2.97 \times 10^{-8}$) not present in any of the other cell-types (Figure 3.4 A). While it is possible for an association to exist because of lack of expression in other cell-types, this was not the case for *PDXNL* which had moderate to high levels of expression in all cell-types, although specifically much higher in inhibitory neurons (Figure 3.4 B). In fact, *PDXNL* expression was significantly associated to 416 SNPs, none of which were associated in any other cell-types (Figure 3.4 C).

In a different example, *rs1693551* was significantly associated with *SNX31* expression in astrocytes (p-value $= 1.08 \times 10^{-46}$, FDR $= 2.97 \times 10^{-8}$), but not in any other cell-type (Figure 3.4 D). In this case, *SNX31* expression was higher in excitatory neurons and inhibitory neurons, while more variable in astrocytes and OPCs (Figure 3.4 E). Despite this, the majority of eQTLs for *SNX31* (n = 119) were found in astrocytes (Figure 3.4 F), and although the particular association between *rs1693551* and *SNX31* was not present in excitatory neurons there were 24 other eQTLs present. However, these shared eQTLs varied drastically in terms of association strength and effect size between excitatory neurons and astrocytes; while astrocytes had 50 of the 119 associations with a significance below 5×10^{-8} (as an arbitrary threshold), all 24 eQTLs for *SNX31* in excitatory neurons had association p-values which ranged between 4.79×10^{-4} - 1.39×10^{-4} (which equated to an FDR between 0.019 - 0.05). These associations are naturally linked to effect size, where the *rs1693551* (the top eQTL for astrocytes) had an effect size of 0.85, compared to 0.16 for the top eQTL in excitatory neurons.

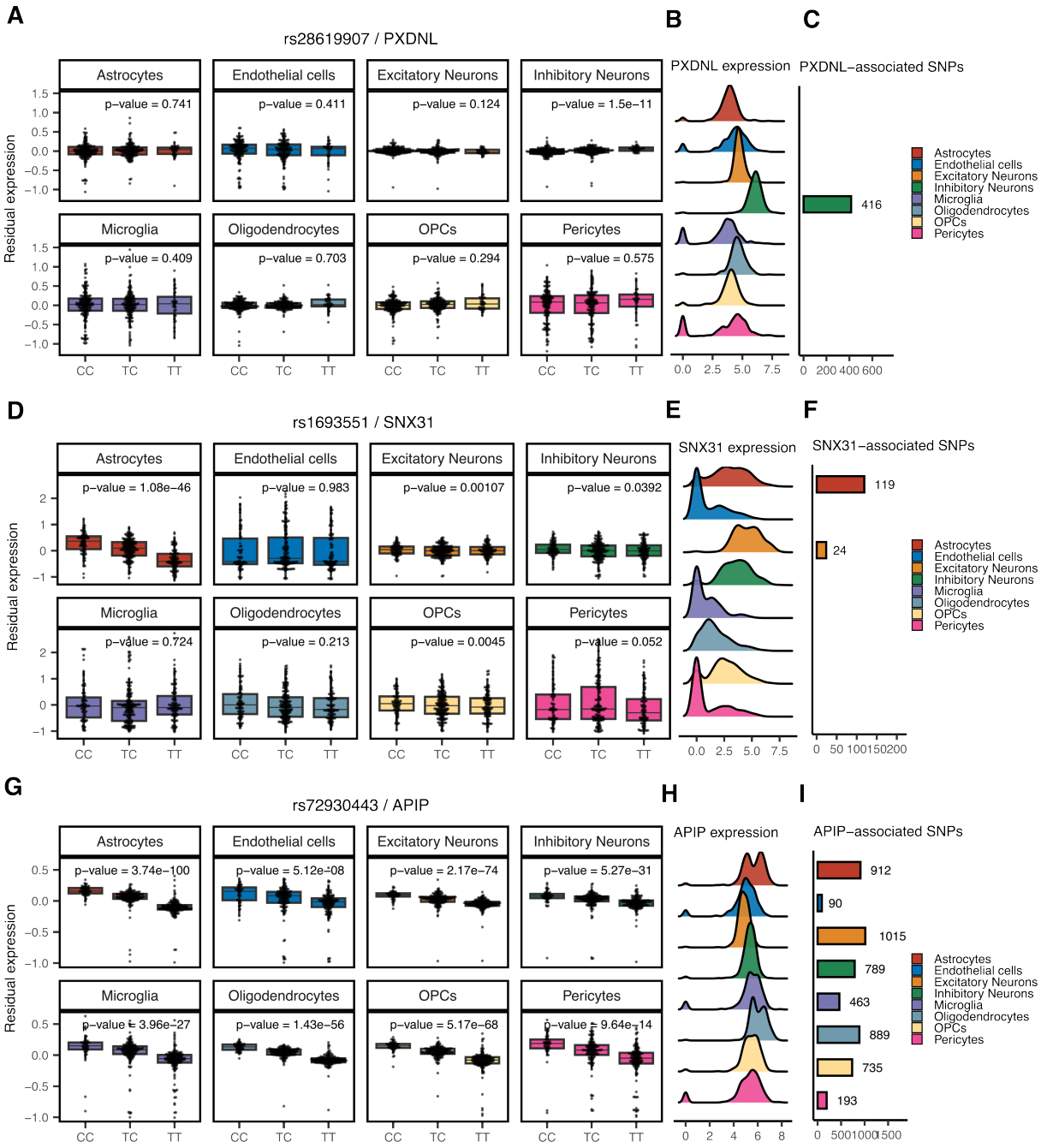


Figure 3.4: Examples of cell-type specific eQTLs. ***A***, Association between *rs28619907* and *PXDNL* expression in inhibitory neurons, compared to 7 other brain cell-types. The y-axis represents residual expression after covariate correction, and are the values used in the MatrixEQTL model. The x-axis represents genotype dosage. ***B***, Overview of normalised *PDXNL* gene expression values in bulk expression matrices across cell-types. ***C***, Number of SNPs (eQTLs) associated with *PDXNL* expression across cell-types. ***D***, Association between *rs1693551* and *SNX31* expression in astrocytes. ***E***, *SNX31* expression across all cell-types. ***F***, Number of SNPs associated with *SNX31* expression across cell-types. ***G***, Association between *rs72930443* and *APIP* expression, significant in all cell-types. ***H***, *APIP* expression across all cell-types. ***I***, Number of SNPs associated with *APIP* expression across cell-types.

Interestingly, only 6 of the 13,939 eGenes discovered were shared across all 8 cell-types, of which one was *APIP*. In this example, *rs72930443* was significantly associated to *APIP* expression in all cell-types, highest in astrocytes (p-value = 3.73×10^{-100} , FDR = 7.93×10^{-94}) (Figure 3.4 G). Overall expression was highest in oligodendrocytes, but generally highly expressed in all cell-types (Figure 3.4 D). There were over 1,000 eQTLs for *APIP* in excitatory neurons, and more than 700 in OPCs, inhibitory neurons, oligodendrocytes and astrocytes, likely due to a high number of correlated SNPs (Figure 3.4 I). In contrast, *APIP* did not have many eQTLs in endothelial cells or pericytes, potentially due to differences in statistical power, as the number of eQTLs for *APIP* across the 8 cell-types generally followed the number of cells assayed per cell-type.

3.3.4 Patterns of eQTL behaviour

In total, the 30,027 eGenes discovered in aggregate in the 8 cell-types represented a total of 1.82 million SNP-gene associations (Figure 3.5 A), with a total of 857,538 unique SNPs across all cell-types. Each eGene was associated to at least one SNP, with an average of 32 - 74 SNPs per gene (varying by cell-type, 60 overall), largely driven by regions with extremely large number of SNPs likely in LD (median number of SNPs per gene in each cell-type: 2 - 15). For instance, there were 48 genes with more than 1,000 SNPs strongly associated to their expression, of which 8 belonged to the *MAPT* locus (chr17:45,309,498-46,903,773) and 22 to the *HLA* locus (chr6:28,510,120-33,480,577), many of which had associations across multiple cell-types. After performing LD pruning SNPs associated to each gene, the median number of independent SNPs across all cell-types was 1, with few regions containing more than one independent SNP.

Although there were shared eGenes between cell-types, they were not necessarily regulated by the same SNPs. For instance, the largest proportion of shared eGenes were shown between the two neuronal cell classes (28%); however, only 1,555 of those (56%) were regulated by the same SNPs. I then asked the question whether the SNPs regulating the same genes across cell-types were consistent in their directionality. Investigation of the effect sizes between these SNP/gene pairs shared between the two cell-types showed that 99.76% had the same direction of effect, and a pearson correlation coefficient of $r = 0.95$ (Figure 3.5 B). This was seen in all cell-type pairs, with correlation coefficients ranging between $r = 0.84$ - 0.98, with 95.1 - 99.9%

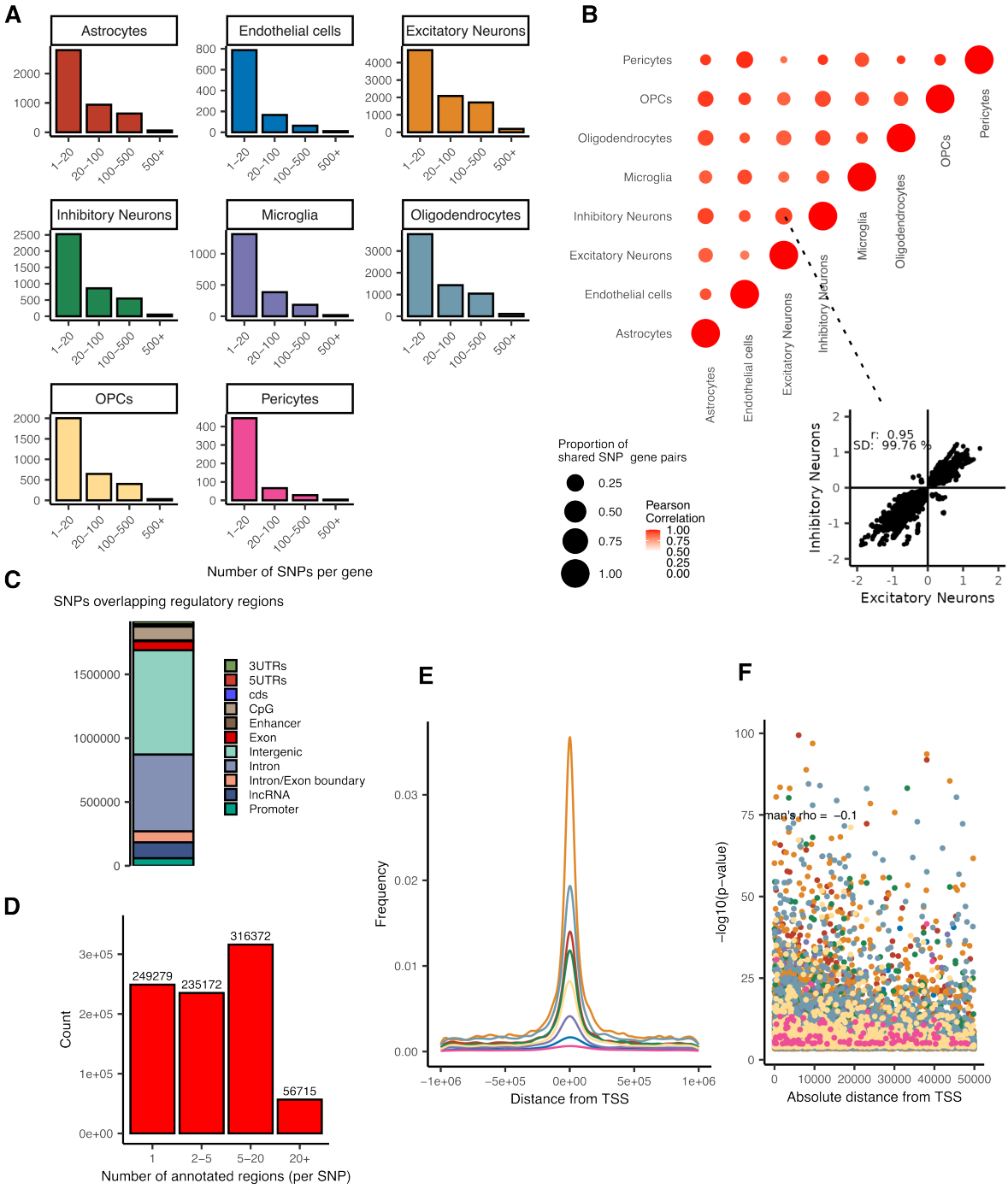


Figure 3.5: Patterns of eQTL behaviour. ***A***, Total number of eQTLs per eGene across all cell-types, binned into 1-20, 20-100, 100-500 and 500+ eQTLs per gene. ***B***, Correlation of effect sizes for shared SNPs between cell-types. The size of the dot represents the proportion of shared eQTLs between two cell-types, and the shade represents the degree of Pearson correlation between the effect sizes for shared eQTLs. As an example, a scatterplot of effect sizes for shared eQTLs between inhibitory neurons and excitatory neurons is shown (SD = same direction). ***C***, Overlap of all unique eQTLs with regulatory regions. The y-axis represents the number of SNPs overlapping a specific type of regulatory region. ***D***, Number of annotated regions per SNP. ***E***, Distance from TSS for all significant eQTLs. ***F***, Scatterplot of -log10(p) (y-axis) versus absolute distance from TSS. Each point represents a SNP colored by the cell-type in which it was found to be significant, and Spearman's rank correlation coefficient is shown.

of SNP-gene pairs with the same direction of effect. Importantly these tests were only possible for shared SNP-gene pairs which only represented a small proportion of the total SNP-gene pairs between two cell-types; only 1.88% of SNP-gene pairs between excitatory neurons and pericytes were shared, for instance, compared to the two neuronal classes which had the highest proportion of overlapping SNP-gene pairs (22%).

To understand the potential function of these eQTLs and how they are likely regulating gene expression, I investigated whether these eQTLs overlapped any known regulatory regions using a combination of databases such as *FANTOM5*, *ENSEMBL* and *GENCODE* as collated in the *annotatr* package[197]. As expected, the large majority of eQTLs overlapped intergenic (42.7%) or intronic (31.4%) regions. There was also a large overlap with lncRNA regions (6.52%), but these SNPs were not necessarily associated to their expression; for example, *rs1016971* was significantly associated with *KNOP1* expression (p-value = 2.00×10^{-66}) in astrocytes (a protein-coding gene) but also overlaps with the *AC0022550* lncRNA. The next most common annotation was "CpG" which encompassess CpG islands, shores and shelves (5.63%), followed by exons (3.56%) and promoters (3.08%) (Figure 3.5 C).

Genomic architecture is complex and more often than not, SNPs overlapped more than one regulatory feature. Of the 857,538 unique SNPs regulating gene expression across the 8 cell-types, 249,279 (29.1%) only overlapped a single feature, while the majority overlapped more than one feature. Interestingly, the largest proportion of these SNPs overlapped as many as 5 - 20 features (316,372, 36.9%)(Figure 3.5 D). In one example, *rs13303057* overlaps an intergenic region, while also overlapping a CpG shore, an exon for the *LINC02593* lncRNA and being on the intron/exon boundary for a different transcript of that same lncRNA.

In a more broad approach less reliant on databases, I also sought to understand the proximity of these eQTLs to transcription start sites (TSS). The majority of eQTLs were enriched around the TSS, in all cell-types and as previously observed by Bryois et al.,[108] (Figure 3.5 E). Proximity to TSS however was only moderately correlated with eQTL p-value and not necessarily a measure of strong association (Spearman's rho = -0.45, p-value versus absolute distance) (Figure 3.5 F). In fact, when reducing the window to +/- 50,000 kb, there is almost no correlation between the two (Spearman's rho = -0.1).

3.3.5 Comparison to bulk data

To assess whether the eQTL discovery was enriched compared to eQTL discovery in bulk brain, I undertook a series of analysis. First, I assessed whether the eQTLs were comparable to those of a well-powered bulk RNA-seq eQTL study by de Klein and colleagues[90], in which eQTLs were mapped in a cohort of 6,523 individuals using prefrontal cortex tissue (the Metabrain dataset) (Figure 3.6 A). Restricting the analysis to SNP/gene pairs that were present in both datasets, I found that 73.0 - 88.7 % of significant eQTLs I discovered (FDR $<$ 0.05) were also significant within the Metabrain study, varying by cell-type. For instance, of the 602,735 significant associations I discovered in excitatory neurons, 480,860 were also FDR significant ($<$ 5%) in the Metabrain study. The highest replication percentage was seen in pericytes (88.7%), although the search space was considerably smaller (only 17,047 associations, of which 15,115 were also significant in Metabrain). In addition, by looking at directionality of effect and harmonizing effect alleles across the two datasets, I found that 90.0 - 98.3% (by cell-type) of these replicated eQTLs had the same direction of effect in both metabrain and the cell-type specific eQTLs presented here.

I also attempted to understand the level of power gained from using a single-cell approach compared to bulk. To this end, I generated a synthetic "bulk" expression profile, in which I summed counts across all cells in all cell-types to use in eQTL discovery. With this synthetic bulk dataset, I discovered a total of 10,166 significant eGenes (FDR $<$ 5%), of which a considerable amount (9,041, 88.9%) was shared with the cell-type specific eGenes (Figure 3.6 B). On the opposite side however, using cell-type specific eQTLs I discovered a substantially larger amount of unique egenes compared to the synthetic bulk (9,041 out of 13,939, 64.9%) meaning that 4,898 (35.1%) genes would not have been discovered in a bulk setting. Conversely, a small proportion of eGenes were only found to be significant when combining expression signals across all cells, and were not present in any of the cell-type specific eQTL results (1,125, 11.1%). When split by cell-type, the more common cell-types shared the most eGenes compared with bulk dataset compared to the less common cell-types, which is expected given that the combined expression signals will be dominated by the more abundant cell-types (Figure 3.6 C). In addition, comparing the eQTL discovery shown here with existing eQTL datasets shows a substantial power gain in eGene discovery despite the comparably modest sample size (Figure 3.6 D).

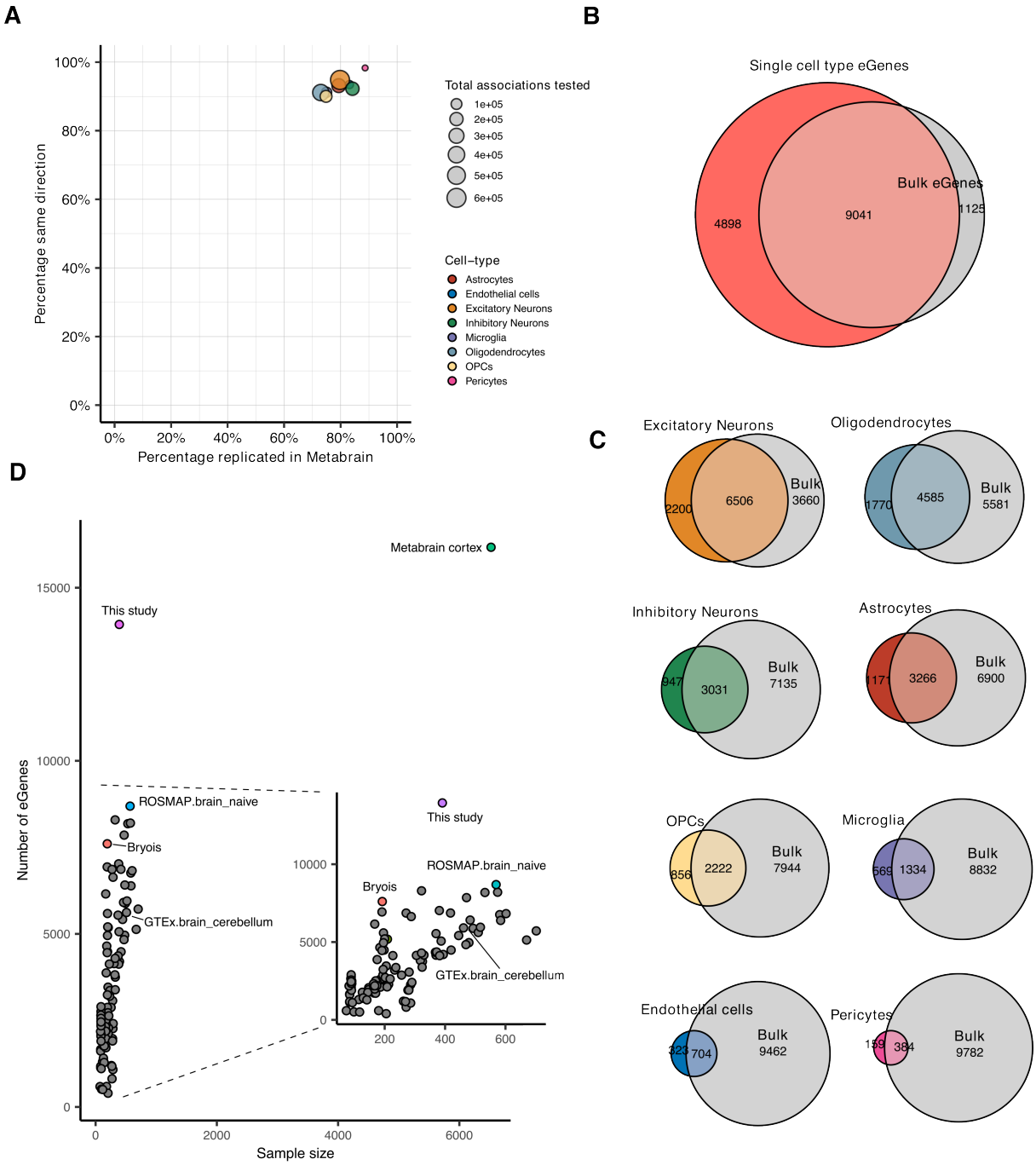


Figure 3.6: Single-cell eQTLs versus bulk eQTLs. ***A***, Summary of replicated eQTLs in the *Metabrain* cohort. Each circle represent a cell-type tested in my dataset, where the size represents the total associations tested for replication (that is, SNP / gene pairs assessed in both datasets) as indicated by the legend. The y-axis represents the percentage of eQTLs that had the same direction of effect as metabrain, after harmonizing for consistent allelic direction. The x-axis represents the percentage of associations tested that were replicated ($< 5\%$ FDR in both datasets). ***B***, Comparison of significant eGenes discovered in a "synthetic" bulk dataset (combining all singlecell expression into a single matrix) versus the combined eGenes discovered for each cell-type (red, left). ***C***, Comparison of synthetic bulk dataset eGenes versus the eGenes of each cell-type. ***D***, Total eGene discovery (y-axis) versus number of individuals (x-axis) for existing eQTL datasets including Bryois (red), metabrain (blue) and this study (pink).

3.4 Discussion

In this chapter, I describe the analyses of genome-wide *cis*-eQTL discovery using snRNA-seq and genotype data generated on post-mortem brains from 391 individuals, which as of January 2024 is the largest eQTL dataset yet conducted in the human brain. I discovered a total of 1.82 million significant SNP-gene associations, representing a total of 13,939 unique genes and 857,538 unique SNPs present in one or more cell-types. Of these eGenes, 5,454 (39.1%) were unique to a single cell-type, demonstrating a high level of cell-type specificity. I further investigate to what degree these eGenes are shared between cell-type, which showed that the two neuronal classes shared the highest overlap of eGenes as compared to other cell-types (28.0%) whilst also highlighting that a substantial proportion of these neuronal genes are specific to their subtype. In addition, I attempt to uncover patterns of regulation by the discovered SNPs by understanding the surrounding epigenetic landscape. Finally, I show the advantages of using a single-cell approach for eQTL discovery compared to bulk, while maintaining a high level of replication with a well-powered bulk dataset.

Compared to the study published by Bryois et al.,[108] , this cohort included almost 200 additional individuals, and is more than twice the size. As a result, I discovered a higher number of significant eGenes (13,939 versus 7,607, representing an 83.2% increase) in the same number of cell-types, and using a similar SNP set (filtered to >5% MAF). There are however several factors to consider in light of these results.

3.4.1 Dataset limitations

As seen in this dataset as well as others, an inherent property of post-mortem brain data is that there is a considerable skew towards older age, as seen by a median/mean age of 77/74 in this dataset. Brain morphology, for instance, is known to change considerably over the course of life, and differences between middle-aged and older adults is well documented[212]. Aging is commonly associated with cognitive decline even in the absence of PD or AD, not directly through neuronal loss but through a series of subtle synaptic alterations across multiple brain regions[213]. Molecular changes have also been observed at the cellular level; in a recent mouse study, Allen and colleagues showed through single-cell transcriptomics that there are multiple altered factors changed as a result of aging including spatial organization, gene expression,

cell-state and subtype proportions. Interestingly these changes were more pronounced for non-neuronal cell-types, highlighting the involvement of glial and immune cells in the aging brain[214].

It is therefore important to consider that it is possible for a level of variance in gene expression between individuals that is not explained by genotypic differences, but rather by age. While efforts have been made in order to capture these imbalances through the use of principal components and age as a covariate in the eQTL modelling, aging is a complex biological process that cannot be comprehensively explained by simple linear covariates. In addition, the brain samples were diagnosed through extensive neuropathological examination to accurately group samples into diagnosis categories. It is however not unlikely that a proportion of these individuals were in fact on a disease trajectory but at a time-point where no neuropathology is observable.

3.4.2 Single-cell analysis

One key consideration in regards to the analyses presented here is the resources needed for the single-cell analyses. Because of the sheer dimension of the dataset, it was initially computationally inefficient (or downright unfeasible) to analyse all of the samples together. Performing single-cell analyses on the entire dataset could help better harmonize cell-type annotations, for instance, where the smaller datasets may be more susceptible inter-individual variability.

An opportunity that is highlighted here for future work is to extend these analyses to specific cell subtypes, such as layer-specific excitatory neurons or glial subclusters. A recent paper by Siletti et al., which comprises the largest single-cell brain atlas to date characterised over 3,000 transcriptionally diverse cell subtype clusters[98]. In the same way bulk obscures cell-type specific signals due to mixing of expression patterns, the annotation of major cell-types risks obscuring subtle transcriptomic signals specific to cell subtypes. However, it is worth noting that with a median of 4,391 cells per individual these rarer cell-types may be difficult to capture. Currently, single-cell eQTL datasets have favoured individual-level statistical power over cell-level power, mostly because to achieve such large datasets all possible single-cell data available is used[107,108]. It is unclear where the balance between the two lies, and how you

measure success from an eQTL study (more eGenes in a few select cell-types, versus less eGenes but in a more diverse set of cell-types). It is likely future studies will report eQTLs in "layers" based on the resolution of the annotations.

3.4.3 eQTL discovery

When it comes to the eQTL discovery, a limitation of this work is the lack of analysis on *trans*-eQTLs to make this a truly genome-wide study. There are however several concerns when performing *trans*-eQTL analysis.

First and foremost, this dataset is likely to be statistically underpowered for accurate discovery of *trans*- effects as has been observed previously. These *trans*- effects of SNPs on distal genes are generally much weaker and harder to detect, whilst exponentially increasing the burden of multiple testing[4,215]. In the Metabrain study, as many as 6,523 individuals were included but only discovered 737 *trans*-eQTLs compared to 16,169 *cis*-eQTLs. Practically, mapping *trans*-eQTLs is also extremely computationally intensive due to the large number of SNP-gene pairs being tested. However, well-powered *trans*-eQTL analysis has the potential to uncover the architecture of genetic networks, where SNPs regulate the expression of multiple genes.

Another key assumption that is made here is that aggregation of counts is the optimal strategy for discovering eQTLs at the cell-type level. An initial systematic study of eQTL discovery from single-cell data by Cuomo et al., in 2021 showed that aggregation is often the best approach, with the highest level of discovered eGenes and replication in bulk datasets[216]. In fact, the majority of eQTL studies to date using single-cell RNA have used aggregation approaches, including Bryois et al.,[108] and Yazar et al.,[107]. These two metrics also highlight two important caveats; first, more eGenes does not necessarily mean a better approach as it may also include a higher degree of false positives. Secondly, replication in bulk-derived eQTLs is common, but it is important to remember that these are inherently biased because of (1) bulk tissue contains multiple cell-types which may obscure biological signal and (2) these are not gold standard datasets free from other unobserved biases.

On the other hand, count aggregation has several main advantages; for instance, it can be interpreted as a rough variance-stabilization method prior to normalization to deal with the

high drop-out rates that plague single-cell data. Aggregation also allows the application of well-established methodologies for eQTL discovery, many of which have been optimised over several years[103] and have track records in multiple tissue types. Modelling eQTLs at the single-cell level has been demonstrated with Poisson Mixed Effects (PME) models but remain computationally prohibitive[109]. However, aggregation of counts may undoubtedly result in the removal of other biological signals and may suffer from the same biases seen in bulk (albeit in a smaller scale).

Finally, use of PCs for variance stabilisation has been a common approach in the large majority of eQTL studies, both at bulk and cell-type levels[74,90,107]. Generally speaking and as has been done here, PCs are selected in a way that maximizes eGene discovery. However, more is not always better and there is always the risk of overfitting by capturing variance that could in fact be biological signal.

3.4.4 Epigenetic annotation

Epigenetic annotations are curated from a large variety of datasets, many of which are the results of experiments carried out in bulk tissue. The *FANTOM5* consortium for example has used thousands of samples for epigenetic characterisation, but mainly conducted in mouse, primary cells and cell lines[217], and may poorly represent the epigenetic regulation of the different cell-types. The *ENCODE* consortium was also almost exclusively cell-line based; while covering an impressive variety of cell and tissue types, these may not be reflective of the cell-types in the post-mortem brain[218].

There have however been several studies in recent years attempting to provide more context-specific epigenetic data. In 2019, Nott and colleagues conducted one of the first studies to characterize the regulatory elements of cell-types in the post-mortem brain, by combining nuclei isolation from cortical brain tissue with fluorescence-activate nuclei sorting (FANS) and then ATAC-, Histone ChIP- and PLAC-seq[209]. Focusing on four majour cell-type populations (oligodendrocytes, astrocytes, microglia and neurons) this approach identified plausible cell-type specific mechanisms around disease risk, exemplified by the evidence around *rs6733839* (the lead SNP in an important AD risk locus) and the *BIN1* gene which not only showed microglia-specific enhancer peaks (through H3K27Ac ChIP-seq) but also a

potential physical chromatin connection between *rs6733839* and *BIN1* through PLAC-seq. In a similar approach, Roussos et al., in 2020 isolated nuclei from three distinct brain regions (anterior cingulate cortex, dorsolateral prefrontal cortex and primary visual cortex) and performed ATAC-seq on four distinct cell-type populations, sorted by FANS (GABAergic neurons, glutamatergic neurons, oligodendrocytes, and microglia/astrocytes). This dataset (also known as *BOCA2*) showed an enrichment of neuron-specific enhancers around schizophrenia risk variants[219].

As these studies become more common, it will be possible to add truly cell-type specific functional significance to identified genetic variants, especially if combined with cell-type specific eQTLs. Not only will the function of the genetic variants become less speculative by drawing from context-specific data, but may also validate the observed eQTL associations in a "denoising" approach. Taking a step further, recent developments have allowed the simultaneous assessment of the open-chromatin landscape and gene expression from the same cell/nuclei[191]. While costs remain high, future large-scale studies will be able to simultaneously map expression and open chromatin QTLs, providing high confidence in the function of these SNPs while allowing for a more strict filtering of associations.

3.4.5 Replication value

Single-cell eQTL approaches are still very recent, and as such it is difficult to say whether the discovered associations are credible. Use of well-powered datasets as a validation of discovered associations (such as Metabrain[90]) is a good sense-check to make sure they are not spurious. However, as is a weakness of the epigenetic annotation, so too is the use of bulk tissue a weakness of replication eQTL datasets due to obscuring expression signals of specific cell-types.

An important question which future studies will consider is whether to prioritize cell numbers per individual versus total number of individuals assayed; cell-level or individual-level statistical power.

Chapter 4

Impact of disease on cell-type specific eQTL associations

4.1 Preface

4.1.1 Appendices

The appendices referenced in this chapter are Appendices 2-4, organised as follows;

- Appendix 2: disease-interaction QTLs
- Appendix 3: eQTLs from the controls-only cohort
- Appendix 4: age-interaction QTLs (controls-only)

4.2 Introduction

To date, eQTL studies in the brain using bulk tissue have seen a steady increase in cohort sizes but because control-tissue is often difficult to obtain many have included disease samples to maximize statistical power. The PsychENCODE consortium for instance organised a resource with expression and genotype data from 1,866 individuals, of which 827 (44%) were diagnosed with a psychiatric disorder[193]. The CommonMind Consortium (CMC) and the Accelerating Medicines Partnership for Alzheimer's Disease (AMP-AD) Consortium were also

used in a recent meta-analysis by Sieberts et al., bringing the study cohort to 1,433 samples with a substantial proportion of individuals with psychiatric as well as AD diagnoses[220]. These combined with the GTEx and Braineac datasets eventually made up Metabrain, which is the largest eQTL study in the brain to date with samples from 6,523 individuals[90]. Few studies have exclusively used control tissue and while it could be argued that control tissue may suffer from unobserved pathology (especially in aged samples, which are common for post-mortem brain), the effect on gene expression is unlikely to be as evident as from samples with observable pathology. In 2021, Porcu et al., systematically investigated the effects of disease pathology on gene expression, finding that the majority of differentially expressed genes (when comparing controls versus disease samples) were the result of disease-induced rather than disease-causing changes in the transcriptome[172].

As of yet it is unclear how the presence of disease samples affect the resulting gene expression, and their impact on SNP-gene relationships in eQTL studies. To circumvent the effect of disease, the standard methodology to date has been to adjust statistical models or even the expression data itself by including diagnosis as a fixed effect, as seen in CMC[220]. Alternatively, variance capturing methods (such as PCA or PEER[125]) are used to derive factors that correlate with technical, demographic, and clinical covariates (including diagnosis), which are then included as covariates in the model, as demonstrated in Metabrain and PsychENCODE[90,193]. Fixed effects models however assume that the impact of diagnosis is consistent across all samples, thereby failing to account for differential effects on gene expression across multiple diagnoses (or other groupings for that matter). This in turn may erroneously identify eQTL associations that may in fact be disproportionately driven by the presence of disease, as well as obscure associations due to opposing effects on gene expression in the case of multiple diagnosis groupings.

This becomes even more important when considering the fact that expression is dynamic and context-specific, such as what is seen in different tissue types as well as cell-types transitioning through different states or developmental trajectories. It is therefore not inconceivable that eQTLs (or any other molecular QTLs for that matter) are also dynamic and context-dependent, where the association between a genetic variant and the expression of a gene is dependent on the context of the expression read-out. Only in the past few years have dynamic eQTLs been an increasingly popular area of research; in 2019, Strober et al., published one of the first dynamic eQTL studies where they mapped the dynamic regulation

of genes during cardiomyocyte differentiation, by using iPSCs from 19 individuals and measuring expression across 16 time points during the differentiation process[221]. Since then, there have been numerous publications on the discovery of dynamic eQTLs including in stimulated versus unstimulated T-cells[110] and a bulk study of psoriatic skin[222]. In addition, Nathan et al., and Yazar et al., (both in 2022) made use of the single-cell dimension to discover pseudo-time ordered eQTLs, showcasing the added advantage of single-cell approaches to derive both context- and state-specific eQTLs[107,109].

Identifying these dynamic effects is important for downstream interpretation. From a cell-type perspective, dynamic eQTLs can highlight genes that drive population shifts over time, be it in the transition of one cell-type to another such as during development, or the trajectory of healthy cell populations towards disease-specific subpopulations such as in the case of disease-associated microglia[223]. In addition, overlapping dynamic eQTLs with GWAS risk variants may offer insights into disease mechanisms.

4.2.1 Aims and objectives:

To date, there has not yet been a study that investigates the impact of disease on eQTL associations in a cohort comprising of both control and disease cases in the brain. In particular, the effects of disease on cell-type specific gene expression and its impact on eQTL discovery remains unknown. To this end, this chapter will aim to perform the following:

- Assess differences in eQTL discovery between cohorts using both disease and control tissue, and a cohort exclusively made up of control subjects.
- Model the interaction between disease status and genotype, and assess the extent to which the eQTLs are impacted by presence of disease.
- Investigate the impact of age on eQTL associations.

4.3 Results

4.3.1 Overview of methodology and random effect modelling

Application of standard linear models is quick and efficient and allows packages such as MatrixEQTL to quickly perform computations at scale[103]. However, these linear models assume that all included covariates are fixed and as such have the same impact across the dataset. In contrast, the addition of random effects using linear mixed-effects models allow modeling of variability across different groups. In addition, linear mixed-effects models allow the addition of interaction terms which allow explicit testing of whether the effect of one variable on the outcome depends on the level of another variable; in this particular case, an interaction term can discern whether the effect of genotype on gene expression varies by disease satus (Figure 4.1 A).

To test whether the inclusion of random effects in eQTL modelling had a significant impact on the overall association between genotype and gene expression, I performed step-wise eQTL modelling using linear mixed effect (LME) models, testing each association between the lead SNP of each discovered cell-type specific eGene (Chapter 3). In brief and as described in Chapter 2 (Methods), this involved first testing the null model (M_0) in which each the gene is tested versus all covariates; the base model (M_1) in which the genotype is included to mimic the discovery MatrixEQTL model; and finally the interaction model (M_2), testing the interaction between genotype and disease status (Figure 4.1 B). A likelihood ratio test (LRT) between a model and its more complex counterpart (M_0 versus M_1, and M_1 versus M_2) then determines whether a specific configuration is better suited to model the association. Importantly, all of these model configurations included a nested random effect to account for diagnosis-specific variation within each sample source. Including these effects is crucial as they address biases from the brain bank itself and the specific diagnoses, which may be influenced by the pathologists' assessments.

In total, 30,027 associations were tested using this step-wise modelling across the 8 cell-types. If M_1 was a better fit than M_0 (determined by a LRT p-value <0.05), this was considered a pass and a validation of the MatrixEQTL results, signifying that addition of the genotype shows an association to the SNP rather than any of the associated covariates. Across these

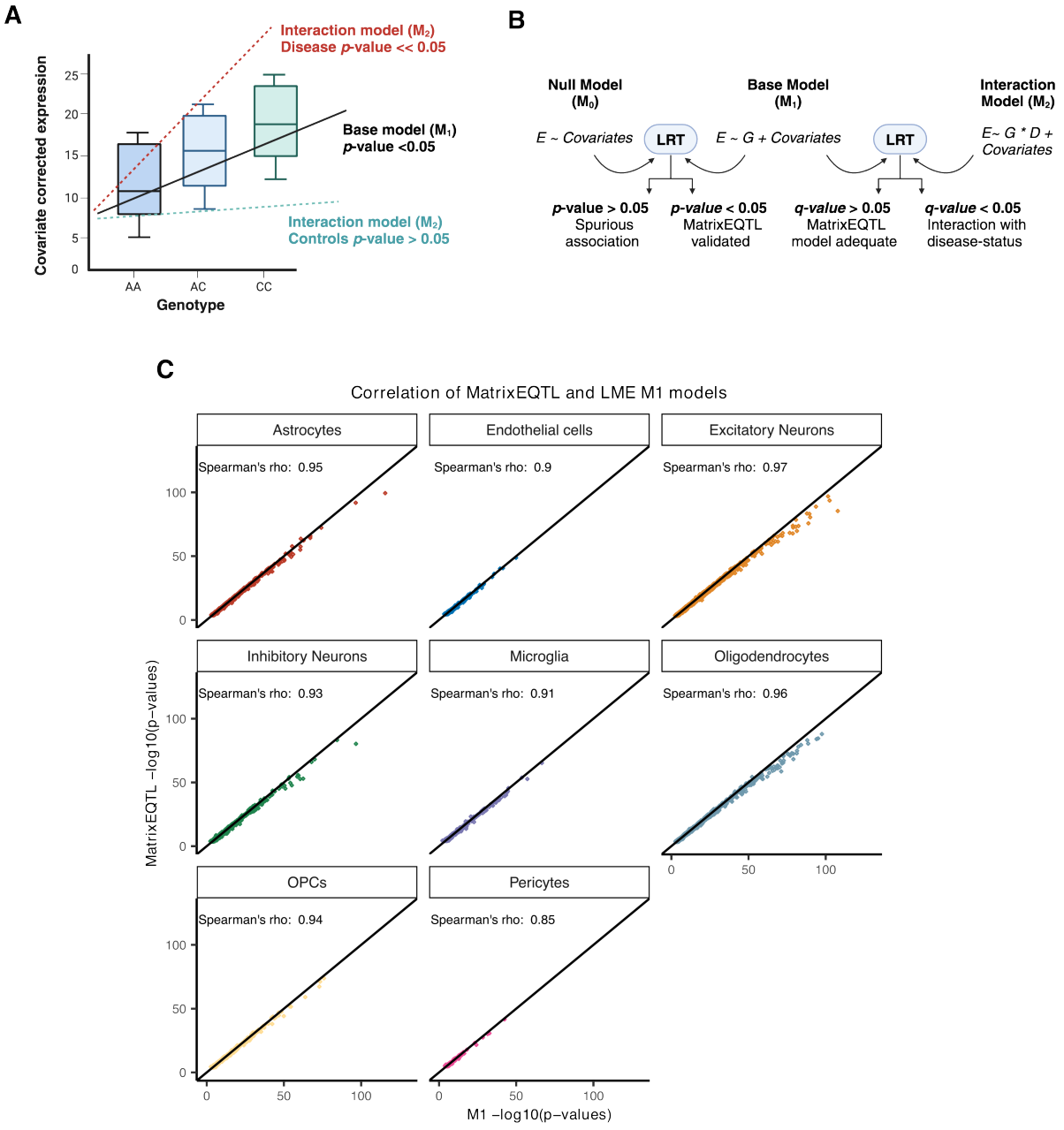


Figure 4.1: Overview of eQTL modelling methodology. ***A***, Rationale behind interaction testing. In a combined dataset, it is possible that an overall significant association ($p < 0.05$, black line) is mostly driven by the presence of disease samples as found in an interaction model ($p << 0.05$, red dotted line line) and would otherwise be non-significant in control-only tissue ($p > 0.05$, blue dotted line). ***B***, Overview of step-wise interaction testing. The null model M_0 tests whether there is an association between the expression of a gene and any associated covariates used in the dataset, whereas the base model (M_1) includes the genotype to model an eQTL association, and the interaction model (M_2) tests whether disease status interacts with genotype in predicting the overall association. A likelihood-ratio test (LRT) between each determines which model is more appropriate (M_0 versus M_1, and M_1 versus M_2). ***C***, Comparison of $\log_{10}$-transformed p-values as determined by MatrixEQTL (y-axis) versus M_1 (x-axis) for each cell-type and their respective Spearman correlation statistics.

30,027 associations, all but one (*rs10762316-HK1* expression in Inhibitory Neurons) were better fit with the M_1 model (with a SNP). In addition, despite the inclusion of a nested random effect the *p*-values from the M_1 models correlated well with the discovery MatrixEQTL *p*-values (Spearman's $rho = 0.85$ - 0.97) (Figure 4.1 C). These correlations were higher in the larger cell-types (oligodendrocytes spearman's $rho = 0.96$, excitatory neurons spearman's $rho = 0.97$) compared to the smaller cell-types (pericytes spearman's $rho = 0.85$, endothelial cells spearman's $rho = 0.90$), suggesting a greater degree of variability in the associations as statistical power decreases. However, these results overall suggest that the inclusion of nested random effects in eQTL models (at least in these examples) only has a small impact on the overall SNP-gene associations.

4.3.2 Modelling of disease interactions

Random effects account for random variability across groups, whereas interaction terms tests the combined effect of variables on the outcome. In particular, if an interaction term is included between a variable with different groupings, the model can extract the different effect each group has on the outcome. In the M_2 model and as described above, each association was tested with an interaction term between genotype and disease status. In combination with a LRT test between this model and the M_1 model, these tests tell us two things; first, whether an interaction model is more appropriate to model the association and second, whether the association is mostly driven by the presence of samples with disease diagnosis (Figure 4.1 A, B).

Applying this framework to all associations described above (except *rs10762316-HK1* in Inhibitory Neurons), there were 16.7 - 40.8% of tested eQTL associations (depending on cell-types) that were better modelled using an interaction model (LRT *q*-value < 0.05) (Figure 4.2 A) (Appendix 2). This value was highest in pericytes (221 out of 542, 40.8%) followed by microglia (697 out of 1,903, 36.7%) and lowest in excitatory neurons (1,455 out of 8,705, 16.7%).

To assess the distribution of *p*-values obtained from the LRTs in this stepwise framework, I generated Quantile-Quantile (QQ) plots to compare the observed LRT *p*-values versus expected *p*-values under the null hypothesis that there is no association. Across all cell-types,

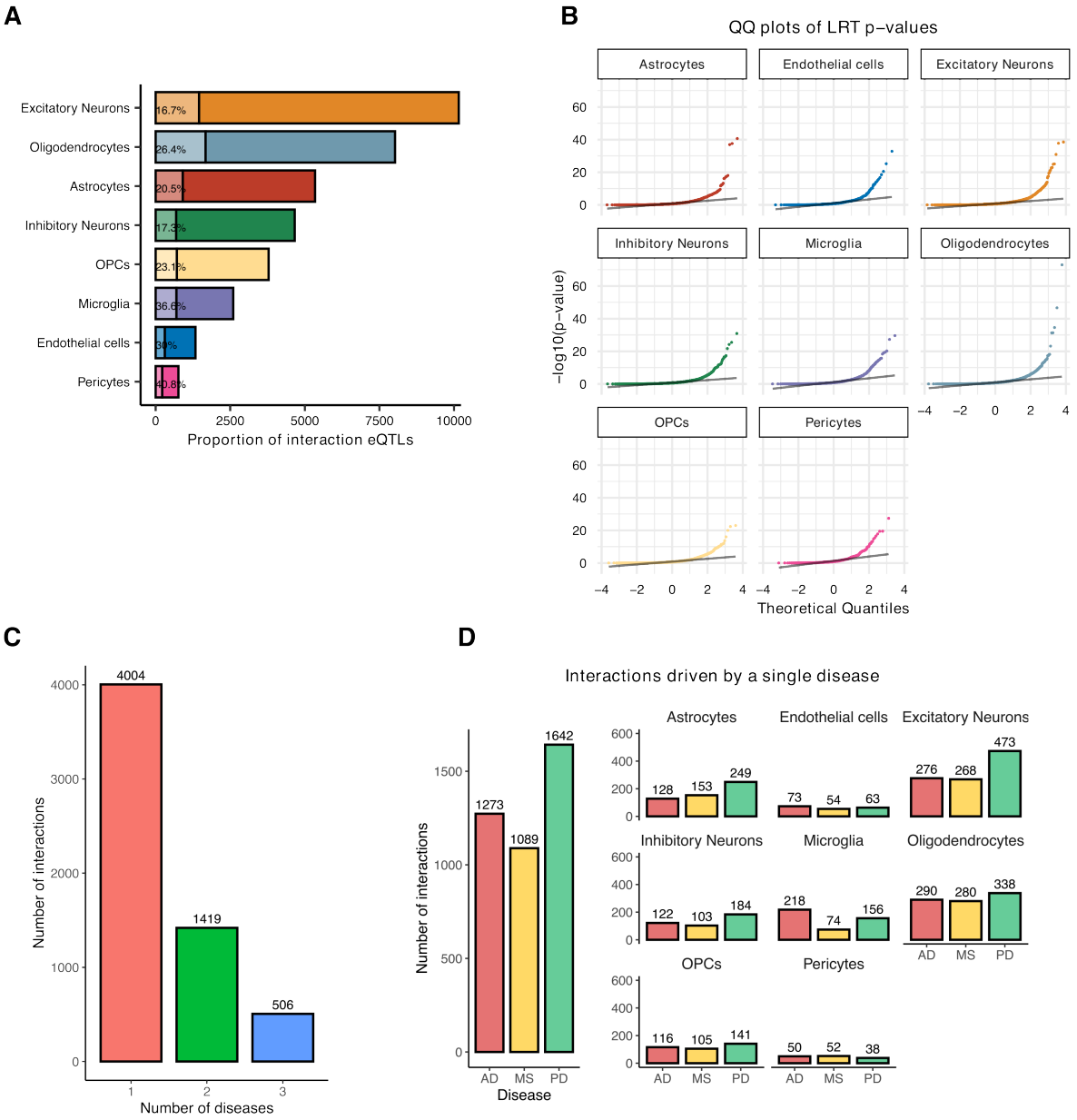


Figure 4.2: Interaction modelling results. ***A***, Proportion of eQTLs (top SNP-gene pair) with a disease interaction. The dark bar represents the total eQTLs discovered for that cell-type. ***B***, QQ-plots of LRT p-values, showing the distribution of p-values versus the expected distribution (black line). ***C***, Number of disease-interacting eQTLs (y-axis) affected by 1, 2 or 3 diseases (determined by the interaction p-value). ***D***, Number disease-interacting eQTLs affected by a single disease group (y-axis), split by category (AD, MS and PD). Left plot represents the total interaction eQTLs for all cell-types, right plot is split by cell-type.

there was a notable deviation from the expected p-value distribution suggesting an enrichment of associatins better modelled using the interaction terms (Figure 4.2 B). The plotted p-values were additionally corrected for false discovery rate within each cell-type (q-value < 0.05).

Of all interaction models showing a significant interaction (q-value < 0.05), the large majority were driven predominantly by a single disease group (4,004 out of 5,929, 67.5%) with a smaller proportion by two (1,419, 23.9%) and all three diseases (506, 8.5%) (Figure 4.2 C). In the interactions driven by a single disease, PD was the most common, (1,642 interactions driven by PD, 41.0%), followed by AD (1,273, 31.8%) and MS (1,089, 27.2%). This pattern repeated itself when looking by cell-type; however, microglia was the only cell-type where the majority of interactions were driven by AD cases, highlighting the impact of AD on microglial gene regulation[224] (Figure 4.2 D).

For example, the association between *rs117934759* and *PTPN12* expression in microglia was significantly better modelled with the interaction model M_2 (Figure 4.3 A) (LRT p-value = 1.72×10^{-16}, q-value = 4.14×10^{-14}). A closer inspection shows that the eQTL association was strongly influenced by AD samples ($AD_{interaction}$ p-value = 1.11×10^{-10}), with an almost non-existent association in the controls ($Controls_{interaction}$ p-value = 0.968) when considering the effect of disease cases. Given that the overall association is significant (M_1 p-value = 7.23×10^{-6}), it is likely that it is mostly driven by the presence of AD disease cases. In another example, this time in astrocytes, the association between *rs7038041* and *SLCO1B3* was also better modelled with an interaction term (LRT p-value = 3.11×10^{-7}, q-value = 1.74×10^{-5}) (Figure 4.3 B). In contrast to the previous example, expression is generally similar across all samples except within the PD cases which seem to be driving the association ($PD_{interaction}$ p-value = 5.08×10^{-4}). This is further supported by the weak association in the controls ($Controls_{interaction}$ p-value = 0.02). In a third example, as a result of the MS cases, the association between *rs62019445* and *FRMD5* in excitatory neurons also harbored a significant interaction (LRT p-value = 8.07×10^{-13}, q-value = 2.11×10^{-10}, $MS_{interaction}$ p-value = 6.25×10^{-10}) with a weak association in controls ($Controls_{interaction}$ p-value = 0.08) (Figure 4.3 C).

Despite the majority of interactions being driven by a single disease, there are many examples with less obvious effects contrasting the ones above. For instance the association between *rs7932358* and *ARHGAP20* expression in oligodendrocytes had a significant interaction with disease (LRT p-value = 4.75×10^{-8}, q-value = 2.10×10^{-6}). However in this example, the

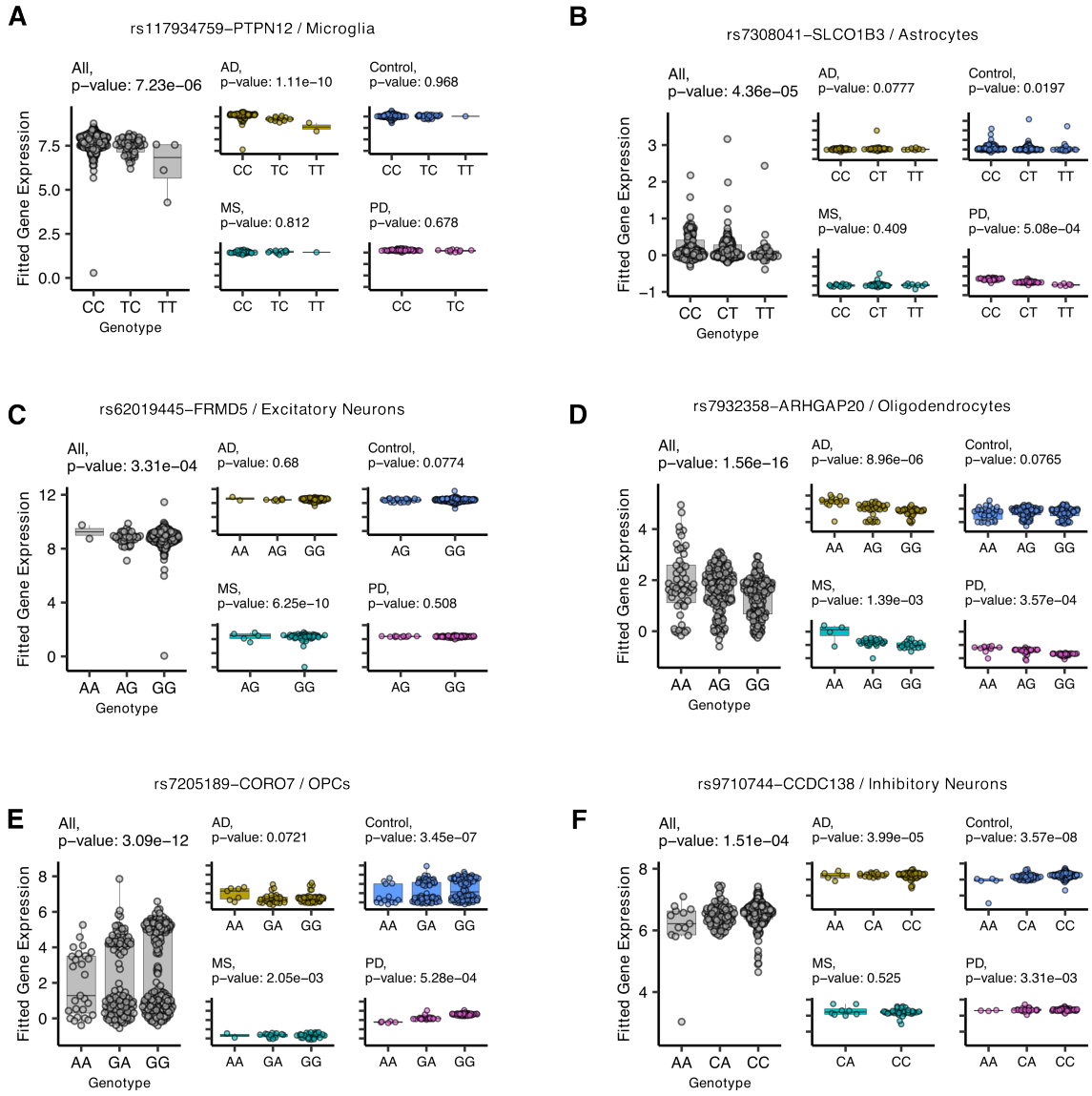


Figure 4.3: Examples of interaction-eQTLs. The left plots represent the fitted residual expression values (y-axis) from the M_1 model versus genotype dosage (x-axis) and its corresponding *p*-value (which is comparable to the MatrixEQTL discovery *p*-value, see Figure 4.1 C). The right plots represent the fitted values from the M_2 model, split by disease-group. For each disease group, the *p*-value represents the significance of the interaction between the genotype G and that disease group, and is interpreted as whether the effect is significantly different from that of the controls. The controls *p*-value reflects the main effect of genotype on expression after accounting for disease interaction effects. ***A***, Example of an interaction-eQTL in *rs117934759* and *PTPN12* expression in microglia. ***B***, Example of an interaction-eQTL in *rs7038041* and *SLCO1B3* expression in astrocytes. ***C***, Example of an interaction-eQTL in *rs62019445* and *FRMD5* expression in excitatory neurons. ***D***, Example of an interaction-eQTL in *rs7932358* and *ARHGAP20* expression in oligodendrocytes. ***E***, Example of an interaction-eQTL in *rs7205189* and *CORO7* expression in OPCs. ***F***, Example of an interaction-eQTL in *rs60935857* and *CCDC138* expression in inhibitory neurons.

association was affected by all three disease states ($AD_{interaction}$ p-value = 8.96×10^{-6}, $MS_{interaction}$ p-value = 1.39×10^{-3}, $PD_{interaction}$ p-value = 3.57×10^{-4}) with an overall weak association in controls ($Controls_{interaction}$ p-value = 0.08), suggesting a combined disease effect that drives the robust association (M_1 p-value = 1.56×10^{-16}) (Figure 4.3 D). It was also possible however for an overall strong association to be retained in the controls; for instance, the association between *rs7205189* and *CORO7* expression in OPCs was also found to have a significant interaction with disease (LRT p-value = 1.99×10^{-10}, q-value = 2.33×10^{-8}). However, this association was affected both by the PD samples ($PD_{interaction}$ p-value = 5.28×10^{-4}) and MS samples ($MS_{interaction}$ p-value = 2.05×10^{-3}), whilst retaining an overall strong association in the controls ($Controls_{interaction}$ p-value = 3.45×10^{-7}). This suggests that rather than the disease samples being the sole driving force, they are more likely contributors to the overall association (M_1 p-value = 3.09×10^{-12}) (Figure 4.3 E).

Finally, in the case of *rs60935857* and *CCDC138* in inhibitory neurons (LRT p-value = 4.75×10^{-6}, q-value = 1.55×10^{-4}), there was also influence from two disease states ($AD_{interaction}$ p-value = 3.99×10^{-5}, $PD_{interaction}$ p-value = 3.31×10^{-3}). Similarly to the previous example, there is an overall significant effect in the controls ($Controls_{interaction}$ p-value = 3.57×10^{-8}). However, the effect of the AD and PD cases are opposed to that of the control samples, resulting in an overall weaker association (M_1 p-value = 1.51×10^{-4}).

4.3.3 Controls versus full-cohort eQTLs

Given the substantial impact the presence of disease cases have on eQTL discovery, I sought to compare the eQTL profile from the full cohort (N = 391) to a controls-only subset (N = 183). Aside from the difference in statistical power, it can be assumed that there are associations in the full cohort that are exclusively driven by the presence of disease samples that would therefore not be present in the controls-only. In addition, the eQTL discovery in the controls-only subset is performed independently and as such results in a different set of optimized PCs to maximize eQTL discovery.

Using this controls-only cohort, I repeated the eQTL discovery using MatrixEQTL and found a total of 7,204 unique eGenes (FDR < 5%, 10,470 in aggregate) (Appendix 3). As previously, a large proportion of these were unique to a single cell-type (5,046, 48.2%) and only 8 eGenes

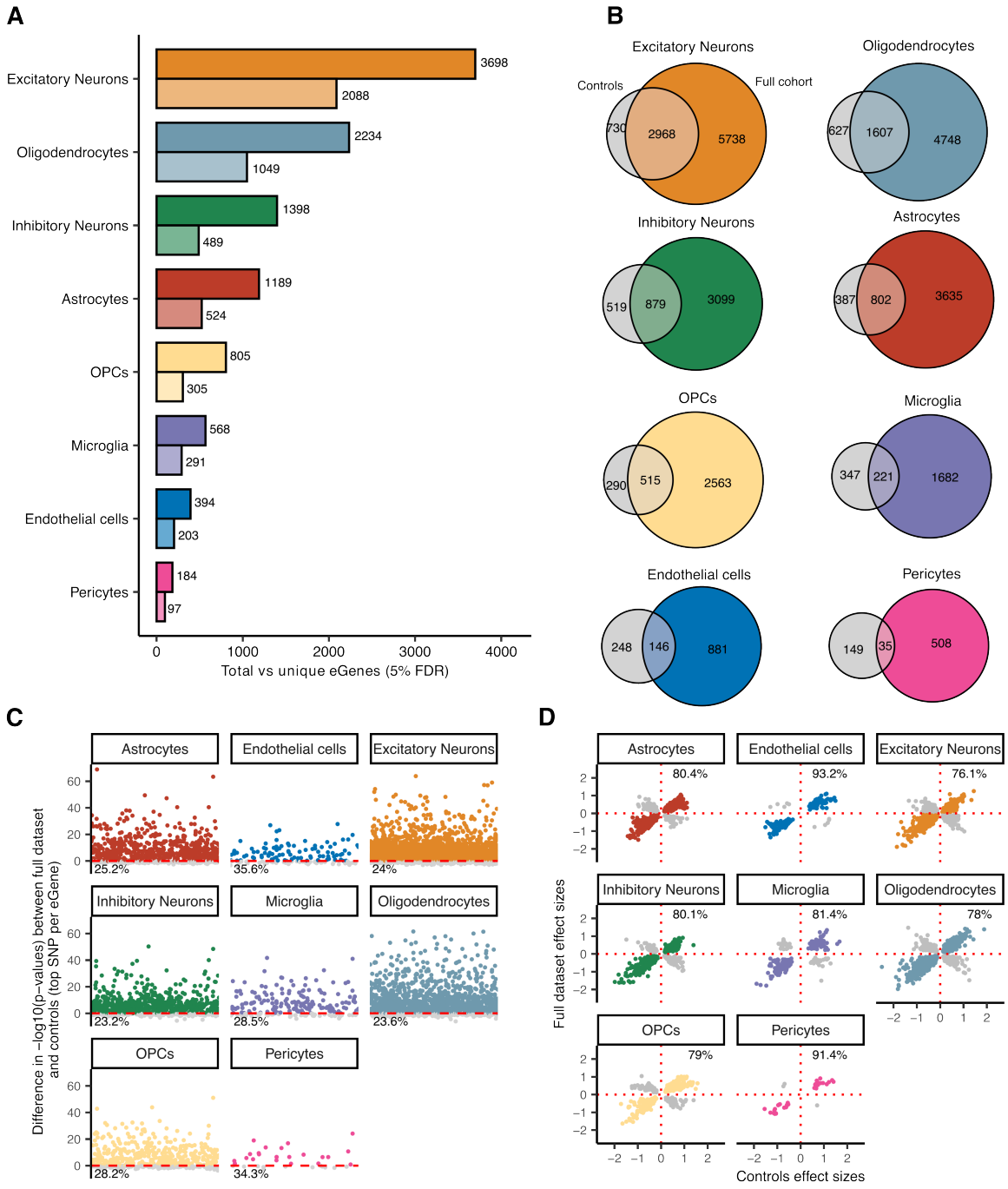


Figure 4.4: Controls-only eQTLs. ***A***, Number of significant eGenes discovered per cell-type in the controls-only cohort ($< 5\%$ FDR). Each set of bars for each cell-type (y-axis) represents the total number of eGenes discovered (darker, top bar) versus the number of eGenes discovered unique to that cell-type (lighter, bottom bar). ***B***, Venn diagrams of eGenes discovered in the full (right, colored circle) versus the controls-only cohort (left, grey circle). The central number represents the number of eGenes discovered in both cohorts. ***C***, Comparison of p-values for the top association of shared eGenes between the controls-only and full cohorts. The y-axis represents the difference in -$\log_{10}$(p-value) between the association in the full versus controls-only cohort, where a negative value (grey dots) means the association was stronger (a higher -$\log_{10}$(p-value) in the controls). Each dot represents one eGene that was present in both cohorts. ***D***, Comparison of effect sizes in the full (y-axis) versus controls-only (x-axis) cohorts. Each dot represents a shared eGene and its top association, where grey dots are highlighted to show opposite directions of effect.

were shared between all 8 cell-types (Figure 4.4 A). Compared to the full cohort, the number of unique eGenes discovered across all cell-types represented an 48.3% decrease (65.1% when considering eGenes in aggregate, i.e total eGenes across all cell-types even if shared). The largest decrease was observed in OPCs (73.8%) and astrocytes (73.2%), with the least number of eGenes lost in excitatory neurons (64.9%). Interestingly, between 19.7 - 81.0% of the eGenes discovered by cell-type in the controls were not present in the larger, better powered cohort. As expected, the largest change was observed in the smaller cell-types; 81.0% of eGenes discovered in pericytes in the controls-only cohort were not present in the larger discovery dataset. In contrast, better powered cell-types such as the excitatory neurons and oligodendrocytes had much smaller proportions of eGenes only discovered in the smaller cohort (19.7% and 28.1% respectively), but still considerable given the large difference in statistical power between the full and controls-only cohort (Figure 4.4 B).

When looking at the top SNP for shared eGenes between the two cohorts, the large majority of associations were much stronger (lower p-values and higher effect sizes) in the full cohort. However, between 23.2 - 35.6% of these had stronger associations in the controls, which could either be the result of disease presence weakening the association or simply that the association is more accurately measured in the full cohort (Figure 4.4 C). Notably, despite the majority of top SNPs differing for the shared eGenes between the controls and the full cohort (only between 22.6 - 42.9% had the same exact top SNP, which could be explained by small numerical differences) the majority of effect sizes were in the same direction (Figure 4.4 D).

Given the differences in discovered eQTLs between the full and controls-only cohort, there are several possibilities that explain these differences; first, the presence of disease samples may be driving associations in the full cohort, which explains why they are not observable in the controls-only cohort. Secondly it is probable that a large number of associations not observed in the controls-only cohort is due to loss of statistical power. Finally on the opposing side, there may be significant associations in the controls-only cohort that are countered by the disease-affected expression, thereby obscuring the overall association.

To understand these differences, I intersected the eGenes that were only discovered in the full cohort with the results from the interaction analysis. Interestingly, between 13.9 - 40.9% (by cell-type) of the eGenes discovered only in the full cohort had a significant interaction with disease (q-value < 0.05) (Figure 4.5 A). The association between *rs6052519* and *SLC52A3* for

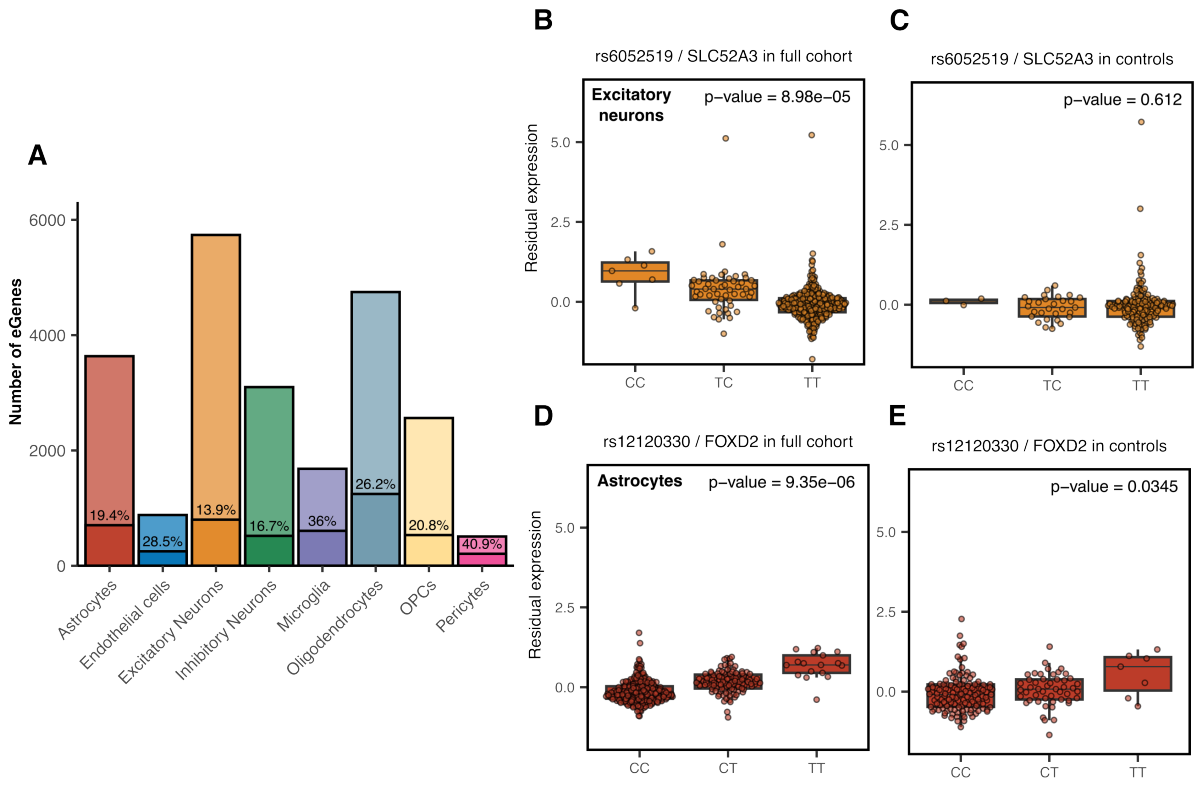


Figure 4.5: Impact of interactions on controls-only eQTLs. ***A***, Number of eGenes discovered only in the full cohort by cell-type (y-axis), and the percentage of which had a significant interaction with disease. ***B***, Association between *rs6052519* and *SLC52A3* in excitatory neurons as seen in the full cohort, as well as the controls-only cohort showing a non-significant association (***C***). ***D***, Association between *rs12120330* and *FOXD2* in astrocytes in the full cohort and the controls-only cohort (***E***).

instance in excitatory neurons was only found in the full cohort (MatrixEQTL p-value = 8.98 $\times$ 10^{-5}) (Figure 4.5 B) but not in controls (MatrixEQTL p-value = 0.61) (Figure 4.5 C). When testing the association in an interaction model, there was a significant interaction from the PD cases (LRT p-value = 2.22 $\times$ 10^{-12}, q-value = 5.00 $\times$ 10^{-10} , $PD_{interaction}$ p-value = 2.31 $\times$ 10^{-8}) and a non-significant main effect ($Controls_{interaction}$ p-value = 0.73). Similarly, the association between *rs12120330* and *FOXD2* in astrocytes was only found in the full cohort (MatrixEQTL p-value = 9.35 $\times$ 10^{-6}) (Figure 4.5 D) but not in controls (MatrixEQTL p-value = 0.03) (Figure 4.5 E), seemingly driven by AD driven expression (LRT p-value = 1.22 $\times$ 10^{-7}, q-value = 8.29 $\times$ 10^{-6}, $AD_{interaction}$ p-value = 3.79 $\times$ 10^{-5}).

4.3.4 Age-related interaction eQTLs

The results above show that the presence of disease cases have a significant impact on the discovery of eQTL associations. In the case of the controls-only cohort, the use of control tissue therefore assumes that the associations discovered are unconfounded by disease-modified expression. It is however possible that the patients harbor disease susceptibility despite neurological assessment of the patients and neuropathological assessment of the tissue. The implication is that some patients will indeed be on the path to disease development (which is not unlikely given the high median age of this cohort) that predates observable pathology or symptomatic expression. I therefore sought to identify eQTLs which interacted with age dynamically.

Using age as the interaction term, I modelled all of the significant associations found in the controls-only cohort between each unique eGene by cell-type and its top SNP. The aim of this interaction model is to capture whether the effect of the genotype on expression changes with age, either through reduced or increased expression. Following the previous approach using the top SNP for each unique significant eGene by cell-type, in all 10,470 tested associations there were 1,605 (15.3%) which were significantly better modelled with age-genotype as an interaction term (q-value $<$ 0.05) (Appendix 4). This proportion as usual varied by cell-type. In the order of least to most impacted, the number of associations that were significantly better modelled using age as an interaction term (M2) were as follows; excitatory neurons (292, 7.9%), inhibitory neurons (146, 10.4%), astrocytes (153, 12.9%), oligodendrocytes (430,

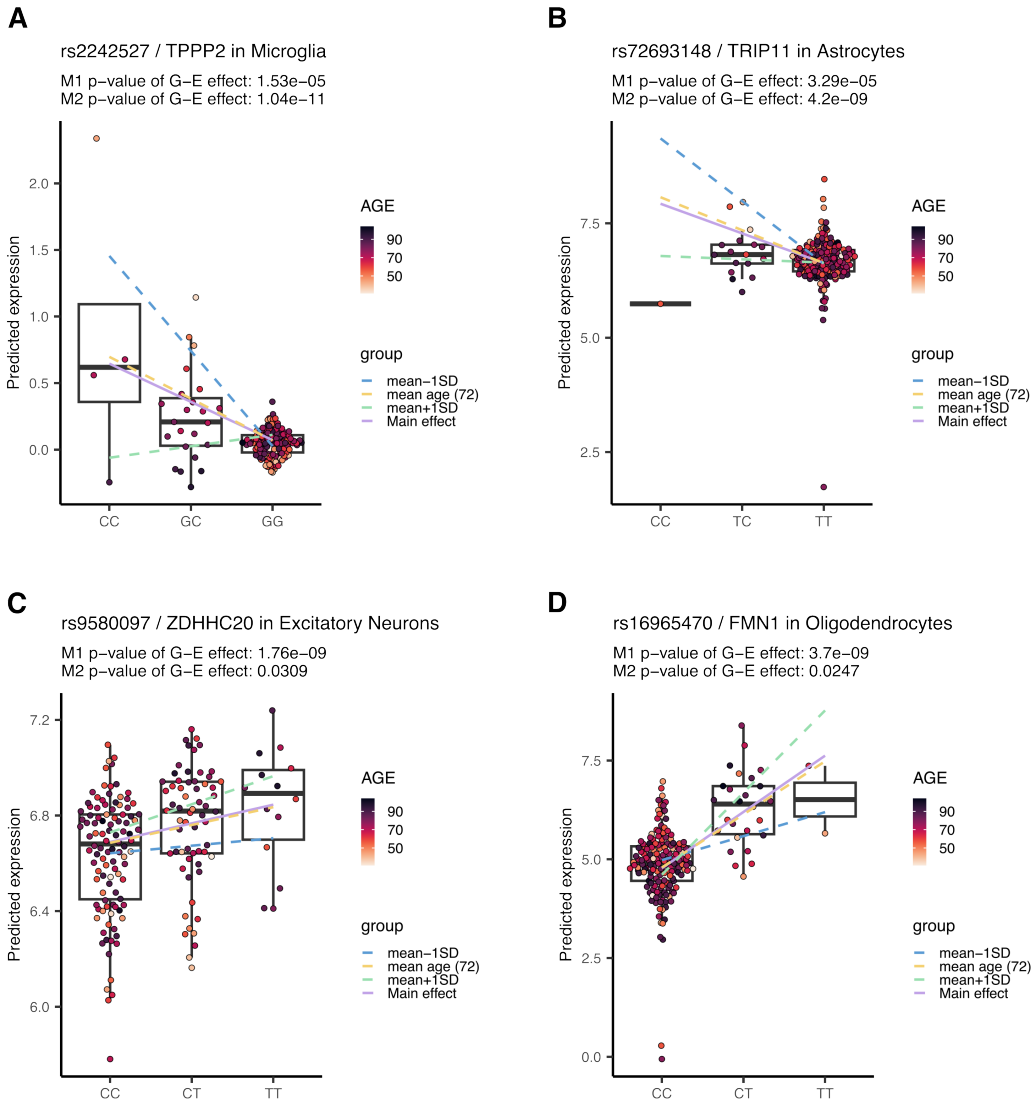


Figure 4.6: Impact of age on eQTL associations. Using a similar configuration to the previous M_2 model, significant associations were re-tested where age was modelled with an interaction term on genotype to identify whether effects of genotype on expression varied with age. ***A***, Association between *rs2242527* and *TPPP2* in Microglia, showing a significant effect of age on the association between genotype and expression. The M_1 model p-value reflects the standard eQTL model. The M_2 model p-value reflects the effect of genotype on expression when accounting for changes in gene expression across age (purple line, "main effect"), signifying in this case a more robust association. The dotted lines represent the predicted value of expression based on genotype at mean (yellow) and -(blue)/+(green) one standard deviation. ***B***, Association between *rs72693148* and *TRIP11* in astrocytes, again showing a more robust association when accounting for age. ***C***, Association between *rs9580097* and *ZDHHC20* in excitatory neurons, showing an association in majority driven by age as well as in ***D***, in the association between *rs16965470* and *FMN1* in oligodendrocytes.

19.3%), microglia (137, 24.2%), OPCs (220, 27.3%), endothelial cells (144, 36.5%) and pericytes (83, 45.1%).

To exemplify these age interactions, four examples are included in Figure 4.6. In interpeting the outputs of these models, the LRT p-value (and its q-value which is corrected for multiple testing) is the significance with which the M_2 model is preferred over the M_1 model. The M_2 age interaction p-value determines the significance with which the effect of genotype on expression changes with age. For instance in microglia, there was a significant interaction with age in the association between *rs2242527* and *TPPP2* (LRT p-value = 1.62×10^{-10}, q-value = 1.33×10^{-8}, M_2 age interaction p-value = 1.76×10^{-9}) (Figure 4.6 A). In the M_1 model (replicating the MatrixEQTL discovery model), there was a significant association (M_1 p-value = 1.53×10^{-5}, MatrixEQTL p-value = 1.07×10^{-5}). When accounting for age as an interaction term however, the SNP-gene association is much stronger (M_2 p-value = 1.04×10^{-11}), signifying that age-related changes in gene expression of *TPPP2* diminish the strength of the association. The same example is apparent in the association between *rs72693148* and *TRIP11* in astrocytes (LRT p-value = 3.6×10^{-10}, q-value = 1.33×10^{-8}, M_2 age interaction p-value = 2.67×10^{-7}), where expression is generally lower in aged samples which inhibit the association (M_1 p-value = 3.29×10^{-5}, MatrixEQTL p-value = 2.43×10^{-5}). When accounting for this effect, the association is much more robust (M_2 p-value = 4.20×10^{-9}) (Figure 4.6 B).

On the other hand, the opposing effect was also observed. In the association between *rs9580097* and *ZDHHC20* in excitatory neurons, the interaction model was also preferred (LRT p-value = 1.20×10^{-6}, q-value = 1.61×10^{-4}, M_2 age interaction p-value = 2.03×10^{-4}). In this case however, age is disproportionately contributing to the effect (M_1 p-value = 1.76×10^{-9}, MatrixEQTL p-value = 2.01×10^{-9}), with aged samples generally having higher expression which when accounted for inhibits the association (M_2 p-value = 0.03) (Figure 4.6 C). A similar example is observed in oligodendrocytes between *rs16965470* and *FMN1* (LRT p-value = 3.99×10^{-9}, q-value = 3.95×10^{-7}, M_2 age interaction p-value = 8.87×10^{-5}). Again here, the strong association in the discovery (M_1 p-value = 3.70×10^{-9}, MatrixEQTL p-value = 4.51×10^{-10}) is mainly driven by differences in expression from higher age individuals (M_2 p-value = 0.03) (Figure 4.6 D).

4.4 Discussion

In this chapter, I show that the presence of disease samples have a substantial impact on eQTL associations despite explicit correction of disease status prior to eQTL discovery using MatrixEQTL. While the use of random effects in linear mixed effects models may capture additional variability and more accurately reflect the impact of sample source on these associations they are largely comparable to initial eQTL discovery. However, use of interaction terms between disease status and genotype to reflect the impact of disease on the overall association shows that a large proportion of associations favour this model (16.7 - 40.8%, depending on cell-type). In addition, comparison of eQTLs discovered in the full cohort (N = 391) versus a controls-only subset (N = 183) shows a drastic loss in eQTL discovery. While the majority of associations lost is likely due to a loss in statistical power, I show here that a proportion of eQTLs discovered in the full dataset that are not present in the controls is likely due to the presence of disease driving or contributing associations. Finally, I show that even in the absence of disease status, age also has an impact on eQTL associations.

4.4.1 Impact of disease on eQTL associations

The results show here that a large proportion of eQTL associations are significantly impacted by disease cases (16.7%-40.8% depending on cell-type). Changes in gene expression in neurons compared to controls is well known[186,211], although in this case they seem to be the cell-types least impacted by interactions (16.7% and 17.3% in excitatory and inhibitory neurons, respectively). Many of these cell-types have all been independently linked to specific diseases including oligodendrocytes and OPCs in AD[225], and glia more broadly in both PD[226] and MS[227]. Identifying the specific roles for these cell-types and their changes in gene expression at scale in relation to disease presence has become much more feasible in recent years through single-cell sequencing; while certain hallmarks are emergning it is difficult to define an expectation into the impact of disease on gene expression and their associations.

In the majority of cases, a single disease group was responsible for the interaction. This in itself is not surprising given generally different genetic architectures across brain disease[12], but whether this is by chance or lack of power in each respective group remains to be understood. In cases where only one disease was responsible, PD was the most common in

almost all cell-types except microglia and pericytes. This result is surprising for two reasons; first, PD samples did not make up the majority of disease cases (68 out of 208) thereby excluding statistical power as a reason. Secondly, the genetic architecture of PD cases from GWAS studies compared to AD or MS is to some degree less complex with less loci involved[40,42,227]. Microglia was the only cell-type where AD had the most interactions which is convincing given the increasing body of evidence in the last few years of its crucial role in AD pathology[224]. In a smaller proportion of interaction eQTLs, more than one disease was responsible for the interaction, which could potentially point towards shared disease pathways. Investigating these genes more closely and whether they have previously been involved in disease overlap will be of interest in future work.

4.4.2 Differences between controls and full cohort

When comparing the results of the full cohort (N = 391) versus the controls-only (N = 183), large differences are expected, especially when considering the drop in statistical power (more than a 50% decrease in sample size). It is safe to assume that the majority of eGenes discovered in the full cohort are therefore the result of better statistical power, be it on the expression side through more accurate gene expression estimates or through more accurate genotype spreads in variants with low minor allele frequency. To some extent however, the results show that some eGenes were purely driven by the presence of disease cases which explains why they would not be observed in a controls-only sample set. This is a crucial point to consider in future studies combining disease and control cases to maximize statistical power, especially if downstream analyses attempt to draw conclusions in relation to disease pathology.

An interesting finding here however was that a proportion of eGenes were only discovered in the controls-only cohort, suggesting the possibility of disease presence obscuring the association in the larger dataset. To know this with certainty however, the interaction modelling should in future work be extended to all possible *cis*-associations across the genome which is a computationally challenging task. For reference, 30,027 associations were tested in these results, representing the top SNP for each significant eGene in each cell-type; counting all possible SNP/gene pairs in *cis* in this dataset, there are more than 60 million associations to test (although they could be restricted to independent SNPs per gene).

The results here do show however that controls-only cohorts have value on their own in finding associations unconfounded by disease presence, and will be of interest when attempting to identify dynamic eQTL states in transitions from control to disease. While this is not possible in the post-mortem human brain which is restricted to a single time point (control or disease upon death), if these dynamic eQTLs are also represented by protein QTLs (pQTLs) as assayed through cerebrospinal fluid or plasma[82] they could provide important biomarkers to track disease progression.

4.4.3 Age-related interactions

Brain ageing is well-known to be an important risk factor in neurodegenerative disease[228] and as such understanding both the linear[229] and more recently non-linear[230] dynamics of aging on multi-omic modalities is crucial in identifying checkpoints for intervention or prevention. Importantly, age-related changes in gene expression that pertain to disease biology likely predate any observable pathology and/or clinical symptoms. Despite using the smaller controls-only cohort, the results show here that indeed there is a large amount of eQTLs that are impacted by changing levels of gene expression across age; in some eQTLs, aged samples may drive associations entirely and would presumably not be discovered in a cohort with younger individuals, while in others they have the opposite effect and in fact decrease the association strength revealing a more robust association when accounting for age.

These results are important for several reasons. First, age-related eQTLs could well overlap with disease-specific eQTLs and provide potential targets for intervention or prevention, assuming a model where age modifies gene expression which in turn promotes disease progression. On the opposite side, they help isolate disease-specific eQTLs which are not affected by age and as such present a more sudden change in expression important for its pathogenesis. Finally, the results show that this framework is broadly applicable to other quantitative covariates which are increasingly accessible from electronic health records, and could uncover novel dynamic eQTLs that pre-date disease pathology[231].

Chapter 5

Colocalisation of cell-type specific eQTLs across the brain phenome

5.1 Preface

5.1.1 Appendices

The appendices referenced in this chapter are Appendices 5-8, organised as follows;

- Appendix 5: List of GWAS included
- Appendix 6: COLOC results (full dataset, N = 391)
- Appendix 7: Interaction COLOCs
- Appendix 8: COLOC results (controls-only cohort, N = 183)

5.2 Introduction

Genetic colocalisation has been used extensively to integrate eQTL data with GWAS in attempts to identify mechanisms that may explain disease risk or phenotypic presentation. Importantly this integration allows the pinpointing of genes in genome-wide significant risk loci, thereby linking tissue-specific expression, putatively causal genetic variant and disease risk.

This approach has been applied in a variety of tissues in recent years, including heart[232], lung[233], liver[234], subcutaneous fat[235], ocular tissue[236] and more linking a range of phenotypic traits through publicly available GWAS summary statistics. It has even prompted the generation of bespoke computational pipelines to perform eQTL / COLOC analysis such as "ColocQuial"[237] and "echolocatoR"[238], making use of large eQTL (such as GTEx) and GWAS summary statistics repositories (such as the openGWAS portal).

These frameworks have also been applied in the brain. In 2021, Kia et al., overlapped eQTL data with the latest PD GWAS using a combination of Braineac, GTEx and CommonMind published eQTLs to uncover 5 genes (*WDR6*, *CD38*, *GPNMB*, *RAB29*, and *TMEM163*) which colocalised with PD risk, with further supporting evidence from methylation and splicing on these genes[137]. Further, the *Metabrain* study used their landmark eQTL study on more than 8,000 individuals to perform colocalisation for 13 brain-related traits, including neurological disorders such as PD, AD, MS and amyotrophic lateral sclerosis (ALS); psychiatric disorders, such as schizophrenia, bipolar disorder and broad depression; and quantitative phenotypes such as intracranial volume, cognitive function and years of schooling. With this methodology, they discovered 186 eQTLs which colocalised with these traits, each linked by the same putatively causal SNP and providing novel potential targets[90].

Lastly and as expected through the recent increase in single-cell eQTL studies, the eQTL research community has been able to link specific cell-types to colocalised genes. In the largest single-cell eQTL study to date (982 individuals), Yazar et al., identified 108 genes colocalised between immune cell-types (530 gene / cell-type pairs) and loci in MS, rheumatoid arthritis, systemic lupus erythematosus, crohn's disease and more[107]. In early 2024, a study using both individual cases of idopathic pulmonary fibrosis (IPF) and controls (66 vs 48) mapped eQTLs across 38 cell-types. Despite this modest sample size, they discovered 108 genes which colocalized with IPF risk using available GWAS summary statistics, with multiple cell-types implicated[239]. Finally in the brain, Bryois et al., using a sample size of 192 individual for cell-type specific eQTL discovery in the brain discovered cell-type specific colocalised signatures between eQTLs and brain disease (such as schizophrenia, PD, AD and MS), including microglia-specific signals in AD. Interestingly, the majority of colocalised loci were cell-type specific, highlighting the value of using cell-type specific eQTLs to pinpoint causal genes[108]. However, the study was of modest statistical power and only focused on a few

specific traits, which presents an opportunity to extend cell-type specific eQTLs to less well-studied phenotypes.

5.2.1 Aims and objectives:

As it stands, there has not yet been a thorough investigation of cell-type specific colocalisation patterns between eQTLs and traits across the human brain phenome. In addition, the impact of state-specific or state-influenced eQTLs on colocalisation, as well as the difference in colocalisation patterns between eQTLs discovered in a control-only cohort compared to a mixed dataset has not been explored. The aims of this chapter are the following;

- Establish a carefully curated resource of GWAS summary statistics across the brain phenome.
- Investigate cell-type specific colocalisation patterns between eQTLs and brain phenotypes.
- Explore differences between colocalisations arising from controls-only versus mixed cohort eQTLs.
- Identify the impact of state-specific eQTLs on colocalisations.

5.3 Results

5.3.1 Overview of GWAS datasets

In total, I identified 42 traits with available genome-wide statistics across the brain phenome, summarised in (Appendix 5). These GWAS spanned 9 major categories, including neurodegenerative diseases, neurological conditions, psychiatric disorders, sensory disorders, substance use disorders, sleep and well-being, cognitive traits and pain-related conditions. Following pre-processing and harmonization using the *MungeSumstats* package[201], most loci of genome-wide significance (defined as regions of 1Mb around a lead SNP in a non-overlapping fashion, with a minimum p-value threshold of 5×10^{-8}) were found in the SCZ GWAS (n = 238), followed by intelligence (INT) (n = 198) and sleep duration (SD) (n = 81) (Figure 5.1 A). On average, this equated to an average of 30 loci per trait (median = 14).

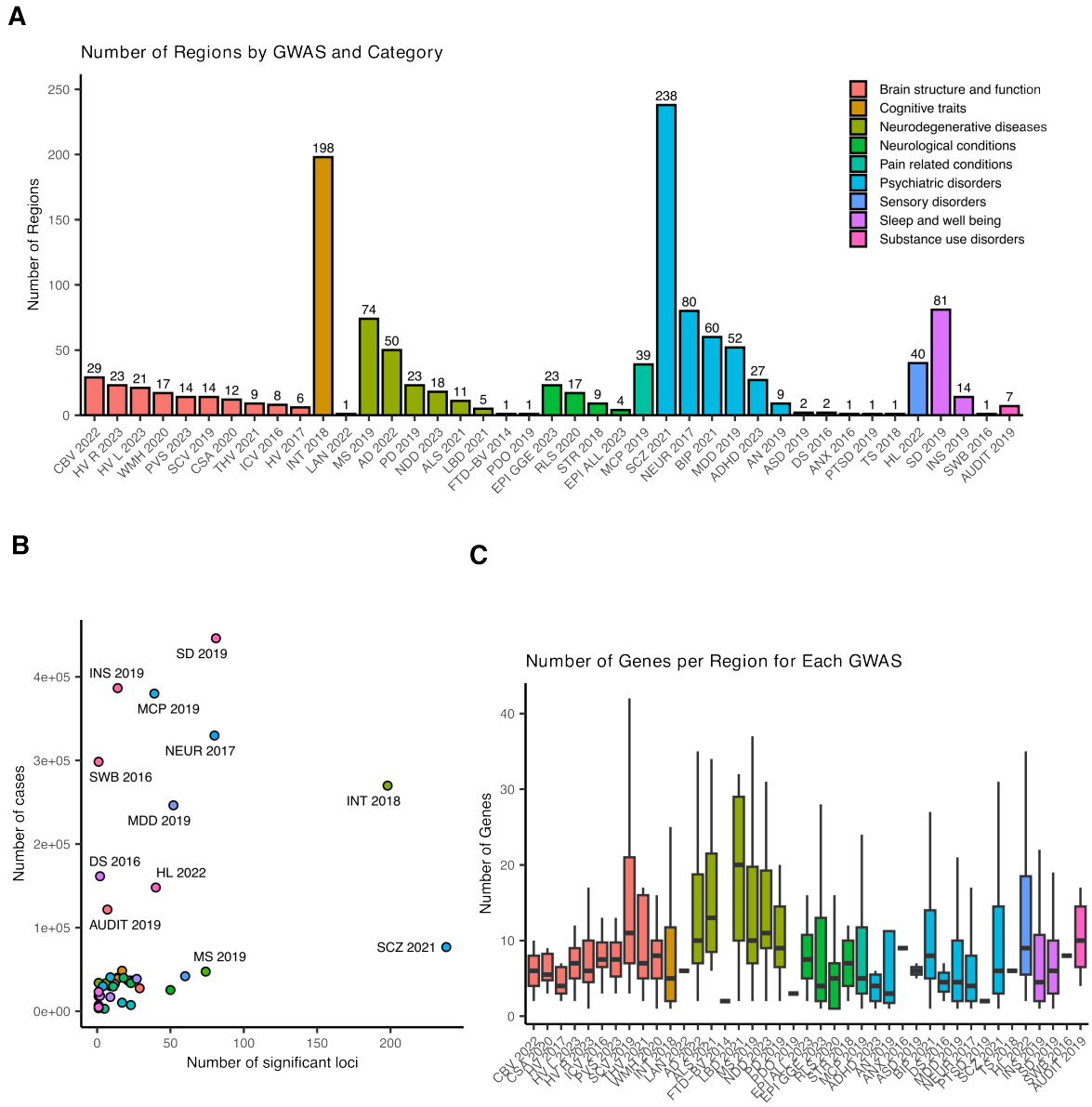


Figure 5.1: Overview of GWAS included. ***A***, Overview of GWAS (x-axis) across 9 major brain-related phenotypic categories (by color), and the total number of genome-wide significant loci assessed (y-axis) for each. ***B***, Number of cases used in study (y-axis) versus number of genome-wide significant loci. Each point represents a GWAS. ***C***, Boxplots of number of genes (y-axis) per locus for each GWAS. The boxplots represent median number of genes, min and maxima represent the lowest number of genes found in a locus and the most genes found in a locus, respectively.

The number of loci retained after processing did not necessarily correlate with number of cases (or total number of individuals for non- case/control traits, Figure 5.1 B). For instance, the major depressive disorder GWAS (MDD) included ~246,000 cases and discovered 39 loci of genome-wide significance, whereas the SCZ GWAS by comparison only included ~77,000 individuals leading to the discovery of 238 significant loci. In a stark contrast, the GWAS on subjective well-being (SWB) included ~298,000 participants but only yielded a single locus. In neurodegenerative phenotypes, MS had the most number of loci (n = 74) despite a modest number of cases (~68,000) followed by AD (n = 50), which included >111,000 cases.

Across the 42 different traits, there were 1,163 unique regions as defined by their lead SNP, with a total of 6,826 unique genes overlapping these regions (11,885 total genes as the sum of all GWAS genes); on average, there were 10 genes overlapping each region (median 7), with a minimum of one gene and a maximum of 53 as seen in the *rs17671352* locus in hearing loss (HL), located on chromosome 7. These numbers however varied greatly by GWAS (Figure 5.1 C).

5.3.2 Colocalisation results using the full eQTL cohort

Using the cell-type specific eQTLs discovered and desdribed in Chapter 3, I performed genetic colocalisation using COLOC[134] between the eQTLs from each cell-type and each of the 42 GWAS traits (see Chapter 2 for full methodology). Briefly, the eQTLs for each gene within a genome-wide significant locus for a trait was tested (per cell-type) for genetic colocalisation, where the posterior probability (PP.H4) greater than 0.8 represents a successful colocalisation.

In total, I discovered 501 colocalisations in 30 different traits (Figure 5.2 A)(Appendix 6). The large majority of obtained colocalisations were observed in intelligence (INT, n = 79) and SCZ (n = 86), followed by AD (n = 35), where each colocalisation represents a gene/cell-type pair. Even traits with a small number of loci such as behavioural frontotemporal dementia (FTD-BV) ($n_{loci} = 1$) , language ability (LAN, $n_{loci} = 1$) and lewy body disease (LBD, $n_{loci} = 5$) were successful in discovering colocalisations, supporting the value of using cell-type specific eQTLs for colocalisation analysis. In addition, the number of discovered

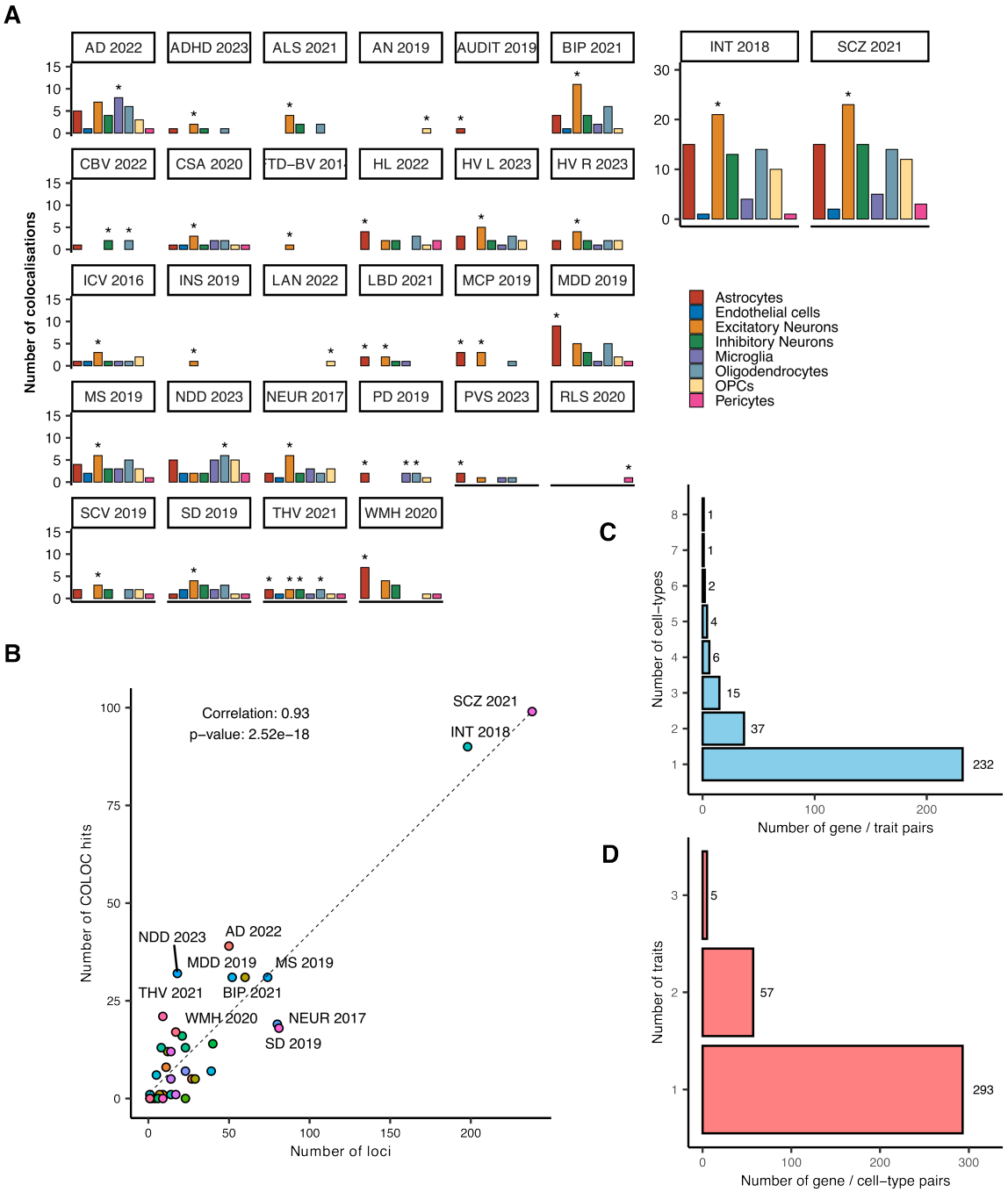


Figure 5.2: Overview of colocalisation results. ***A***, Summary of colocalisations (PP.H4 > 0.8) per cell-type and trait (see appendix 5 for each GWAS trait code). Each bar chart is coloured by cell-type. Y-axes indicate the number of colocalisations in that cell-type. Stars indicate the cell-type with most colocalisations with a particular trait. ***B***, Correlation of number of colocalisations (cell-type / gene / trait triplets) per trait (y-axis) versus number of genome-wide significant loci for that trait (x-axis). ***C***, Number of gene / trait pairs (x-axis) and the number of cell-types they are found in (y-axis). ***D***, Number of gene / cell-type pairs (x-axis) and the number of traits in which they appear (y-axis). For example, only 5 gene / cell-type pairs appear in 3 traits.

colocalisations correlated strongly with the number of loci per trait (Pearson $r = 0.93$, p-value $= 2.52 \times 10^{-18}$ (Figure 5.2 B).

The majority of colocalisations were found in excitatory neurons ($n_{colocs} = 124$), followed by astrocytes ($n_{colocs} = 93$), oligodendrocytes ($n_{colocs} = 84$) and inhibitory neurons ($n_{colocs} = 70$); OPCs and microglia had 55 and 43 colocalisations, respectively, whereas endothelial cells ($n_{colocs} = 17$) and pericytes ($n_{colocs} = 15$) had the lowest amount of colocalisations, as expected by the low number of discovered eQTLs. Generally speaking, the larger the eQTL discovery for each cell-type the more colocalisations uncovered, with the exception of astrocytes (3rd most eQTLs, 2nd most colocalisations). Interestingly, excitatory neurons was the cell-type with the most colocalisations in 15 traits, followed by astrocytes with the most colocalisations in 5 traits. Echoing multiple streams of evidence, microglia was the cell-type with the most colocalisations in AD[108,224].

A substantial proportion of these colocalisations were assigned to a single cell-type. Of the 298 gene/trait pairs (excluding *HLA* and *MAPT* regions), 232 (77.9%) were assigned to a single cell-type, showing a high degree of cell-type specificity (Figure 5.2 C). Conversely, 293 of the 355 gene/cell-type pairs (79.0%) were found in a single trait (again, excluding HLA and MAPT regions) (Figure 5.2 D). Interestingly, only 5 gene/cell-type pairs were found in 3 traits, although many of these are the result of overlapping loci from meta-analysis GWAS such as "common factor" neurodegenerative diseases (NDD) and AD, or hippocampus volume left (HV-L) versus right (HV-R). For instance, the association between *VCAN* expression in Astrocytes colocalised with HV-L (PP.H4 = 0.90), HV-R (PP.H4=0.90) and white matter hyperintensities (WMH) (PP.H4=0.90), although this locus had a different lead SNP (as proposed by COLOC) for each trait.

5.3.3 Exampes of cell-type specific colocalisations

One of the main advantages in using cell-type specific eQTLs for colocalisation with phenotypic traits is the ability to link specific cell-types to genes that are genetically linked to phenotypic risk. As observed in AD, and as is the case for the majority of traits, colocalisations are highly cell-type specific (Figure 5.3 A). Of the 35 colocalisations (PP.H4 > 0.8) in AD, 25 (71.4%) were found in a single cell-type. Only 3 genes (*PRSS36* , *CR1* and

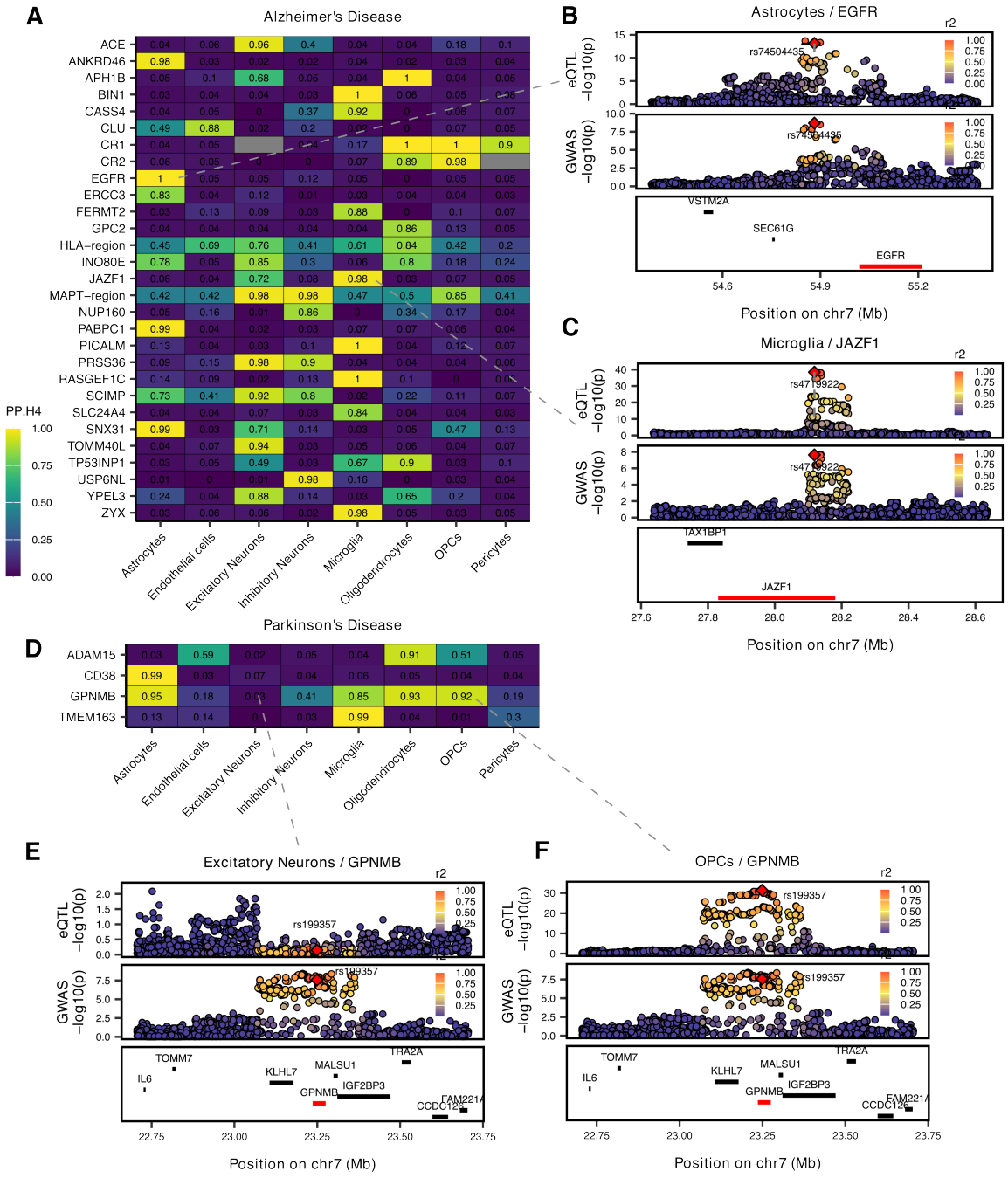


Figure 5.3: Examples of cell-type specific colocalisations. ***A***, Colocalisation results for AD. Each tile represents a posterior probability (PP.H4) for genes (y-axis) in all 8 cell-types (x-axis), where PP.H4 > 0.8 indicates colocalisation. Grey tiles represent genes that were not tested in that cell-type due to low expression. ***B***, Example regional association plot for *EGFR* in astrocytes and AD (PP.H4 = 1.00). Each point represents a single association for a SNP and either the gene expression (top plot) or the trait (AD risk, bottom plot) and its position along the genome (x-axis). The y-axis represents the strength of the association (negative $\log_{10}$ transformed p-values). The color of the points represents the degree of correlation (r^2) with the lead SNP (red, diamond shaped and labelled point). ***C***, Regional association plot for *JAZF1* in Microglia and AD risk (PP.H4 = 0.98). ***D***, Colocalisation results for PD. ***E***, Regional association plot for *GPNMB* in excitatory neurons, demonstrating lack of colocalisation with PD (PP.H4 = 0.03). ***F***, Regional association plot for *GPNMB* in OPCs and PD risk (PP.H4 = 0.92).

CR2) were found in multiple cell-types as well the *MAPT*-region. For example, genetic regulation of *EGFR* expression had a high posterior probability of colocalisation (PP.H4 = 1.00) but exclusively in astrocytes (Figure 5.3 B). Similarly, *JAZF1* expression had a high posterior probability of colocalisation (PP.H4 = 0.98) with AD in microglia and no other cell-type except a marginal colocalisation probability in excitatory neurons (PP.H4 = 0.72) (Figure 5.3 C). Many of these genes, aside from being highlighted as being important in the original GWAS[40] have been highlighted in other studies, including *EGFR*[240], *JAZF1*[241] and *BIN1*[108,209]specifically in microglia, as well as *PICALM*[242] and many more[243,244] that are for the first time highlighted here with cell-type specific context.

Albeit less prevalent, the same examples can be seen in PD. Interestingly, *GPNMB* colocalised with all glial cell types ($PP.H4_{astrocytes} = 0.95$, $PP.H4_{oligo} = 0.93$, $PP.H4_{OPCs} = 0.92$, $PP.H4_{microglia} = 0.85$), but not in neurons (Figure 5.3 E), pericytes or endothelial cells. As an example, I show here the colocalisation pattern for genetic regulation of *GPNMB* expression in OPCs with PD risk (Figure 5.3 F). Despite all colocalising in the same region, each celltype had a different proposed "lead SNP" by the coloc framework. All other genes (*CD38* , *TMEM163* and *ADAM15*) were only found in specific cell-types (astrocytes, oligodendrocytes and microglia, respectively) and have also been previously linked to PD risk in other studies[90,108].

5.3.4 Shared patterns of colocalisation

From the 501 colocalisations discovered, there were 249 unique genes across the 30 traits. Of these, 47 genes appeared in more than one GWAS, with the *MAPT* and *HLA* regions being highly ubiquitous, found in 8 and 14 traits out of 30, respectively (Figure 5.4 A). The *ACE* gene, for instance, was found to colocalise exclusively in excitatory neurons for both AD (PP.H4 = 0.96) as well as SCZ (PP.H4 = 0.98), suggesting a potential shared mechanism of action (Figure 5.4 B, C). Both have been described separately in AD[243] and a mendelian randomisation study in SCZ[245] as potential clinical targets.

In another example, perivascular space burden (PVS) and white matter hyperintensities (WMH) have previously been correlated in numerous imaging studies comparing topological data[246,247]. In line with this data, the results presented here show that these two traits shared

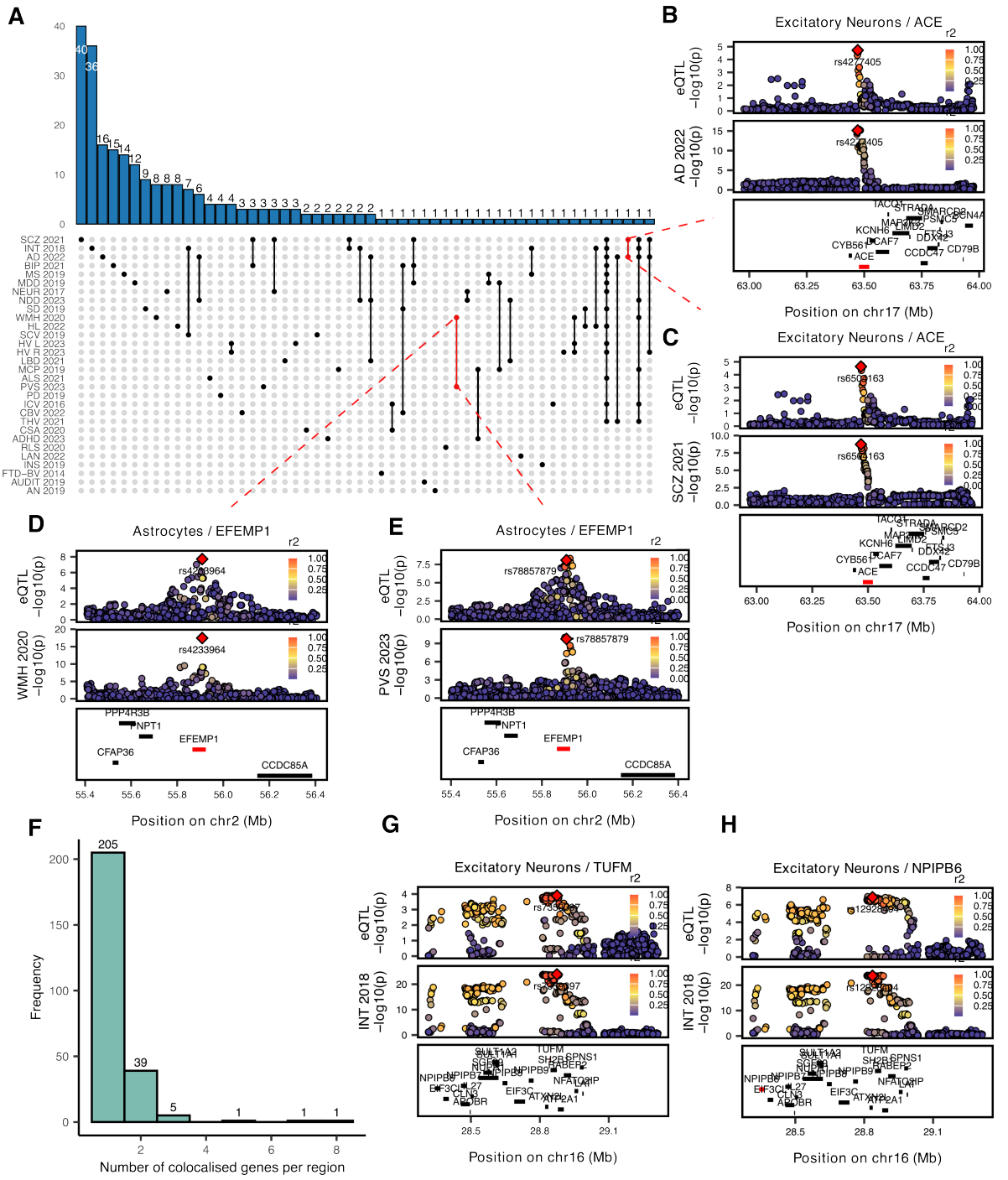


Figure 5.4: Shared colocalised genes. ***A***, UpSet plot of unique colocalised genes (n = 249) across 30 brain traits. Each set "column" represents a set of genes; for example, 40 genes were only found in schizophrenia (SCZ), and 7 genes were shared betweeen intelligence (INT) and subcortical volume (SCV). ***B***, Colocalisation between *ACE* expression in excitatory neurons and alzheimer's disease (AD), also present in schizophrenia (SCZ) (***C***). ***D***, Colocalisation between *EFEMP1* expression in astrocytes and white matter hyperintensities (WMH), also present in perivascular space burden (PVS) (***E***). ***F***, Number of colocalised genes per genome-wide significant loci containing at least one colocalisation. ***G***, Colocalisation pattern of *TUFM* expression in excitatory neurons with INT in the *rs2008514* locus. ***H***, Colocalisation pattern of *NPIPB6* expression in excitatory neurons with INT in the *rs2008514* locus.

a colocalisation with *EFEMP1* expression specifically in astrocytes, which could potentially reveal insights into their relationship (Figure 5.4 C, D). The highest number of shared genes between two traits was beween intelligence (INT) and subcortical volume (SCV), sharing 7 genes total. Naturally, some of the genes shared between traits were expected due to study design. For instance, several genes colocalised both for the AD GWAS and the "common factor" NDD summary statistics, including *BIN1*, *PICALM*, *ERCC3*, *CR1*, *CR2*, *CLU*, *SLC24A4* and *USP6NL*. The high number of overlapping colocalised genes is expected given that the NDD summary statistics are generated through a common factor model using AD, PD, ALS and LBD GWAS[248]. For the majority of traits with shared colocalisations, only 1-3 genes were shared.

There were also several colocalisations with genes from the same locus. In total, of the 252 unique loci with colocalisations across the 30 traits, 47 colocalised with more than one gene in one or more cell-types. The large majority (39 loci) only included 2 genes (Figure 5.4 F). A notable example of this is the locus containing *CR1* / *CR2* which colocalised in multiple cell-types with AD risk. In another interesting case, the *rs10043984* locus harbored two colocalisations for bipolar disorder (BIP), *FAM53C* / oligodendrocytes (PP.H4 = 0.92) and *WNT8A* / inhibitory neurons (PP.H4 = 0.88). Curiously, this relationship was not present when flipping the cell-type; *FAM53C* had no eQTLs in inhibitory neurons, nor did *WNT8A* have any eQTLs in oligodendrocytes.

One of the most pleiotropic loci was *rs2008514*, where as many as 7 genes were colocalised with intelligence (INT) GWAS in numerous cell-types, including *TUFM* (Figure 5.4 G), *EIF3CL*, *SH2B1*, *CLN3*, *NPIPB6* (Figure 5.4 H), *NPIPB9* and *SGF29*. However, a closer inspection shows an extremely gene-dense region, indicating a potential LD block based on the observed SNP correlation which might bias expression and/or SNP dosage estimates. In addition, several genetic variants are missing in sections of this region; while it could be explained by only including common genetic variants in the eQTL study (MAF > 5%), this pattern was also visible in the original GWAS despite inclusion of SNPs far below this threshold[47]. This complex locus behaviour was also present in the *rs1987471* locus (SCV GWAS, 8 colocalised genes) and *rs4787495* (HV-R GWAS, 5 colocalised genes).

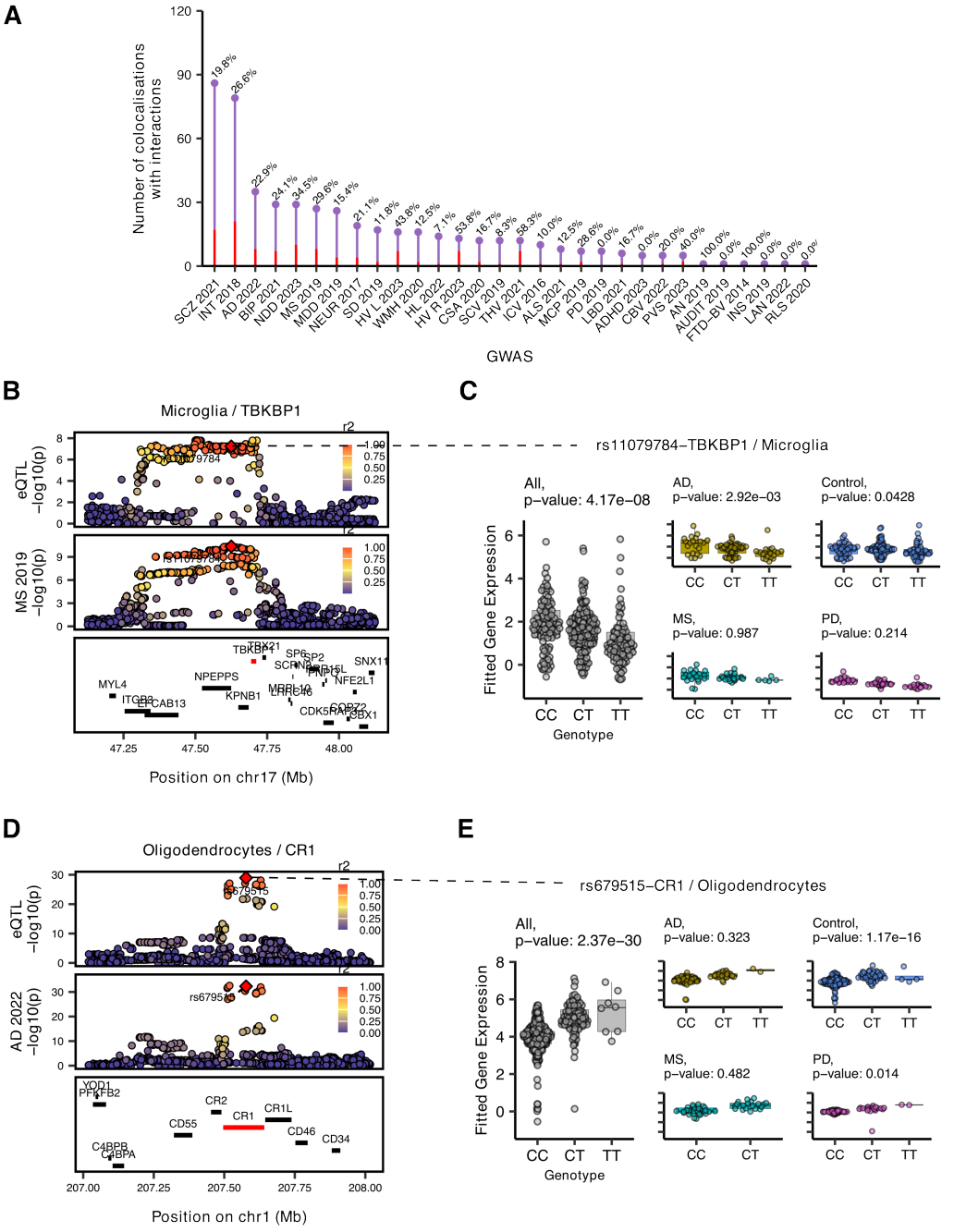


Figure 5.5: Colocalisations with disease interaction. ***A***, Number of colocalisations (y-axis) for each trait, and the proportion of colocalisations where the lead SNP association demonstrates interaction with disease state (demonstrated by the red bar and percentage labels). ***B***, Regional association plot of *TBKBP1* in microglia (top), which colocalises with MS risk (bottom). ***C***, eQTL plot showing the interaction model outputs between the colocalisation lead SNP (*rs1079784*) and *TBKBP1* in microglia. The left, full plot ("all") represents the overall association in the base model (M_1), where x-axis represents genotype and y-axis represents the fitted expression values. The right plots represents the fitted expression values, grouped by diagnosis. The control *p*-value is the association *p*-value when accounting for disease effects, whereas the *p*-value for each diagnosis represents the degree to which the effect differs from the control group (if < 0.05, there is a significant difference in effect on expression to the controls in that group). ***D***, Regional association plot of *CR1* in oligodendrocytes (top), which colocalises with AD risk (bottom). ***E***, eQTL plot showing the interaction model outputs between the colocalisation lead SNP (*rs679515*) and *CR1* in oligodendrocytes.

5.3.5 Impact of disease-interacting eQTLs on colocalisations

Given the proportion of eQTLs discovered to be impacted by the presence of disease samples (as described in Chapter 4), I sought to understand whether this impact extended to colocalisations. To do this, I repeated the interaction testing (Methods) between the lead SNP and each colocalised genes in their respective cell-types. For each lead SNP from a colocalisation, the eQTL association was re-tested with the interaction term between genotype dosage and disease status; however, because this was performed on a small number of genes as compared to the number of eGenes tested for each cell-type, a nominal p-value of the LRT was used here to determine whether an interaction was present. Of the 501 colocalisations, there were 118 (23.6%) where the lead SNP had a nominally significant interaction (LRT p-value < 0.05) with disease status in the eQTL association (Figure 5.5 A) (Appendix 7). For instance, out of the 27 colocalisations found in MS, 8 (29.6%) showed an interaction with disease when testing the lead SNP.

An example of this can be seen in the colocalisation for *TBKP1* in microglia and MS (PP.H4 = 0.87), where interaction with disease testing was tested with the lead SNP *rs11079784* (Figure 5.5 B). The overall association p-value without considering interaction with disease status (the M_1 model) was 4.17×10^{-8} , also passing FDR significance in the MatrixEQTL model (p-value $= 5.74 \times 10^{-8}$, FDR $= 1.23 \times 10^{-4}$). There was a significant preference for the interaction model M_2 based on the LRT outcome (LRT p-value $= 1.59 \times 10^{-2}$). A closer inspection shows that the reference p-value (i.e, for the controls) was comparably much higher (controls p-value $= 4.28 \times 10^{-2}$) than the M_1 and MatrixEQTL p-values, indicating that the association is much weaker when accounting for the presence of disease samples. In addition, there was a significant difference in the effect of the AD samples on the association compared to the controls (AD p-value $= 2.92 \times 10^{-3}$), which seem to be the primary drivers of this association observed in the eQTL discovery and consequently the colocalisation (Figure 5.5 C).

Similarly, the LRT showed a preference for the interaction model M_2 with *rs679515 / CR1* in Oligodendrocytes (LRT p-value $= 5.19 \times 10^{-3}$), which colocalised in AD (PP.H4 = 1.00) (Figure 5.5 D). In contrast to the previous example, there was still an overall strong association in the control samples after accounting for interaction with disease samples (controls p-value $= 1.17 \times 10^{-16}$) albeit weaker than the M_1 model (p-value $= 2.37 \times 10^{-30}$).

In this case, the effect of the PD samples on expression was significantly different than the controls only ($PD_{interaction}$ p-value $= 1.14 \times 10^{-2}$), and may in this case contribute to the overall signal instead of being the primary driver (Figure 5.5 E). It should therefore be possible to replicate this eQTL (and therefore colocalisation) in a control-only cohort. In total, 8 of the colocalisations found in AD were determined to interact with disease status, including *CR2, RASGEF1C, SNX31* and *TP53INP1*.

5.3.6 Differential colocalisations; controls versus full cohort

The presence of disease cases can in some cases contribute to the overall association eQTL association, consequently contributing to colocalisation signal against a number of traits, or in even driving the association entirely in some cases. Whether this results in differences in colocalisation patterns between eQTLs obtained from a cohort restricted to controls and one including disease cases remains unanswered.

To answer this question, I utilised the eQTLs discovered in the controls-only cohort (n = 183, the results of which are described in Chapter 4) and repeated the colocalisation analysis against the same 42 phenotypic traits. In total, I discovered 256 colocalisations representing 134 unique genes in 26 different traits (Appendix 8). Similarly to the colocalisations using the eQTLs from the full dataset, 80.6% of these were in a single cell-type, and the majority of colocalisations were found in the *MAPT* and *HLA* loci. Outside of these loci, the most colocalised gene was *PCCB* (n = 7 colocalisations), appearing in oligodendrocytes and both neuronal classes in SCZ, neuroticism (NEUR) and sleep duration (SD).

Of the 256 colocalisations discovered in the controls-only cohort, 165 (64%) were also present in the full cohort. Curiously, despite the substantial difference in sample size (391 in the full cohort versus 183 in the controls-only cohort) and therefore statistical power, 91 colocalisations were only found in the controls-only cohort representing as many as 61 gene / trait pairs (Figure 5.6 A). For example, there was a colocalisation between *ZEB2* in astrocytes with SCZ (PP.H4 = 0.82) using the control-only cohort eQTLs that was not colocalised using in the full dataset (PP.H4 = 0.05). This was likely driven by the differing strength in association between *ZEB2* expression and the lead SNP (*rs12991836*, MatrixEQTL p-$value_{controls} = 6.78 \times 10^{-5}$, $FDR_{controls} = 0.32$; p-$value_{full} = 0.10$, FDR_{full}=

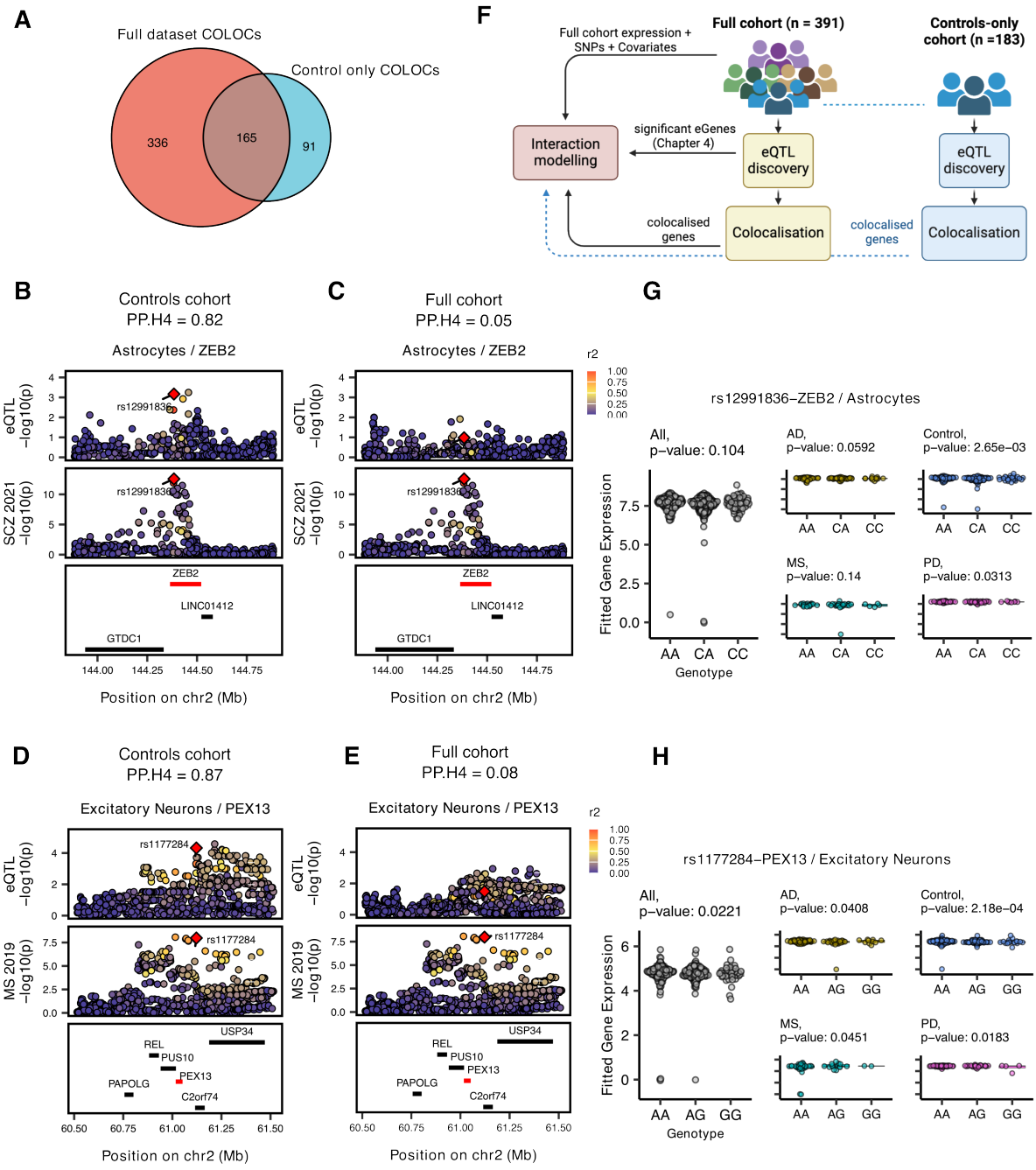


Figure 5.6: Differences in colocalisation patterns between controls and disease. ***A***, Overlap of colocalisations discovered in the full cohort (left, red circle) and those in the controls-only cohort (right, blue circle). ***B***, Colocalisation between *ZEB2* eQTLs in astrocytes with SCZ in the controls-only cohort not present in the full cohort (***C***). ***D***, Colocalisation between *PEX13* eQTLs in excitatory neurons MS not present in the full cohort (***E***). ***F***, Overview of methodology for testing interactions with disease for colocalisations in both the full and controls-only colocalisations. ***G***, Association between *rs12991836* and *ZEB2* in astrocytes with a significant disease interaction when tested in the full cohort. ***H***, Association between *rs1177284* and *PEX13* in excitatory neurons with a significant disease interaction when tested in the full cohort.

0.85). This particular colocalisation also exemplifies that eQTL associations without FDR significance can still colocalise.

Again, using the controls-only eQTLs I discovered a colocalisation for *PEX13* in excitatory neurons and MS (PP.H4 = 0.87) that was not colocalised using the eQTLs in the full dataset (PP.H4 = 0.08) (Figure 5.6 B, C). A closer inspection of the association driving the colocalisation between *rs1177284* (the lead colocalised SNP) and *PEX13* expression explains it; in the full cohort, the association is weak (MatrixEQTL *p*-value = 0.03 , FDR = 0.55) in comparison to the association found in the control-only cohort (MatrixEQTL *p*-value = 4.74×10^{-5} , FDR = 0.02).

The question then becomes why these differences in colocalisation patterns occur between eQTLs from a control-only cohort versus one that includes disease samples, aside from pure statistical power. Given that a large proportion of colocalisations were impacted by the presence of disease samples (23.6%), it might be possible for the presence of disease to obscure associations and lead to a non-colocalisation which would otherwise be present using control tissue. I therefore repeated the interaction analysis using the full dataset, restricting the analysis to genes that only appeared in the control cohort (summarised in Figure 5.6 F). Out of the 91 colocalisations that only appeared in the control cohort, 53 (58.2%) harbored an interaction (meaning the interaction model was favoured over the base model) when tested in the full cohort.

In the example of *ZEB2*, the interaction testing showed a preference for the interaction model over the base model (LRT *p*-value = 0.02) a clear non-significant association in the base model (M_1 *p*-value = 0.10) (Figure 5.6 G). However, in the interaction modelling, the association in the reference group (the controls) is significant (M_2 controls *p*-value = 2.65×10^{-3}), meaning that the presence of disease samples obscure the overall association in the full cohort, leading to a non-colocalisation.

Similarly, testing the association between *rs1177284* and *PEX13* in excitatory also showed a preference for the interaction model (LRT *p*-value = 3.66×10^{-3}), where the base model showed a weak overall association (M_1 *p*-value = 0.02) which is much stronger when accounting for disease effects in the interaction model (M_2 controls *p*-value = 2.18×10^{-4}) (Figure 5.6 H). In addition, the *p*-values for each diagnosis indicate that they each have a

significantly different effect on expression compared to the controls ($AD_{interaction}$ p-value = 4.08×10^{-2}, $MS_{interaction}$ p-value = 4.51×10^{-2} , $PD_{interaction}$ p-value = 1.83×10^{-2}).

5.3.7 Interaction COLOCs not present in controls

On the opposite side of the colocalisations that are present in controls but not in the full cohort, are colocalisations that are only found in the full dataset. One likely explanation for this is that the full dataset is larger and better powered statistically for eQTL discoveries, hence why more colocalisations are found. However, it is also possible in contrast to the results of the previous section that there are colocalisations in the full cohort that are exclusively driven by the presence of disease samples, and would therefore not be visible in the controls no matter the statistical power.

To identify these colocalisations, I analysed the 118 colocalisations in the full cohort which were indicative of disease interaction, based on interaction modelling of the lead SNP (and an LRT p-value < 0.05 to indicate a better fit for this interaction model compared to the base model). Of these 118 colocalisations present in the full dataset, 65 were not present in the controls-only cohort. For instance, *TP53INP1* expression in oligodendrocytes colocalised with AD (PP.H4 = 0.91) in the full dataset but not in controls (PP.H4 = 0.04) (Figure 5.7 A, B). This is explained by significant association between *rs4582532* and *TP53INP1* (MatrixEQTL p-value = 3.14×10^{-6}, FDR = 5.00×10^{-3}) in the full dataset, but not in the controls (MatrixEQTL p-value = 0.58, FDR = 0.99). Further, interaction modelling of the association on the lead SNP *rs4582532* showed a significant preference for the interaction model over the base model (LRT p-value = 0.01) - in this scenario, the interaction model reveals that when accounting for the effects of disease groups on the association there is a non-significant association in controls (controls p-value = 0.32). Finally, there is a significant difference of the effect from the PD samples compared to the controls ($PD_{interaction}$ p-value = 8.83×10^{-3}), suggesting that this association (and consequently the colocalisation) is driven by the PD cases which offers an explanation as to why this colocalisation is not present in the controls-only cohort (Figure 5.7 C).

This situation is repeated in *RAB38*, which colocalises in excitatory neurons with FTD-BV. Again, there is a colocalisation in the full dataset (PP.H4 = 0.81) but not in the controls

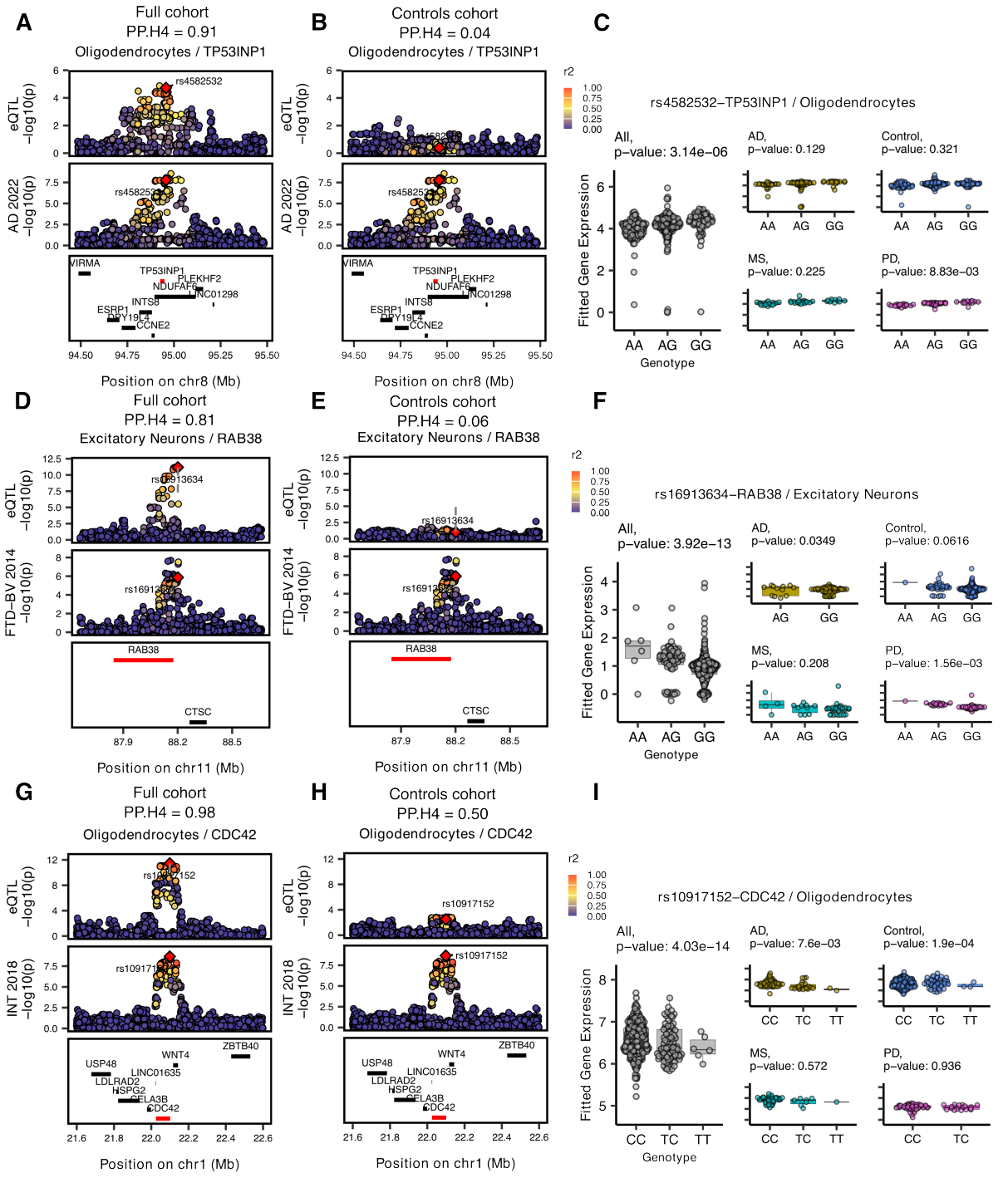


Figure 5.7: Colocalisations driven by disease. ***A***, Colocalisation between *TP53INP1* eQTLs in oligodendrocytes with AD in the full cohort, not present in the controls-only cohort (***B***). ***C***, Interaction eQTL between *rs4582532* and *TP53INP1* in oligodendrocytes. ***D***, Colocalisation between *RAB38* eQTLs in excitatory neurons with FTD-BV in the full cohort, not present in the controls-only cohort (***E***). ***F***, Interaction eQTL between *rs16913634* and *RAB38* in excitatory neurons. ***G***, Colocalisation between *CDC42* eQTLs in oligodendrocytes with INT in the full cohort, not present in the controls-only cohort (***H***). ***I***, Interaction eQTL between *rs10917152* and *CDC42* in oligodendrocytes.

(PP.H4 = 0.06) (Figure 5.7 D, E). The lead association in the region for the full dataset is strong (MatrixEQTL p-value = 6.63×10^{-12}, FDR = 4.99×10^{-9}) compared to the controls (MatrixEQTL p-value = 0.43, FDR = 0.96). Modelling the interaction between *rs16913634* and *RAB38* in the full dataset shows a preference for the interaction model (LRT p-value = 0.01), and shows that the disease-adjusted association in controls is non-significant (controls p-value = 0.06). Just as the previous example, this association is likely driven by PD cases ($PD_{interaction}$ p-value = 1.56×10^{-3}) (Figure 5.7 F).

There are however also examples of disease-driven colocalisations that still retain a significant effect in the interaction testing when accounting for the effect of disease, but yet do not appear in the control-only cohort. For example, *CDC42* in oligodendrocytes colocalises with INT (PP.H4 = 0.98, lead eQTL p-value = 3.53×10^{-12}, FDR = 3.74×10^{-9}) in the full dataset but not in the controls-only cohort (PP.H4 = 0.50, lead eQTL MatrixEQTL p-value = 0.03, FDR = 0.45) (Figure 5.7 G, H). Modelling the interaction with disease cases between the lead SNP *rs10917152* and *CDC42* expression in oligodendrocytes shows a significant preference for the interaction model (LRT p-value = 7.08×10^{-3}) over the base model, being largely driven by the presence of AD cases ($AD_{interaction}$ p-value = 7.06×10^{-3}). However, there is an overall association retained in the control group when accounting for disease effect (M_2 controls p-value = 1.9×10^{-4}), that is much stronger than the eQTL association found when exclusively using the controls-only cohort (Figure 5.7 I). This suggests that there are other factors at play that are not modelled here, leading to a difference in the association discovered in the controls-only cohort compared to the M_2 controls association (accounting for disease status).

5.4 Discussion

In this chapter, I show for the first time genetic colocalisations using cell-type specific eQTLs across the neurological, neuropsychiatric, behavioural and structural brain phenome with a strict posterior probability threshold (PP.H4 > 0.8). I show that in large part, these colocalisations are cell-type specific (77.9 - 80.6% across tested cohorts) and in most regions concern a single gene with the trait in question (81.3% of regions in these results only had a single colocalised gene). In addition, I show for the first time that colocalisations can also be

dynamic as a result of investigating state-specific eQTLs. In particular, the findings described here also show the value of restricting analyses to samples free from presence of disease pathology; despite more than half the size, the use of a controls-only cohort allowed the discovery of 91 colocalisations not present in the full dataset free from confounding disease effects. These results provide an important resource for target prioritisation in disease risk, and an understanding on the cell-types behind gene-trait relationships in the brain.

5.4.1 Choice of GWAS

There is an ever-growing resource of publicly available GWAS summary statistics with multiple public portals for download, such as the NHGRI-EBI GWAS catalogue (https://www.ebi.ac.uk/gwas/), the openGWAS database (https://gwas.mrcieu.ac.uk/).

In this chapter, I have opted for a carefully curated dataset of 42 traits instead of using up to hundreds or even thousands of traits. One of the main reasons for this is because only a few number of GWAS traits in the brain have a large number of loci to test against and as shown in Figure 5.2 B, the number of colocalisations discovered is highly correlated with the number of loci tested. As such, increasing the number of traits tested may become counter-productive as well as unnecessarily increasing the need for computational resources.

This approach is however entirely feasible and has been done in the past; in 2020, Pividori et al., published their *PhenomeXcan* resource, where they applied a novel colocalisation method to prioritize candidate gene targets in 4,091 GWAS traits with GTEx eQTL data in 49 tissues[249]. Similarly, Rocheleau et al., in 2022 performed a phenome-wide colocalisation approach, using the eQTLs from 48 tissues as published by GTEx against 1,411 phenotypes collated in the UK Biobank, discovering more than 9,100 colocalised genes which they used to create a tissue-level phenome-wide network map[250]. In addition, brain-specific resources have also been created recently, such as the brain catalog which claims to host 500 GWAS summary statistics for brain-related traits as well 58 molecular QTL datasets[251] but lacks the resolution (i.e cell-type context) shown in this chapter.

5.4.2 Cell-type specificity of colocalised genes

A key finding of this work, echoing the work of Bryois et al.[108] is the degree to which the majority of colocalisations are cell-type specific (~80%). This is particularly interesting considering that the majority of eQTLs are shared across cell-types (~40% are unique to a cell-type). Given that the GWAS used here are much better powered statistically than this comparably small eQTL dataset, it is likely they pinpoint a group of strongly associated SNPs of which one or multiple are presumably causal for the trait. One potential reason for the cell-type specificity is that these genetic variants will in the large majority of cases cluster in or around a single gene target, and in turn overlap cell-type specific regulatory elements such as enhancers, promoters and transcription factors as well as cell-type specific patterns of chromatin accessibility.

Another potential reason which is more technical is that some of the cell-types used in this study are simply underpowered for eQTL discovery and consequently colocalisation, and that colocalisations which are cell-type specific here are not "truly" cell-type specific. This is evidenced by the fact that the number of colocalisations discovered here by cell-type followed almost the same order as the number of eGenes discovered for each cell-type (with the exception of astrocytes in more colocalisation were found than oligodendrocytes). Sequencing a larger number of cells per individual will not only allow a more robust confirmation of the cell-type specificity of these colocalisations, but also potentially uncover more colocalisations shared across cell-types as well as the ability to map rarer cell-types.

One unavoidable problem of these analyses is the setting of pre-defined thresholds. Using a PP.H4 above 0.8 allows a strong degree of confidence when reporting colocalisations, but this threshold does not preclude marginal colocalisation (0.5-0.8) from further investigations. It is possible that there are shared colocalisations between cell-types that are simply not considered here because of this cut-off (one cell-type for a gene may be in the 0.5 - 0.8 range whereas the other may be well above 0.8). Carefully exploring more relaxed posterior probability thresholds may shed some light on shared signals across cell-types.

There were however a smaller proportion of colocalised genes which were shared between more than one cell-type (23.1%), such as *PCCB* which was colocalised in SCZ, SD and INT within both neuronal classes as well as oligodendrocytes. In this chapter, the overlaps have been

determined by a straight intersection; it is worth noting however, that more elaborate methods allows proper statistical inference of shared colocalisations. For instance, *HyPrColoc* could be well suited here testing all genes in a region in parallel (or all cell-types for one gene) as it is designed to handle a large number of traits simultaneously[139]. Multi-trait colocalisation could uncover statistically supported pleiotropic pathways and allow a more nuanced target prioritization and drug design approach.

5.4.3 Disease-interacting colocalisations and use of control tissue

Another important finding of this chapter is the degree to which eQTLs influenced by the presence of disease samples (presumably leading to differences in expression) also influence genetic colocalisations. These results show that nearly a quarter of colocalisations discovered here have some interaction with disease when considering the lead SNP. There are however several factors to consider when evaluating these observations.

One is that despite the large proportion of disease-interacting colocalisations, the measured interactions retain significance on the controls (the reference group) when discounting the disease effect. This means that despite impact of disease on the association, it is likely that even without use of disease samples these colocalisations would have been discovered. Indeed many of the disease-interacting colocalisations (53 out of 118) were also found in the colocalisations using the control-only cohort, meaning that not all of the disease-interacting colocalisations are completely disease-driven.

On the other side of this however is the interesting finding of colocalisations that only appear in the controls-only cohort (n = 91), despite the substantial decrease in statistical power. The natural question to ask here is whether the associations found in the controls-only cohort not found in the full cohort is due to the disease samples obscuring the association, where disease-influenced expression estimates work against or weaken the overall association found in controls. Testing whether this was the case, I found that 53 (58.2%) of these 91 colocalisations were significantly better modelled in the interaction model in the full cohort, suggesting that indeed the presence of disease samples can obscure associations and consequently colocalisations. However, this begs the question as to why the other colocalisations uniquely found in the controls-only cohort were not found in the full cohort,

given that these were not found to interact with disease. As always, pre-defined cut-offs (in this case an LRT p-value <0.05, and a PP.H4 > 0.8) may themselves obscure the whole picture; it is likely that some colocalisation scores will in reality be much closer together across the controls-only and full cohorts than what a strict cut-off shows.

It is also important to consider that as described in this chapter, only the lead SNP from each colocalisation was tested; it is unclear how the combined effect of disease on eQTL associations (across multiple correlated SNPs for instance, such as with the example of *rs11079784* and *TBKBP1*) affects the colocalisation signal as a whole. Naturally, other concerns that are present in the analysis of disease-interaction eQTLs as described in Chapter 4 also carry over in this analysis, such as the use of PCs to correct across multiple disease states. In an optimal setting, a comparison of colocalisations using eQTLs derived from control-only tissue versus a disease-specific tissue (such as exclusively AD) that are similarly powered might uncover truly state-specific colocalisations.

Chapter 6

Mendelian Randomisation at single-cell level in the post-mortem brain

6.1 Preface

6.1.1 Appendices

The appendices referenced in this chapter are Appendices 9-11 organised as follows;

- Appendix 9: single-cell MR results
- Appendix 10: pQTL MR results (UKB-PPP)
- Appendix 11: pQTL MR results (Robins et al.,)

6.2 Introduction

MR has in recent years enjoyed a growing popularity with applications in a variety of fields. In its most basic form, MR can be applied using publicly available summary statistics and in a two-sample setting has allowed to infer directional relationships between thousands of traits[252]. Recently this approach has also been applied to eQTLs in the brain[90] as well as tissue-wide with the use of GTEx[253]. It has also been a popular method for pQTLs such as assayed in brain[163], plasma to uncover COVID-19 targets[174] and cardiovascular risk targets[75] as well as

cancer[254]. As of yet, only one study has applied MR at single-cell level. Yazar et al., as mentioned previously, applied summary-based mendelian randomization (SMR) on cell-type specific plasma eQTLs to uncover putatively causal associations in blood-related traits[107].

To translate genome-wide significant loci into potential therapeutic targets for drug development, the direction of effect informs whether an antagonistic or agonistic therapeutic strategy is appropriate; in particular, use of single-cell eQTLs also provide with an important cellular context and pave the way for more targeted therapy. To date, this approach has not yet been applied to eQTLs derived from single-cell expression in the brain, and promises to identify novel targets with therapeutic directionality.

6.2.1 Aims and objectives

- Infer putatively causal relationships in the first MR study using single-cell brain eQTLs.
- Validate MR hits using independent cross-tissue datasets to strengthen inferences.
- Investigate the epigenetic landscape of instrumental variables and infer putative mechanisms of action.

6.3 Results

6.3.1 Instrumental variable (IV) selection

To infer putatively causal relationships through MR between cell-type specific eQTLs and brain traits, control tissue must be used to avoid the potential for reverse causality through potential confounding as a result of disease presence. In addition, MR requires careful selection of genetic variants as instrumental variables (IVs) to infer these relationships. To meet these criteria, I performed IV selection using the eQTLs from the controls-only cohort, summarized in Figure 6.1 A. Briefly, eQTL discovery was performed on the controls-only cohort (Chapter 4) followed by genetic colocalisation using COLOC (Chapter 5). Colocalisation helps in the selection of IVs by prioritizing instruments that are directly relevant to the trait; they are both significantly associated to the exposure (the expression of the gene, determined by FDR) and to the trait through a positive colocalisation signal (high

PP.H4). The assumption is then that the association of the IVs to the trait is through the exposure which in this case is the gene expression.

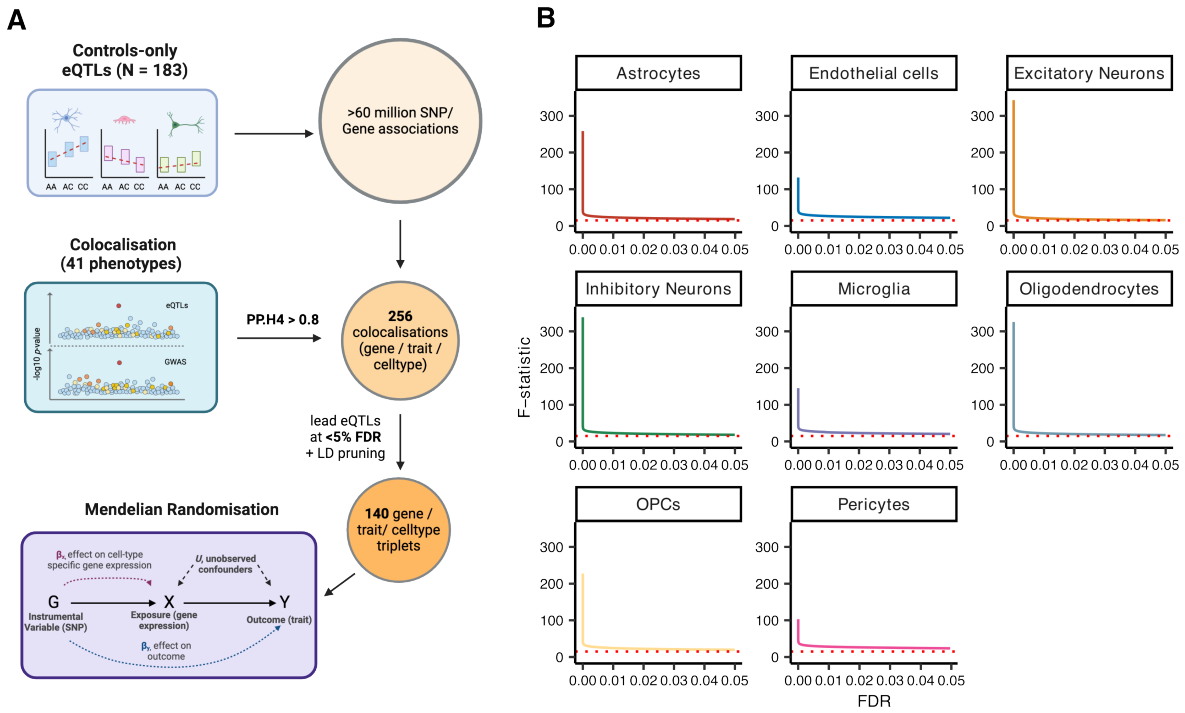


Figure 6.1: Instrumental variable selection. ***A***, Overview of instrumental variable (IV) selection. Using the full set of eQTLs discovered in the controls-only cohort (N = 183), colocalisation was performed with 41 brain-related phenotypic traits. Colocalisations with a PP.H4 > 0.8 were considered significant, and kept for IV selection, where the entire 1Mb of a successfully colocalised region was taken forward. Regions with no eQTLs < 5% were removed. ***B***, Distribution of *F*-statistic for FDR significant eQTLs. Each plot represents a cell-type, with the x-axis representing the FDR compared to the corresponding *F*-statistic (y-axis).

Using the controls-only dataset, there were 256 colocalisations across 26 unique phenotypic traits (out of the 41 traits tested), representing 134 unique genes. To select IVs, the lead SNP from each colocalisation window (1Mb) was selected and filtered based first on its eQTL association strength (< 5% FDR) after which all correlated SNPs were removed ($r^2 > 0.01$, full details in methods, Chapter 2). In addition, age-related eQTLs (as described in Chapter 4) discovered in the controls-only cohort which overlapped with disease-interaction eQTLs in the full cohort were removed, given the possibility of driving unobserved disease. In total, I retained 140 IVs for as many gene / cell-type / trait combinations (triplets) to test in a MR framework. In the majority of cases (135 out of 140, 96.4%) only a single IV was retained. I also confirmed that all selected instruments had an *F*-statistic measure above 15 to avoid weak instrument bias[156]. This measure is important because it captures the ratio of the

variance explained by the model versus variance outside of it, where a high value suggests the model adequately captures the variance around the association and therefore is a measure of a strong association between SNP and gene. Across the entire eQTL dataset in all cell-types, associations below an FDR < 5% all had an F-statistic > 15 (Figure 6.1 B).

6.3.2 Single-cell mendelian randomisation

Given that the majority of triplets had few IVs, MR was applied using the Inverse Variance Weighting (IVW) method (which for single IVs defaults to the Wald Ratio as described in the introduction). Out of all 140 tests conducted, all were significant (IVW p-value < 0.05) (Figure 6.2 A) (Appendix 9) representing 94 unique genes in 22 different phenotypic traits. The majority of genes appeared in a single cell-type (77 out of 94, 81.9%). Following the previous trend in the colocalisation results, the majority of hits were in SCZ (n = 31), followed by INT (n = 21) and MS (n = 11). Several hits were found even for more underpowered traits such as ALS (n = 2), ADHD (n = 2) and AUDIT (n = 1). In addition, many inferences were made in relation to brain volume and structure including thalamus volume (THV, n = 1), hippocampal volume (left L, n = 4 and right R, n = 2), subcortical volume (SCV, n = 2), cerebellum volume (CBV, n = 3) and cortical surface area (CSA, n = 4).

MR allows inferences of effect size and directionality, as exemplified with AD (Figure 6.2 B). For instance, the association between *EGFR* and astrocytes in AD (IVW p-value = 2.35×10^{-9}, IVW beta = 0.14) implies that increase in *EGFR* expression confers an increase in AD risk. Both *JAZF1* (IVW p-value = 2.02×10^{-8}, IVW beta = -0.07) and *RASGEF1C* (IVW p-value = 5.40×10^{-12}, IVW beta = -0.10) expression in microglia show protective effects in relation to AD risk, while *CR1* expression has the opposite effect while acting in two cell-types, OPCs (IVW p-value = 1.99×10^{-33}, IVW beta = 0.12) and oligodendrocytes (IVW p-value = 1.99×10^{-3}, IVW beta = 0.14).

To exemplify the value of cell-type specific MR, the top genes for each trait (by IVW p-value) and their effect sizes are shown (Figure 6.2 C). Several of these genes have previous evidence to the trait in question except here for the first time with cell-type context. For example, expression of *NEGR1* in oligodendrocytes was found to decrease risk of major depressive disorder (MDD) (IVW p-value = 4.75×10^{-18}, IVW beta = -0.09). It has been previously

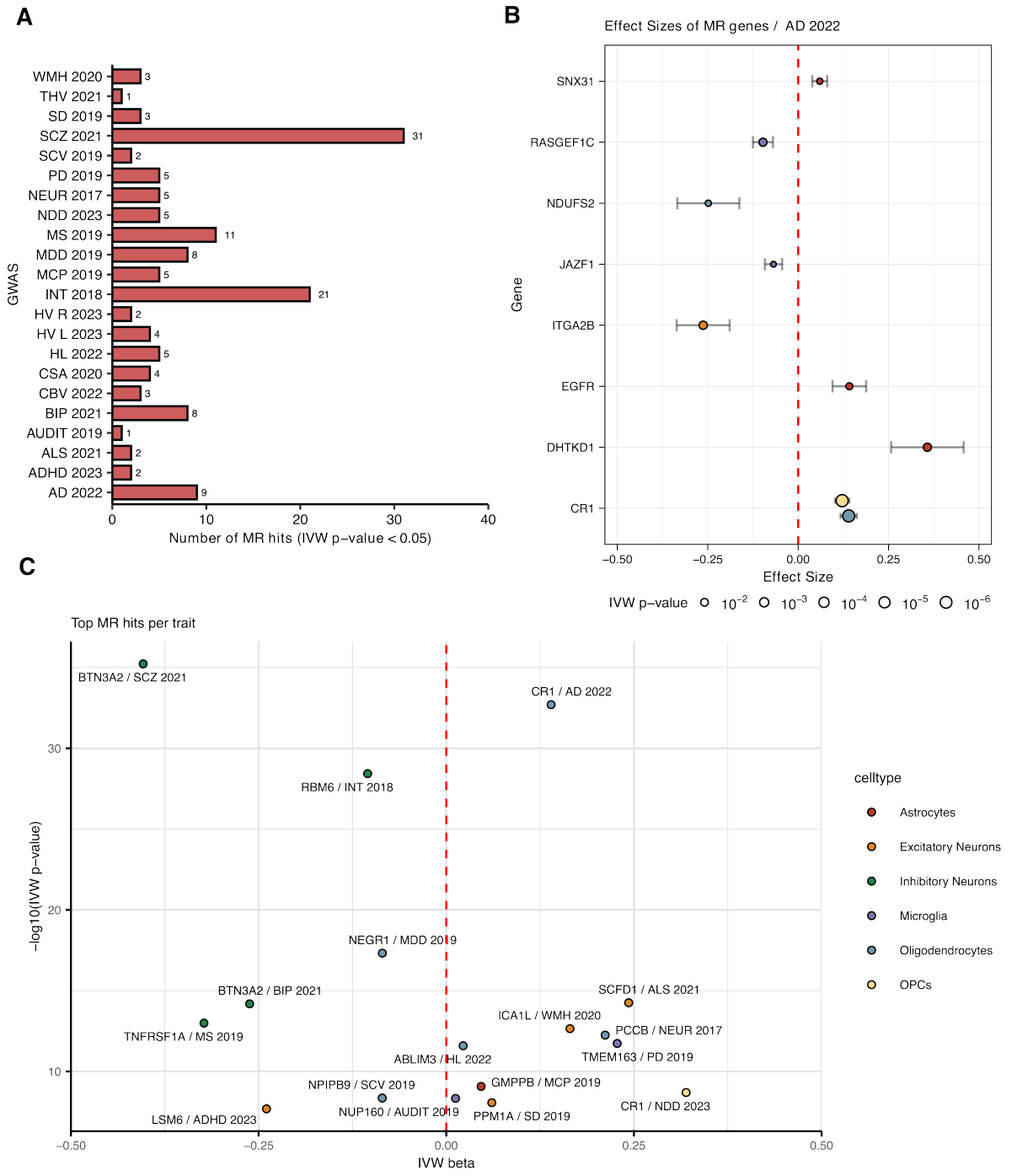


Figure 6.2: Overview of MR results. ***A***, Number of MR inferences (IVW p-value < 0.05) per trait. Each "hit" represents a cell-type / gene pair. ***B***, MR inferences for AD. The IVW effect size is shown for each gene (x-axis), coloured by the cell-type in which the inference was made. Size of each point is scaled by -log10(IVW p-value). A positive effect size indicates that increased expression increases AD risk, whereas a negative effect size indicates that increased expression decreases AD risk. ***C***, Top MR hits for each trait based on p-value (y-axis), showing their respective effect sizes (x-axis). Traits with extremely large betas (CSA, THV, HV) are excluded from this figure.

indentified as a gene of interest in an independent causal inference study using the MDD GWAS and GTEx eQTL data[255], as well as linked to several other psychiatric disorders (including MDD) through a behavioural study of knock-out mice[256]. The expression of *TMEM163* in microglia confers risk towards PD (IVW p-value = 1.85×10^{-12}, IVW beta = 0.23), and was recently labelled as a candidate disease gene in a study integrating brain-specific epigenetic data with the original GWAS[137]. Other genes such as *CR1* , *SCFD1* (ALS)[257] , *ICA1L*[258] (WMH) have also been implicated in their respective traits previously. On the other hand, several of the top genes have a previously unknown role to their respective traits. For instance, *PPM1A* expression in excitatory neurons was linked with lower sleep duration (IVW p-value = 8.49×10^{-9}, IVW beta = 0.06); *LSM6* expression, also in excitatory neurons, was found to increase ADHD risk (IVW p-value = 2.06×10^{-8}, IVW beta = -0.24). These few examples showcase the value of finding novel associated genes, their direction of effect in each respective trait as well as the cell-type in which they act and provide an important resource for therapeutic targeting.

A proportion of these genes also appeared more than once; of the 94 unique genes found across the results, 17 appeared in more than one cell-type and 13 appeared in more than one trait (Figure 6.3 A). A good example of this is *PCCB*, with MR hits in oligodendrocytes and both neuronal classes across NEUR and SCZ (Figure 6.3 B), as well as SD in the case of inhibitory neurons (not shown). In all three cell-types, *PCCB* expression is linked with a decreased risk in SCZ ($IVW_{oligodendrocytes}$ beta = -0.34, $IVW_{exc.neur}$ beta = -0.50 , $IVW_{inh.neur}$ beta = -0.32) while an increased risk in neuroticism (NEUR, $IVW_{oligodendrocytes}$ beta = 0.21, $IVW_{exc.neur}$ beta = 0.31, $IVW_{inh.neur}$ beta = 0.20). Other examples of genes with opposing effects in different traits include *GMPPB* expression in astrocytes (IVW_{ADHD} beta = -0.14, IVW_{MCP} beta = 0.05) (Figure 6.3 C) and *C1QTNF4* expression in both OPCs (IVW_{NDD} beta = 0.67, IVW_{INT} beta = -0.05) and oligodendrocytes (IVW_{NDD} beta = 0.48, IVW_{INT} beta = -0.03) (Figure 6.3 D).

Similarly, several genes shared the same direction of effect in the same cell-type across traits. In the case of *SLC12A5*, its expression in excitatory neurons was linked through MR to decrease both MS (IVW beta = -0.53) and MDD risk (IVW beta = -0.13) (Figure 6.3 E). Two other genes, *NPIPB6* and *NPIPB9* had opposing effects in different cell-types but for the same two traits; *NPIPB6* expression in excitatory neurons was linked to higher INT and SCV (IVW_{INT} beta = 0.09, IVW_{SCV} beta = 0.14) (Figure 6.3 F) whereas *NPIPB9* expression in

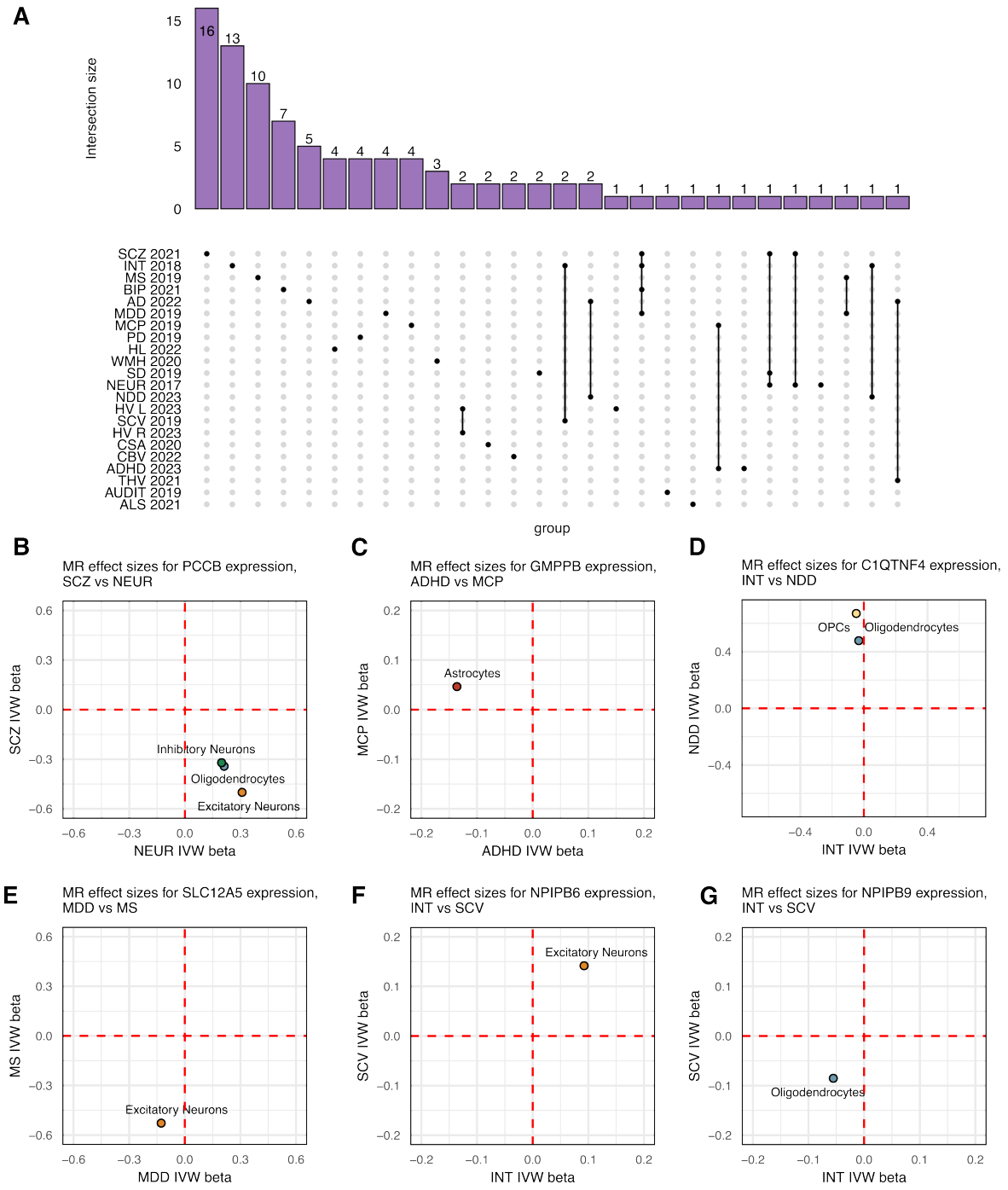


Figure 6.3: Shared MR genes. ***A***, UpSetR plot of the unique genes per trait. For instance, there were 16 genes inferred to be putatively causal in one or more cell-types exclusively for schizophrenia (SCZ). Two genes overlapped between subcortical volume (SCV) and intelligence (INT). ***B,*** Comparison of IVW effect sizes for *PCCB* in SCZ (y-axis) and neuroticism (NEUR, x-axis). ***C***, comparison of IVW effect sizes for *GMPPB* between multisite chronic pain (MCP) and attention deficit hyperactivity disorder (ADHD). ***D***, comparison of IVW effect sizes for *C1QTNF4* between common factor for neurodegenerative diseases (NDD) and INT. ***E***, comparison of IVW effect sizes for *SLC12A5* between multiple sclerosis (MS) and major depressive disorder (MDD). ***F***, comparison of IVW effect sizes for *NPIPB6* between SCV and INT. ***G***, comparison of IVW effect sizes for *NPIPB9* between SCV and INT.

oligodendrocytes had the opposing (albeit smaller) effect (IVW_{INT} beta = -0.06, IVW_{SCV} beta = -0.09) (Figure 6.3 G).

6.3.3 Multi-instrument MR

Careful selection of IVs ensures robust MR estimates and protects them from potential bias. However, given that in this case the majority of MR inferences are done with a single IV (which is common in studies performing MR with pQTL[75,174,259]and eQTL data[90,253]), there is a sensitivity around the selection of a single (or few) variants as IVs. To validate these MR inferences I therefore implemented an additional method termed "PCA-IVW", a method developed by Burgess et al.,[260] that makes use of the entire set of genetic variants in the colocalised region. Briefly, this method decomposes the effect sizes across the region by PCA (weighted by the SNP-SNP correlations) into PCs which are then used as independent instruments for MR using IVW. With this approach, I selected PCs for the colocalised regions of each cell-type / gene / trait triplet that explained 99.9% of the variance in the region. This resulted in an average of 33 IVs (range 10 - 91) per triplet that were then used for IVW-based MR inference. Out of the 140 significant MR hits discovered, 138 (98.6%) were validated with the same direction of effect. In addition, the effect sizes correlated well between both the two MR methods (Pearson $r = 0.97$, p-value $= 7.71 \times 10^{-84}$).

I show here 3 examples that successfully replicate the previous MR inferences; in Figure 6.4 A, the association between *SNX31* expression in astrocytes and AD is shown ($IVW_{singleIV}$ p-value $= 1.46 \times 10^{-8}$, beta = 0.06). Using the effect sizes and the SNP correlation matrix for the entirety of the SNPs in the region (n = 1,710), the effect sizes were decomposed in 66 PCs explaining 99.9% of the variance. Applying the PCA-IVW methodology, there was a significant effect using these PCs as IVs (PCA-IVW p-value $< 2 \times 10^{-16}$, beta = 0.03). In both cases, single IV and PCA-IVW show that *SNX31* expression is associated to increased AD risk. Showcasing an example of the opposite effect size, *FSCN2* expression in oligodendrocytes decreased risk in neuroticism in both single IV MR ($IVW_{singleIV}$ p-value $= 4.45 \times 10^{-11}$, beta = -0.05) and PCA-IVW, where the effect sizes for 2,768 SNPs were decomposed into as many as 91 PCs (PCA-IVW p-value $< 2 \times 10^{-16}$, beta = 0.03) (Figure 6.4 B). While this was the largest number of PCs used compared to other inferences, it was also the region with the largest number of SNPs tested. With SCZ, the results also show that

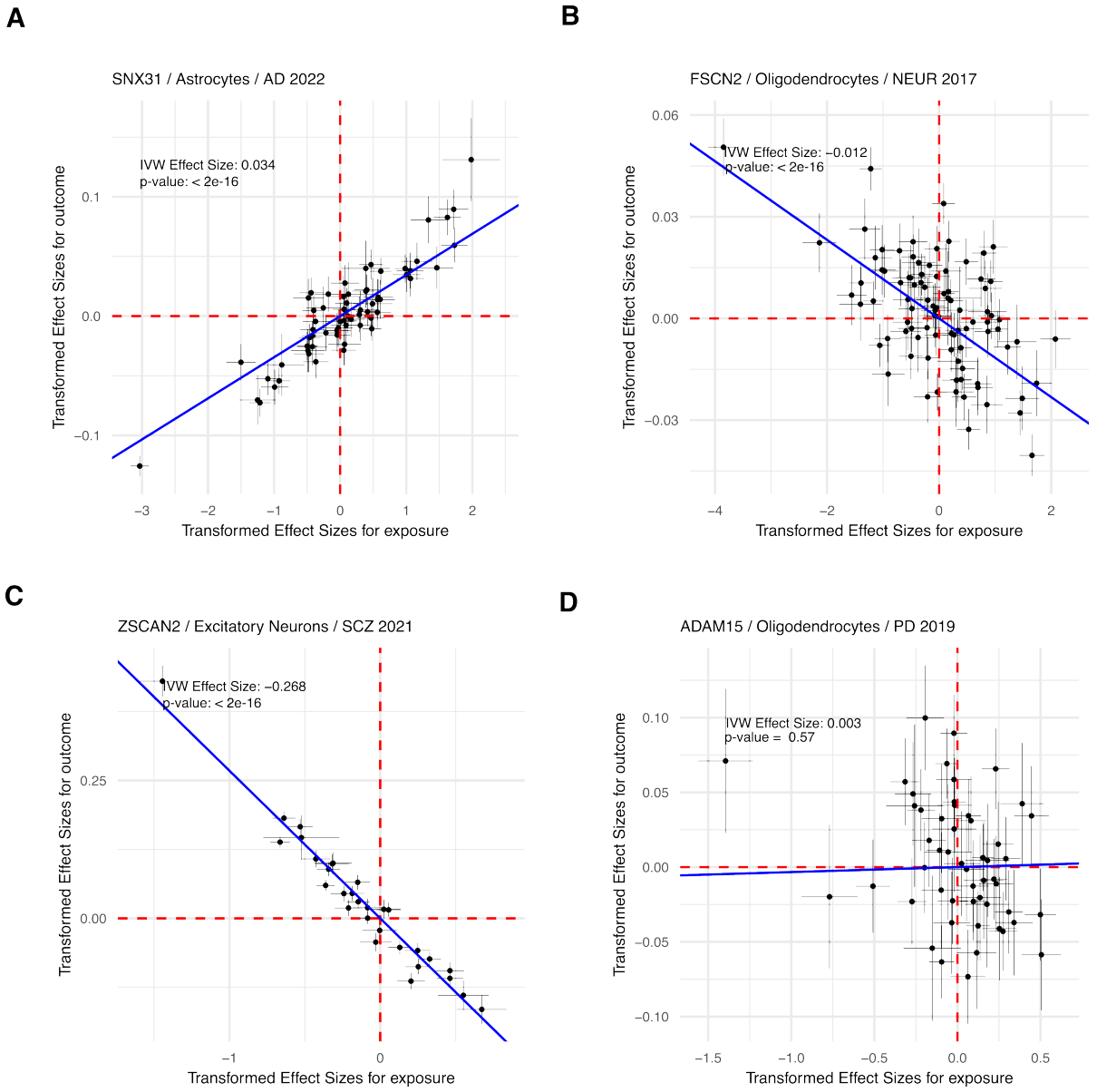


Figure 6.4: Multi-instrument MR using PCA-IVW. ***A***, Example of PCA-IVW method applied to the colocalised region *SNX31* / Astrocytes and AD. Each point represents an IV, obtained through PCA decomposition of effect sizes (y-axis is the GWAS effect size, x-axis is the single-cell eQTL effect size). Bars represent standard errors. ***B***, Example of PCA-IVW data points for *FSCN2* / Oligodendrocytes and NEUR. ***C***, Example of PCA-IVW data points for *ZSCAN2* / excitatory neurons and SCZ. ***D***, Example of PCA-IVW data points for *ADAM15* / oligodendrocytes and PD, showing a non-significant association.

expression of *ZSCAN2* in excitatory neurons was associated with decreased overall risk in both methods ($IVW_{singleIV}$ p-value = 2.37×10^{-12}, beta = -0.29, PCA-IVW p-value < 2×10^{-16}, beta = -0.27) (Figure 6.4 C). The application of this method also showcases the identification of potentially spurious results by single-IV MR; here the association between *ADAM15* expression in oligodendrocytes and PD ($IVW_{singleIV}$ p-value = 5.04×10^{-7}, beta = -0.32) as determined by the PCA-IVW method shows largely competing effects across the region, resulting in an overall non-significant effect opposite to that of the single IV method and two orders of magnitude smaller (PCA-IVW p-value = 0.57, beta = 0.003) (Figure 6.4 D).

6.3.4 Protein QTLs for validation of single-cell MR

The identification of putatively causal inferences between gene expression in a given cell-type and a phenotypic trait is a potentially powerful framework for target identification. For these inferences to be valid targets however, it is expected that the association is replicated at protein level and to this end I sought to replicate these findings using external pQTL datasets. In the first instance, I used the recently published large-scale pQTL study performed on protein levels assayed in circulating blood from UK BioBank participants (N = 54,219), termed the UK BioBank Pharma Proteomics Project (UKB-PPP)[207]. In total, 2,923 proteins were assayed in this dataset, which overlapped with 15 (16.0%) of the 94 unique genes discovered in the single-cell MR inferences, representing 20 gene-trait pairs.

Given the availability of full summary statistics for this dataset, I first independently performed a colocalisation analysis for each of the 20 gene-trait pairs with a relaxed prior probability (methods). Of these, 16 (80%) gene-trait pairs were replicated by colocalisation (PP.H4 > 0.8). Employing the same method as earlier on IV selection (<5% FDR and retention of independent SNPs, $r^2 < 0.01$), I retained between 1-9 IVs per test (median 7) to perform MR and replicated 11 (55%) of the 20 gene-trait pairs available (Figure 6.5 A, E) (Appendix 10). For instance, *CR1* expression in oligodendrocytes (and OPCs) is putatively causal for AD risk as shown earlier (IVW p-value = 1.99×10^{-3}, IVW beta = 0.14). Colocalisation using the UKB-PPP dataset shows a strong signal (PP.H4 = 1.00) (Figure 6.5 B), replicating the cell-type specific colocalisation (PP.H4 = 1.00) (Figure 6.5 C). The MR inference also shows a significant effect in the same direction ($IVW_{UKB\text{-}PPP}$ p-value = 0.02,

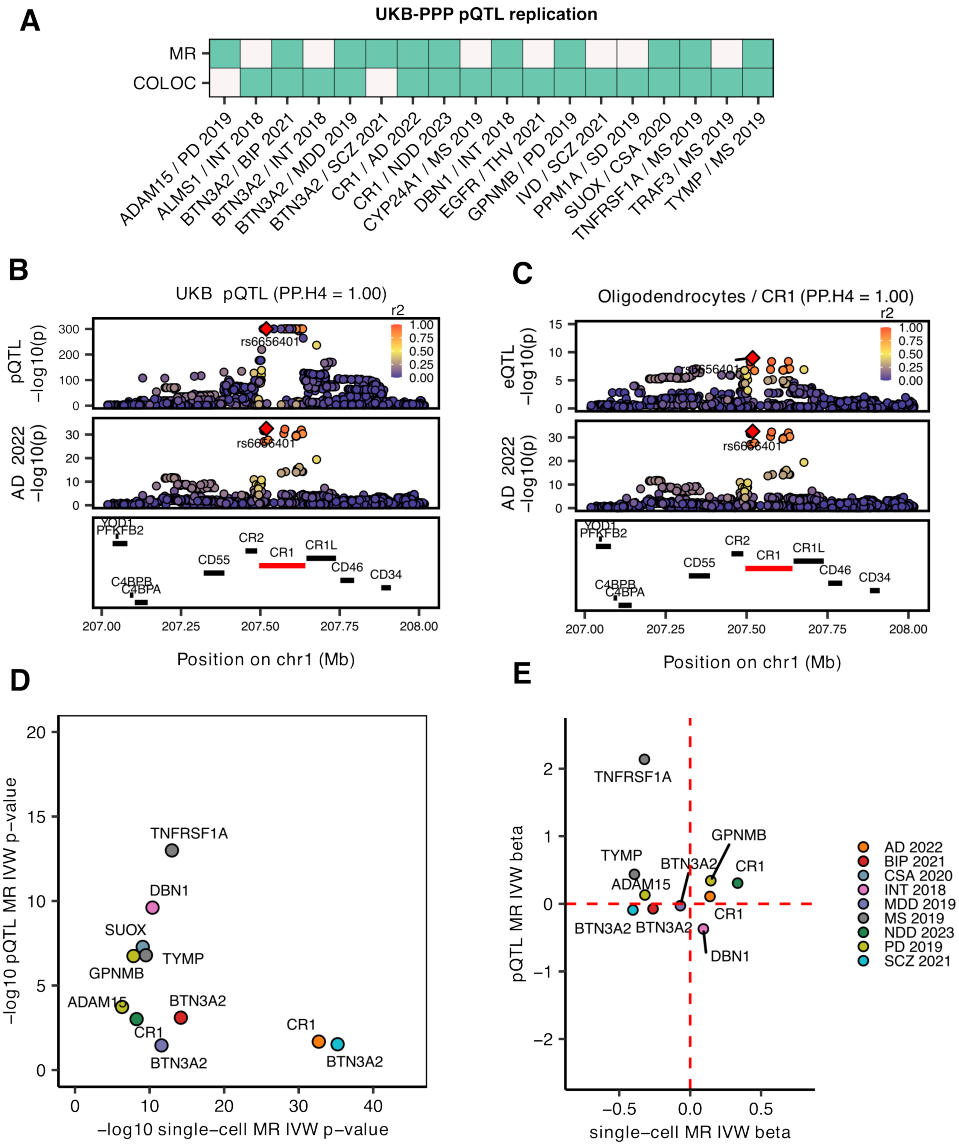


Figure 6.5: Validation of MR inferences using UKB-PPP. ***A***, Overview of UKB-PPP inferences validated by COLOC (PP.H4 > 0.8) and MR (IVW p-value < 0.05), indicated by colored vs non-colored tiles (non-significant). ***B***, Colocalisation between *CR1* pQTLs in UKB-PPP and AD. ***C***, Colocalisation between *CR1* eQTLs in oligodendrocytes and AD. ***D***, Significant MR inferences (IVW p-value < 0.05) using both single-cell eQTLs (x-axis) and UKB-PPP pQTLs (y-axis). Points colored by trait as per panel *E*. ***E***, Comparison of effect sizes for replicated MR inferences using single-cell eQTLs (x-axis) and UKB-PPP pQTLs (y-axis) IVW effect sizes.

beta = 0.11) (Figure 6.5 D), suggesting a potentially useful biomarker for AD given its presence in circulating blood.

For the most part, the MR inferences using the UKB-PPP data reflected the same direction of effects as in the single-cell MR (Figure 6.5 E); *GPNMB* expression for example in OPCs is putatively causal for increased PD risk ($IVW_{\text{single-cellMR}}$ p-value = 1.47×10^{-8}, beta = 0.14) as well as at protein level in circulating blood ($IVW_{\text{UKB-PPP}}$ p-value = 1.80×10^{-7}, beta = 0.34). In other cases, the opposite effect was seen such as for *TYMP* / MS, showing that an increase in expression in excitatory neurons is protective ($IVW_{\text{single-cellMR}}$ p-value = 3.15×10^{-10}, beta = -0.39) but increased protein levels in circulating blood has the opposite effect, putatively causal for increased MS risk ($IVW_{\text{UKB-PPP}}$ p-value = 1.62×10^{-7}, beta = 2.13). This was also the case for *ADAM15* / PD, *TNFRSF1A* / MS , *DBN1* / INT and *SUOX* / cortical surface area (CSA, not shown due to large effect sizes), where the MR inferences had opposite effect sizes at protein versus single-cell expression level. These examples uncover potentially complex pleiotropic pathways that differ between tissues.

Because of the sparsity of assayed proteins in UKB-PPP, I additionally extended this replication to another pQTL study performed in brain by Robins et al[206]. This study while being similarly powered to the single-cell eQTL dataset (N = 330 aged individuals) assayed a much larger number of proteins (n = 7,376, 152.3% increase from UKB-PPP). However, it also made use of both control and disease tissue and as such is of limited replicability because of potential dynamic pQTL effects (as described in Chapter 4). Of the 94 unique genes discovered in the single-cell MR results, 44 (46.8%) had were also assayed in the Robins study representing 50 gene-trait pairs. In contrast to UKB-PPP, the summary statistics available for this study were filtered for independent variants and as such precluded colocalisation analysis.

After IV selection as above, 22 out of the 50 gene-trait pairs were retained for use in MR, the majority of which were excluded due to lack of significance (> 5% FDR). However, of the 22 gene-trait pairs tested, 20 (90.9%) had a significant MR association p-value by IVW (IVW p-value < 0.05) (Figure 6.6 A) (Appendix 11), the majority of which (18, 80%) had the same direction of effect (Figure 6.6 B). For instance, *SCFD1* expression in excitatory neurons was putatively causal for increased ALS risk ($IVW_{\text{single-cellMR}}$ p-value = 5.56×10^{-15}, beta = 0.24) as well as at bulk protein level ($IVW_{\text{RobinspQTL}}$ p-value = 1.14×10^{-13}, beta = 1.61). In

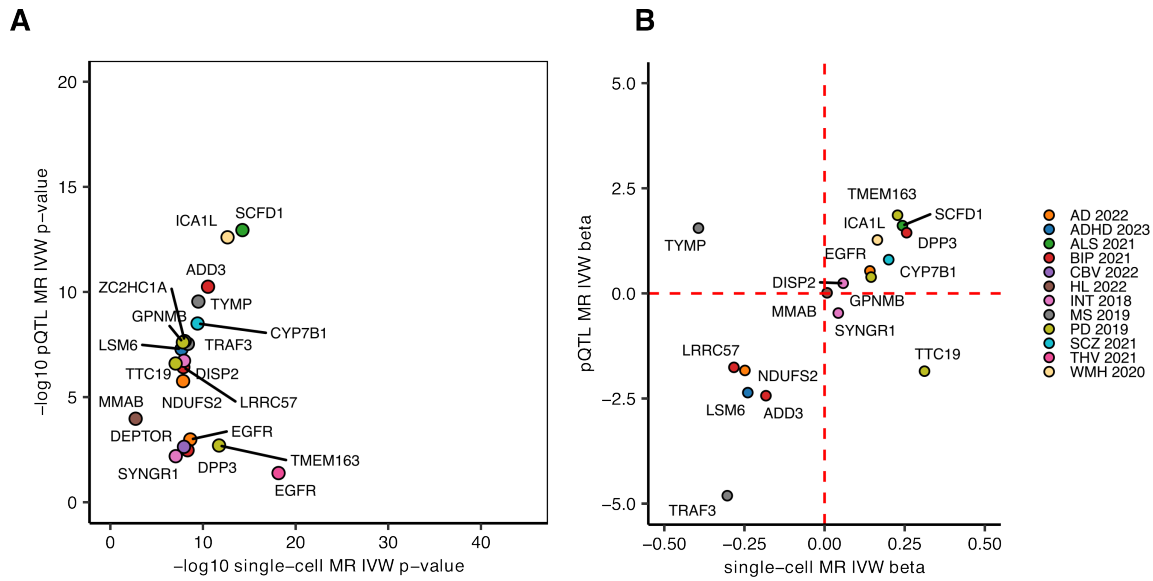


Figure 6.6: Validation of MR inferences using Robins et al., brain proteomics. ***A***, Significant MR inferences (IVW p-value < 0.05) using both single-cell eQTLs (x-axis) and Robins et al.,[206] pQTLs (y-axis). Points colored by trait as per panel *B*. ***B***, Comparison of effect sizes for replicated MR inferences using single-cell eQTLs (x-axis) and Robins et al., pQTLs (y-axis) IVW effect sizes.

addition, *GPNMB* expression in OPCs (also validated by UKB-PPP analyses) was also replicated at bulk level with the same direction of effect ($IVW_{\text{single-cellMR}}$ p-value $= 1.47 \times 10^{-8}$, beta $= 0.15$, $IVW_{\text{RobinspQTL}}$ p-value $= 2.48 \times 10^{-8}$, beta $= 0.39$). Finally *TYMP* was the only other gene (with *GPNMB*) found to be putatively causal for MS in all three QTL datasets ($IVW_{\text{single-cellMR}}$ p-value $= 1.47 \times 10^{-8}$, $IVW_{\text{RobinspQTL}}$ p-value $= 2.48 \times 10^{-8}$, $IVW_{\text{UKB-PPP}}$ p-value $= 1.62 \times 10^{-8}$).

6.3.5 Epigenetic interrogation of MR IVs

The selection of IVs follows a relatively simple series of selections based on thresholds; first, the colocalised region (PP.H4 > 0.8) identifies that the region of interest for both the eQTL and the GWAS. Second, the lead SNP proposed by COLOC is checked for association strength ($< 5\%$ FDR) and finally correlated SNPs are removed ($r^2 > 0.01$) to finally retain a set of independent, robustly associated genetic variants as IVs (in this case usually a single one). Because of the presence of LD however, it is likely that substituting the selected IV with any strongly correlated variant would also lead to a putatively causal association, and therefore begs the question as to which SNP is the likely causal variant. To investigate this question, I

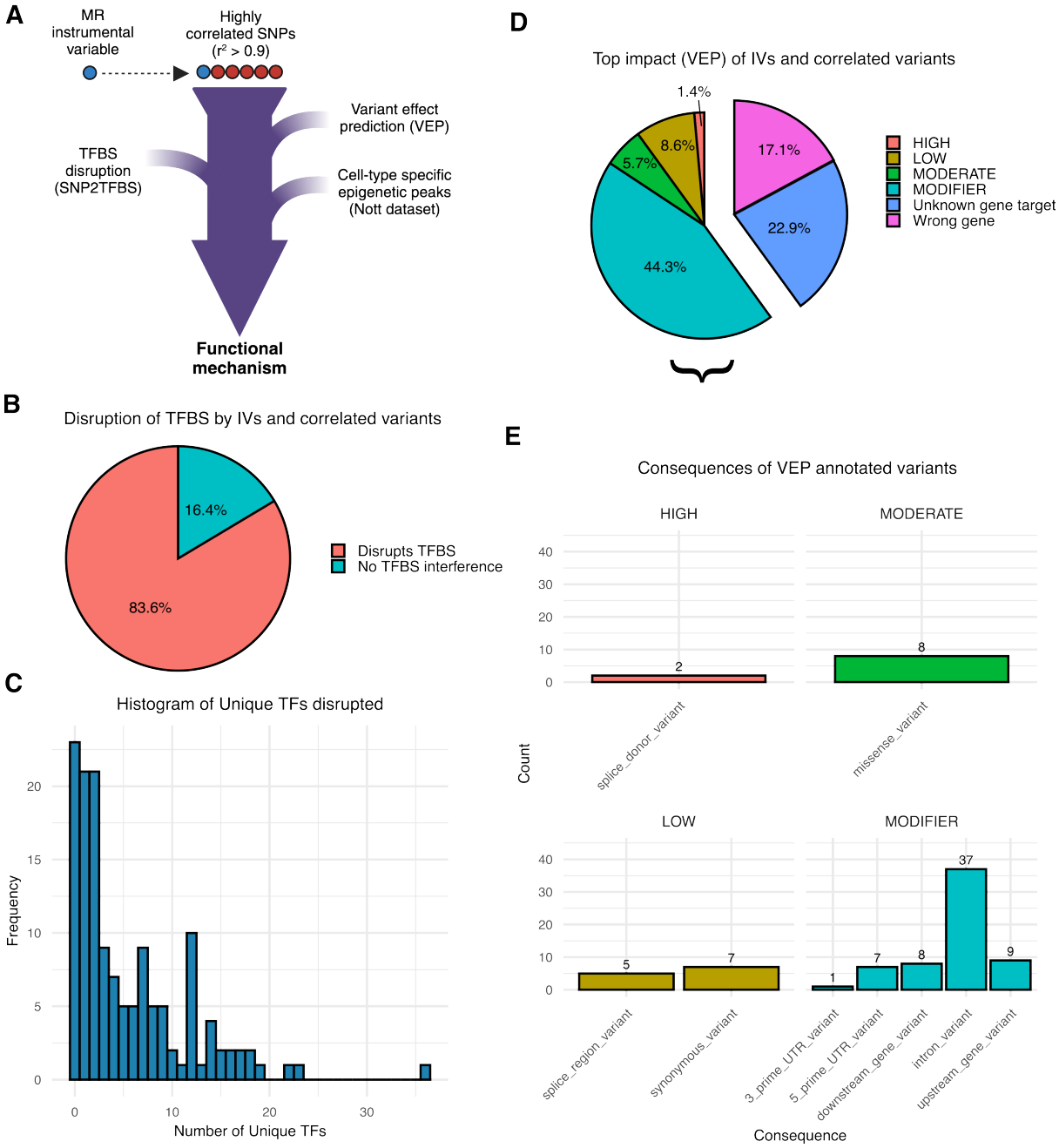


Figure 6.7: Functional assessment of IVs. ***A***, Strategy implemented to query potential function of IVs, extending the search to correlated variants ($r^2 > 0.9$). ***B***, Summary of transcription factor binding site (TFBS) disruption by presence of IV or correlated variants. ***C***, Histogram of number of unique transcription factors (TFs) each set of correlated SNPs are predicted by SNP2TFBS to disrupt. ***D***, Distribution of highest impact score for IVs and their correlated genetic variants on gene expression as determined by VEP (see methods). ***E***, Distribution of variant consequences for variants with the highest impact score.

attempted to interrogate the IVs used in the MR analyses using a series of epigenetic data, including "SNP2TFBS"[208], "Variant Effect Predictor" (VEP)[69] and a published dataset on the enhancer-promoter interactome in the brain by Nott et al.,[209] (Figure 6.7 A).

First, I investigated the overlap of the IVs used in the MR analyses using "SNP2TFBS" which is a curated database of SNPs predicted to disrupt transcription factor (TF) binding at transcription factor binding sites (TFBS), based on position weight matrices curated by JASPAR and other databases of epigenetic data[208]. Of the 140 MR hits, 117 (83.6%) were instrumented by IVs (or in high LD with genetic variants, $r^2 > 0.9$) predicted to disrupt TF binding (Figure 6.7 B). However, given the complexity of these regions despite strict inclusion of highly correlated genetic variants there was often more than one TF which the genetic variants were predicted to disrupt (Figure 6.7 C). Only 21 MR hits predicted a single TF; for example, the MR inference made between *KCTD13* expression in oligodendrocytes and decreased MS risk was instrumented by two independent IVs, *rs1157494* and *rs7542*. The former was predicted through a correlated variant (*rs11865086*, $r^2 = 0.97$) to disrupt *IRF1* binding, which has been linked to the regulation of inflammatory demyelination and oligodendrocyte injury, both of which have an important role in MS[262].

Extending the search for functional annotations using *VEP* and using the scoring annotations used by Ensembl as guidance (Methods), I found that the majority of MR hits were instrumented by IVs (or correlated variants) which were predicted to have a functional consequence in the same gene (84 out of 140, 60%). A smaller proportion had an "unknown" function or gene target (32 out of 140, 22.9%) whereas 24 hits (17.1%) were instrumented by genetic variants for which *VEP* had predicted function in an alternative gene (Figure 6.7 D). Taking the top functional consequence for each set of genetic variants per IV predicted to impact the correct gene, the majority were classified as "modifier" variants, where intronic variants were the most common (Figure 6.7 E). Only two MR hits were instrumented by IVs correlated to genetic variants with a "high" functional consequence; *BTN3A2*, whose expression in both OPCs and Oligodendrocytes were putatively causal for a decreased risk in MDD, was instrumented in those cell-types with two different genetic variants (*rs72841536* and *rs58825580*), both of which correlated with *rs71557335* which was predicted to be a splice donor variant for *BTN3A2*. Despite using the same gene, treating each MR hit separately is important as demonstrated here, because *BTN3A2* was also found to be putatively causal in

inhibitory neurons for other traits (BIP, SCZ, INT). However, the IVs used for the MR inference in inhibitory neurons did not correlate with any functional genetic variants.

Using the data published by Nott et al., 109 of the IVs used for the 140 hits (77.9%) overlapped or were in strong LD ($r^2 > 0.9$) with another SNP which overlapped an epigenetic mark in the correct cell-type. For instance, in oligodendrocytes, epigenetic data was available in all four modalities (H3K4me3- and H3K27Ac-ChIP-seq, ATAC-seq and PLAC-seq). Of the 36 gene / celltype / trait pairs, epigenetic evidence in any modality was found for 33 of the triplets (88.9%) (Figure 6.8 A). As an example, the IV used (*rs79073127*) for the putatively causal inference between *FOXN2* expression and increased SCZ risk ($IVW_{\text{single-cellMR}}$ *p*-value = 1.57×10^{-7}, beta = 0.06) overlapped an oligodendrocyte-specific H3K27Ac peak, indicating a potential enhancer region. Interestingly, this genetic variant was in high LD ($r^2 = 0.98$) with another variant upstream (*rs79247094*) which resided both within another potential enhancer (as denoted by the H3K27Ac peak) and a chromatin contact region as defined by PLAC-seq. These PLAC-seq contacts were enriched for another potential set of enhancer regions downstream of these two variants, suggesting a potential mechanism for these two genetic variants in their regulation of *FOXN2* and consequently SCZ risk (Figure 6.8 B). Curiously, the MR analyses found *FOXN2* expression to be putatively causal for SCZ risk in four other cell-types including OPCs, both classes of neurons and astrocytes, although none of these celltypes showcase the same epigenetic marks in these results.

Epigenetic intersection was successful even for smaller cell-types such as microglia; all five of the IVs in the MR hits for microglia overlapped epigenetic marks (either by itself or through a SNP in high LD) (Figure 6.8 C). For instance, *RASGEF1C* expression was associated with decreased AD risk ($IVW_{\text{single-cellMR}}$ *p*-value = 5.40×10^{-12}, beta = -0.10). The inferences was instrumented by the SNP *rs76792388*, which was situated in a microglia-specific H3K27Ac peak and PLAC-seq loop (Figure 6.8 D). In addition, this SNP was in perfect LD ($r^2 = 1.00$) with three other genetic variants (*rs116347724*, *rs111274391*,*rs113706587*); one of these (*rs113706587*) was the only one to also overlap a microglia-specific ATAC-seq peak. While it is difficult to point towards an exact mechanism, the epigenetic data here shows an enrichment of epigenetic marks specifically for microglia, both as seen by the H3K27Ac and ATAC-seq peak enrichment around the gene start (where the genetic variants are situated) as well as the high concentration of PLAC-seq connections not present in any of the other cell-types.

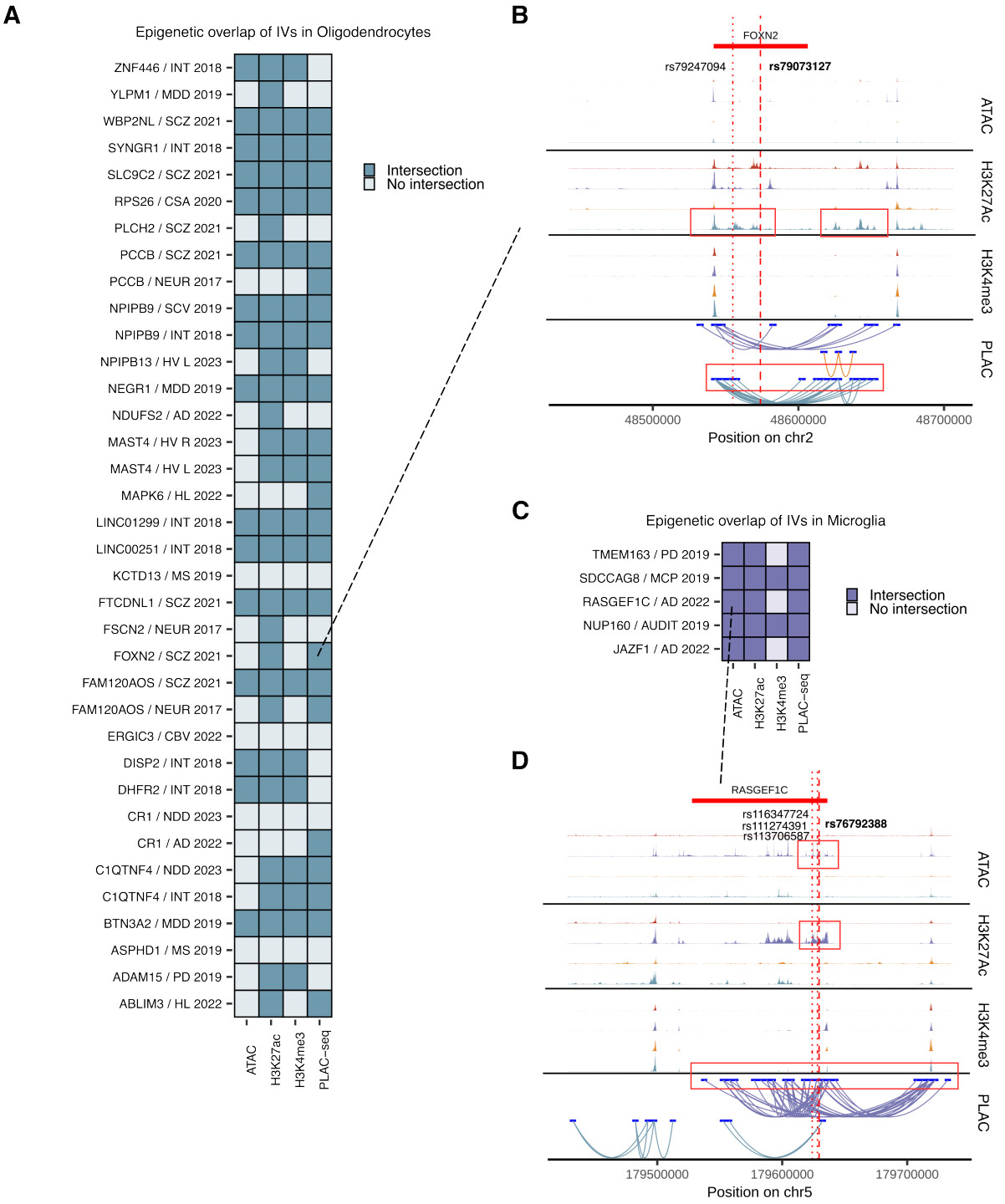


Figure 6.8: Epigenetic overlaps of IVs. ***A***, Overlaps of IVs for each MR hit in oligodendrocytes and their correlated genetic variants ($r^2 > 0.9$) with epigenetic data from Nott et al.,[209]. Dark color represents a positive intersection. ***B***, Example of epigenetic overlap of IVs and correlated variants for the MR hit *FOXN2* / SCZ showing an overlap with a H3K27Ac peak and a PLAC-seq connection. The genetic variant in bold (long red dashes) represents the IV selected for MR. ***C***, Overlaps of IVs for each MR hit in microglia and their correlated genetic variants with epigenetic data. ***D***, Example of epigenetic overlap of IVs and correlated variants for the MR hit *RASGEF1C* / AD showing an overlap with a H3K27Ac peak, an ATAC peak and a PLAC-seq connection.

6.4 Discussion

The results of this chapter show the first MR analyses using cell-type specific eQTLs in the human brain. Using eQTLs anchored in control-tissue, I discovered 94 putatively causal genes in one or more cell-types in 22 brain-related traits across behavioural, structural, psychiatric and neurological spectrum. I also show that the majority of genes are cell-type specific (81.9%) and only a few (13.8%) appear in more than one trait. Moreover, the results here are to some extent independently validated using published pQTL data and show that use of eQTLs as proxies for their downstream protein products is a fruitful approach in causal inference. Finally, use of external epigenetic data suggests potential mechanistic functions for the genetic variants instrumenting these putatively causal genes and identifies plausible causal variants.

6.4.1 Cell-type specific MR

Use of MR to uncover directional effects has important implications for therapeutic targeting as it informs whether activation or inactivation is desired. In particular, context-specificity is important because it underpins the cell-types that contribute to disease pathology. In this work, several genes have been uncovered with directional effects in specific cell-types that are potentially of interest. For instance, the putatively causal association between *EGFR* in astrocytes in AD is particularly interesting given the recent interest in repurposing *EGFR* inhibitors for use in AD[263,264]. In another example, the link between *TMEM163* (which is involved in maintaining zinc homeostasis) and PD was also shown in a recent massive parallel reporter assay study of PD variants[265], but was also only mapped recently to microglia by Bryois et al.,[108]. In ALS, it is known that genetic variants associated with *SCFD1* confer risk[257], and the results here show that this is specific to neuronal cell types.

In an interesting finding one of the genes to be putatively causal for AD was *CR1*, specifically in oligodendrocytes and OPCs. While this oligodendrocyte-specific inference has also been discovered in similar studies (as well as in OPCs, and sometimes including *CR2*)[108,266], it is curious given that *CR1* and *CR2* encode complement receptors and have a well-documented role within immune cells[267]. It is additionally perplexing when considering that *CR1*, given its role in immune mechanisms, does not have strong associations within microglia which play the role of the resident immune cells in the brain; this lack of association is also the case in

the related studies mentioned. Aside from these genetic association studies, *CR1* expression is enriched in oligodendrocytes according to the human protein atlas[268] although to a higher degree in microglia according to the recent brain cell atlas[269]. Some studies have shown that oligodendrocytes play a potentially important immunomodulatory role in response to inflammation beyond their traditional role of maintaining synaptic plasticity, both in response to injury and neuroinflammation in the case of MS[270,271]. While the role of *CR1* (and *CR2*) within oligodendrocytes is yet unclear, the emerging evidence supports a homeostatic maintenance role for oligodendrocytes in which complement pathway receptors may play an important role for AD pathology, recently suggested to play a dual role of neuroprotection and neuroinflammation upon dysregulation[272].

Shared genes between traits mediated by the same cell-type are also particularly interesting, offering potentially dual actions of intervention methods. The finding that *PCCB* expression in multiple cell-types is putatively causal for both SCZ and NEUR supports the well-known relationship between NEUR as a risk factor for SCZ[273]. While in this case this link was found in both classes of neurons and oligodendrocytes (crossing the glia/neuronal span) it may very well be cell-type agnostic given the lack of power in less common cell-types in this dataset.

6.4.2 Protein replications

Replicating MR inferences at protein level using pQTL datasets is desirable for several reasons; first, there is an inherent assumption that the relationship through a gene and a trait is through its protein product, exerting its functions to modify cellular behaviour and therefore the desired target in therapeutic development[274]. Second, it provides with an independent source of validation, where genetic associations to genes are replicated in proteins assayed in individuals independent of the study cohort. Finally, the replication of inferences in other tissues such as in plasma as done here with UKB-PPP provides a potentially important resource of biomarkers that may point towards predisposition of a trait. In contrast to the single-cell eQTL dataset, UKB-PPP is very well-powered (N > 54,000 individuals) and should represent accurate genetic associations, but suffers instead from smaller proteomic panels as less than 3,000 proteins were assessed. Despite this, the results here achieve a strong replication rate (80% in COLOC and 55% in MR) despite 1. being in a

different tissue type at bulk-level and 2. only 20 gene-trait pairs tested. Larger protein panels would therefore be desirable to link more putative gene-trait biomarkers in blood.

Unfortunately sample acquisition in brain is still difficult at scale compared to plasma, and as such the replication using the independent proteomic brain resource by Robins et al.,[206] is of limited value. First, the dataset is of similar size to single-cell dataset (N = 330) and therefore of comparable statistical power, but also includes disease cases (31% of samples were from individuals diagnosed with AD). Despite this, a substantial proportion of gene-trait pairs were validated here (20 out of 22, 90.9% of selected protein-trait pairs), although many were lost due to lack of significant associations in the IV selection step (only 22 out of 50 gene-trait pairs had a significant pQTL). The improvement of proteomic assays, although still expensive, promises to lower the bar for large-scale proteomic assays in the brain should sample acquisition become easier; in addition, combining brain proteomics with plasma as well as cerebrospinal fluid could uncover robust potential drug targets[82].

Several genes validated here are of significant interest however; *GPNMB* for instance, was validated both with UKB-PPP and the Robins proteomic dataset with the same direction of effect (increased expression is putatively causal for increased PD risk). This result is consistent with recent findings that *GPNMB* is selectively elevated in PD patients[275] (although specifically in substantia nigra, which was not present in this study). In addition, knock-out studies in iPSC-derived neurons by Diaz-Ortiz et al., showed that *GPNMB* interacts with alpha-synuclein and is required for the uptake of its fibrillar forms, contributing to alpha-synuclein pathology in PD[276].

6.4.3 Epigenetic overlaps

In this work, I was able to interrogate the potential mechanisms by which the genetic variants affect the gene expression of their gene target through the use of VEP, SNP2TFBs and the cell-type specific epigenetic data published by Nott et al.,[209]. Extending the search to correlated genetic variants is important because they are likely to be co-inherited at conception. In addition, inclusion of these genetic variants relaxes the strict thresholds placed on the IV selection, where slight numerical differences may lead to the exclusion of potentially causal variants. Because they are highly correlated, the effect sizes of their eQTL associations

to the target gene are likely extremely similar, and would therefore be equally valid to use as IVs in an MR setting all whilst obtaining a similar MR-IVW effect size.

In the large majority of variants, the results here show putative mechanistic functions and prove a potential explanation into their impact on gene expression (and consequently phenotypic trait risk or predisposition). In particular, the epigenetic data by Nott et al., is an important resource for this analysis because of both the specificity to the brain as well as cell-type. Validating these epigenetic overlaps with other epigenetic datasets such as *BOCA2*[136], or datasets containing disease-specific epigenetic patterns such as in AD[277] would be valuable in strengthening these observations.

It is important to note that these analyses present a first look at the possible functions of these variants but are not in any sense exhaustive. There is a plethora of tools and databases that have been compiled for the functional annotation of genetic variants; in fact, the *VEP* tool hosts more than 70 add-on packages that can extract further information surrounding genetic variants including pathogenicity prediction, species and ancestral conservation assessment and de novo variant identification to name a few (https://www.ensembl.org/info/docs/tools/vep/script/vep_plugins.html). New methods are also frequently integrated; for example, a recent method called *spliceAI* makes use of neural networks to predict possible splice donor and acceptor sites from reference pre-mRNA transcripts, scoring the probability of a variant being splice-altering[278]. Google Deepmind's deep learning architecture *Enformer*[279]for gene expression predicition and *AlphaMissense*[280]for predicting the impact of genetic variation on protein structure (and potential pathogenicity) are also available as extensions.

Chapter 7

Discussion

The work conducted in this thesis has aimed to provide with a comprehensive framework to link genetics to cell-type specific gene expression in the brain and the relation of such genes to trait predisposition across the brain phenome. To briefly summarise this thesis:

- Using snRNA-seq in the brain, I present 13,939 unique genes regulated by one or more genetic variants in one or more of 8 brain cell-types. In addition, eQTLs are cell-type specific to a high degree (39.1% of all eGenes discovered were only found in one cell-type) (**Chapter 3**).
- Around 16.7-40.8% of eQTLs (depending on cell-type) display a significant impact of disease not captured by standard eQTL methodology. Using a controls-only cohort, a further 7.9% - 45.1% of eQTLs are dynamically regulated by age (**Chapter 4**).
- Employing phenome-wide genetic colocalisation approach, I uncover 501 colocalisations across 30 phenotypic traits, linking genetic susceptibility to a specific gene and cell-type context. A substantial proportion (77.9%) of gene-trait pairs were only found in a single cell-type, highlighting the importance of rooting these inferences in single-cell data. In addition, 118 out of the 501 (23.6%) of colocalisations are lead by dynamically regulated eQTLs, showing that disease-interaction effects can impact downstream associations (**Chapter 5**).
- Finally, using eQTLs unbiased from the presence of disease, I apply MR to find 140 directional and putatively causal effects between cell-type, gene and 22 brain-related

traits. A proportion of these are independently replicated in external proteomic datasets, identifying potentially robust candidates for therapeutic targeting (**Chapter 6**).

While I have shown how powerful this framework can be in uncovering putatively causal drug targets and potential biomarkers, several considerations present themselves along the way of which the most important is the complex interaction of disease on observed gene expression. Identifying these effects however is crucial in understanding drivers of phenotypic presentations, and will help in fine-tuning future target identification efforts that are rooted in genetics. This chapter will attempt to dissect this thesis into a broader discussion surrounding the main findings and their value in the field of QTL-driven causal inference.

7.1 eQTLs at single-cell level; where to next

In **Chapter 3**, I discovered 13,939 unique genes associated to common genetic variation across 8 cell-types in the brain. I showed that there is a high degree of cell-type specificity as 5,454 (39.1%) of eGenes were unique to a single cell-type. Compared to previous single-cell eQTL studies in the brain as done by Bryois et al.,[108], this represented an increase of 83.2% (from 7,607 eGenes in the same 8 cell-types) in large part due to the massive increase in sample size (almost double). Another recent study published in March of 2024 by Fujita et al.,[266] used a similar sized cohort (N = 424) to uncover 10,024 eGenes at the same annotation level, but takes these results one step further by annotating 64 cell subtypes within the 8 major cell-types (uncovering an additional 8,099 eGenes).

The most pressing question is, where to next? The first and most obvious answer is to follow in the footsteps of traditional bulk eQTL analyses and continue increasing sample sizes as a simple (yet costly and sometimes difficult) approach to uncover a larger set of genetically regulated genes that can be overlapped with GWAS signals. This is becoming increasingly feasible with decreasing sequencing costs and a mature single-cell market with competitive pricing[281]. However, only increasing sample sizes will likely plateau as shown in GTEx as well as Metabrain[90,282]. In addition, the recent work by Mostafavi et al.,[283] systematically comparing GTEx eQTLs with 44 traits from the UK BioBank shows that eQTL and GWAS signals are distinct, each powered to map different sets of variants; GWAS genes are generally more constrained and have complex regulatory landscapes across tissues, whereas strong

eQTLs cluster near transcription start sites and as a whole generally are devoid of functional annotations compared to GWAS genes. In a way, this is by design because eQTL studies can be treated as thousands of independent GWAS that aim to uncover associated variants to single specific genes, which will naturally bias the selection to more direct genetic regulators such as within transcription start sites, enhancers or promoters. On the other hand, GWAS find associated variants that have strong effect sizes in relation to trait predisposition at the organism level and is consequently agnostic to gene-specific association bias.

Bridging that gap however is crucial in identifying targetable genes responsible for trait predisposition or disease pathology. Sample size is still relevant, but will take time at single-cell level to reach the same levels currently in bulk and even then may still suffer from the biases outlined above. One of the main benefits of single-cell eQTLs however is the presence of an additional dimension, which is the number of cells that can be sequenced by individual. As with bulk, pseudo-bulking expression also suffer from mixing signals from potentially multiple subpopulations of cells, some of which have a greater importance in disease biology. While cell-subtypes have been explored in the brain by Fujita et al.,[266] they are the result of iterative subclustering methods and not necessarily representative or even "real" cell populations (especially considering a median of 3,824 nuclei per individual). Rather, future studies will greatly benefit from deeper cell-level sequencing per individual to capture rarer cell-type populations, with new technologies allowing the capture of up to 30,000 cells per sample[281]. These sub cell-type populations are additionally increasingly easy to map, thanks to large-scale efforts to generate consensus single-cell brain atlases (most recently by Chen et al., in August of 2024)[98,269].

Beyond eQTLs, it would also make sense to look within other molecular traits. As mentioned in the introduction, there are various methods to uncover genetically regulated epigenetic marks such as chromatin QTLs[78], splicing QTLs[77] and pQTLs[75] that can all help strengthen associations with a full set of surrounding biological information. Because they rely on the same philosophy as eQTLs however they are equally prone to their biases. As such, alternative functional methods have also been used in order to uncover potential regulatory variant functions; for instance, massively parallel report assays (MPRAs) allow quantification of the effect of regulatory variation on a readout such as expression or epigenetic marks and have been used both to understand GWAS signals[284] and tissue-specific functions, most recently at single-cell level[285]. In addition, directly modulating a genomic sequence through

targeted delivery methods such as CRISPR/Cas9 may better represent their effects. For instance, Morris et al., developed "STING-seq", designed to interrogate GWAS signals through massively parallel CRISPR screens at single-cell level[286].

7.2 Why using disease cases can still be valuable

In the majority of large-scale studies, the convention has been to use any sample available (disease or control) in order to maximize statistical power for eQTL discovery, and adjust for clinical covariates during the modelling as fixed covariates[66,90,220]. In **Chapter 4** however, I show that at least 16.7-40.8% of eQTLs tested are dynamically regulated by disease status, even after the standard approach to correct for the presence of disease. Some of these were affected by a single disease state, potentially pointing towards disease-specific genes whereas other interaction eQTLs were affected by more than one disease state which could explain converging disease pathways. In addition, a further 7.9% - 45.1% of eQTLs in the controls-only cohort unbiased by presence of disease were dynamically regulated by age, which is another factor to contend with when putting together large cohorts.

These results show that eQTL discovery through the use of disease tissue, especially when combined with other disease states and controls, should be interpreted with caution given that their presence can bias association estimates. This is important because, as is later shown in **Chapter 5**, downstream causal inferences between gene and trait can be biased simply because the initial discovery contains disease-affected associations. As noted earlier, it is also important to note that the analyses here have been restricted to the significant SNP-gene associations, and only for the top SNP of each gene within each cell-type. Given that the majority of eGenes are associated to more than one SNP (although many of these are due to LD), it is possible that the impact of disease affects the association of each SNP differently. In addition, it is unclear how many associations are "obscured" by the presence of disease, explained by disease causing an opposing effect on expression to that of control-level expression. Only be applying this interaction framework genome-wide can this be clarified, which requires substantial computing resources.

There is however clear value in the inclusion of disease samples. At scale, just as differential expression is routinely applied, so too can differential eQTL analyses provided control and

case cohorts are similarly powered. For instance, such an approach could identify eQTLs which are consistent across controls and disease and are unaffected by disease pathology which becomes valuable in a causal inference setting. It could also identify eQTLs that are dynamic, where presence or absence of disease pathology confers a change in gene expression that is reflected in eQTL associations. While this has been shown in **Chapter 4**, the difference in statistical power (N = 391 versus N = 183) could also explain a large proportion of the observed differential eQTLs.

Importantly, establishing a causal link between gene and disease would require one of two scenarios; either, the effects of genetic variation on a gene are only uncovered in the presence of disease pathology, or, that the effect is consistent across control and disease tissue. An example of this was recently shown in frontotemporal dementia (FTD), where risk SNPs associated to the *UNC13A* gene confer an increase in pathogenic exon inclusion but is not present in controls samples[287]. There might be several other examples like this where risk variants accelerate disease progression, but only in the presence of disease pathology which in turn "unlocks" disease risk. On the other hand, should an effect be consistent across control and disease tissue then the possibility of reverse causation can be eliminated, and as such can be tested in a directional causal framework (such as with MR)[172].

7.3 Colocalisation and MR for putative causal inference

Using eQTL-based colocalisation has become a popular approach to link genetic predispositions with gene expression, most recently at single-cell level as exemplified here in **Chapter 5-6** as well as by previous studies in plasma[107]. While colocalisation has also been performed in the brain using cell-type specific eQTLs as shown in Bryois et al., as well as Fujita et al.,[108,266], the traits included in their studies are restricted to PD, AD, MS, ALS and SCZ. Employing a brain phenome-wide approach, I present 501 genetic colocalisations in 30 phenotypic traits (of 41 tested) across the psychiatric, behavioural, structural and neurological spectrum. The cell-type specificity is importantly highlighted here, and shows the value of using single-cell inferred eQTLs; 77.9% of gene-trait pairs were only present in a single cell-type (at PP.H4 > 0.8).

An important finding as mentioned above is that a substantial proportion of eQTLs are affected by the presence of disease and bias downstream results. In total, 118 colocalisations were driven by a shared SNP (the lead causal SNP by the COLOC framework) that had a disease-interaction in the association to the gene. While it is true that at this stage a more liberal threshold was used (LRT p-value < 0.05), it is unclear how the effect of disease exerts itself on multiple variants in a region and their consequent effect on the colocalisation. In particular, correlated variants are likely to be equally affected by the presence of disease and as such may collectively contribute to a colocalisation signal. The opposite is of course also possible, where a combined disease effect in the region may abolish any signal preventing colocalisation. Just as dynamic eQTLs have been of interest recently, so too should dynamic colocalisations[109,110].

The results presented here show the first MR framework applied to single-cell eQTLs in the brain, using a sub-cohort restricted to controls and therefore unbiased by disease presence. I show here 94 unique genes with directional (putative) causal effect and cell-type specific context for 22 different phenotypic traits, offering targets for risk mitigation. Some important results are shown here such as *EGFR* for AD (specifically in astrocytes), which have previous evidence as a potential dual target for both cancer and AD[263]. Another target shown here which has also an evolving role in PD is *GPNMB* (OPCs and oligodendrocytes), shown to play a role in alpha-synuclein pathology by facilitating uptake of its fibrillar forms which is a hallmark of PD pathology[276]. Several other targets have little previous evidence in their respective traits, such as *JAZF1* / AD, *RBM6* / INT , *LSM6* / ADHD, *ABLIM3* / HL to name a few. In addition 11 inferences were replicated in UKB-PPP, a majority of which have the same direction of effect. This is important because it shows that a genes protein product in plasma mirrors that of the effects in post-mortem brain, and could serve as useful biomarkers for diagnostic or preventative targeting. The replication using the post-mortem brain proteomic dataset by Robins et al., is also encouraging, given that 80% of inferences have the same direction of effect at the protein level[206].

It is important to note however that "unbiased by disease presence" is used here to describe samples which upon death had no observable disease pathology, but it is possible that unobservable biological mechanisms are already placing these patients on a disease trajectory. In fact, the results show in **Chapter 4** that even in the controls-only cohort, a large proportion of eQTLs (7.9% - 45.1%, by cell-type) are dynamically regulated by age which is

the largest risk factor for neurodegenerative disease, some of which make up a large part of the cohort used here[228]. While these age-interacting eQTLs were overlapped with the disease-interacting eQTLs discovered in the full cohort and removed prior to downstream MR analyses it is possible that more subtle effects remain.

Several methods now extend the capabilities of standard colocalisation and MR. As shown in these results, many colocalisations were shared both between cell-types as well as traits based on a simple intersection. There are however tools that can explicitly test these relationships in a statistical framework. For example, *HyPrColoc*[139] was recently developed to test shared colocalisation across multiple traits at once by increasing the number of colocalisation hypotheses to be tested simultaneously. Applying this method could highlight shared genetic susceptibility for multiple traits, and their colocalisation to a specific gene (or more) with cell-type specific context. It could also be applied at the cell-type eQTL level alone to query whether shared significant eGenes in multiple cell-types share the same genetic regulation, or whether they are associated to distinct sets of genetic variants. If they are, this could in turn highlight situations where an eGene is present in multiple cell-types but only colocalise with a trait in a single one.

One of the main caveats with standard MR as used here is its vulnerability to horizontal pleiotropy, whereby an IV affects the outcome through other pathways than the tested risk factor, for example through its effect on another gene. This is important because such a situation would violate the assumption in MR that the IV affects the outcome only via the exposure which in this case is the tested gene. In addition and similarly to the colocalisation results, the interpretation of a single exposure (should there be no horizontal pleiotropy) affecting multiple outcomes becomes difficult using this methodology. In 2023, we developed MR^2 , a multivariable MR framework that is specifically designed to identify exposures associated to multiple outcomes, and conversely multiple exposures that confer risk to distinct outcomes[288].

7.4 Beyond colocalisation and MR

Genetic colocalisation and MR has become the gold standard to perform inference of putative causal associations between gene and trait[149,174,289]. However, there are several other methods

that have garnered interest in the last few years which could complement this work.

Mediation analysis for example can be used as a complementary method to MR, whereby you explicitly test for a mediating effect between cause and effect[290,291]. In its simplest setting, where the exposure E is causal for the outcome O, traditional mediation analysis aims to model three different effects; first, the total effect from E to O through all potential pathways. Second, the remaining effect of E to O after removing the effect from any other potential pathways (also called the direct effect). And finally, the mediating effect, which measures the effect of E to O explicitly through the mediator. With this information and assuming that the effects all act in the same direction, it is then possible to calculate the proportion of the total effect that can be explained by the mediator. For instance, a randomised trial conducted by Freeman et al., showed that sleep quality had a mediating effect on the observed effect of behavioural therapy on psychotic experiences[292]. In observational studies such as with eQTLs for causal inference using MR, mediation analysis could be implemented to test whether a directional putative causal effect (such as gene expression in a certain cell-type and its association to AD risk) is mediated by other covariates such as age or sex. While similar to multivariable MR as described above (such as to test the effect of multiple genes on an outcome), the main difference is generally the lack of IVs to anchor the inferences[288,293].

Transcriptome-wide association studies (TWAS), which uses a gene-centric approach to identify the relationship between predicted gene expression and a GWAS trait[294], is also of potential interest here. The main difference between this method and colocalisation is the use of a model trained on eQTL data to impute probable gene expression signatures from a GWAS trait, without the need of individual data or expression measurements. Beyond measurements of molecular traits entirely, phenome-wide association studies (PheWAS) makes use of the increasing presence of electronic health records, where a host of clinical phenotypes are recorded such as vital signs, past health issues and diagnostic procedures[295]. In addition, cohorts which collect these information at large scale (such as UK BioBank, FINNGEN, 23andme) have powered PheWAS studies to map genetic relationships between thousands of clinical endpoints; Diogo et al., in 2018 demonstrated its use by uncovering a genetic relationship between gout, gallstones, acne and adverse drug events with a variant in the vicinity of the *PNPLA3* gene[231].

7.5 Future work

The results shown in this thesis provide several avenues for future studies building on the inferences made here. On the computational side, it would be interesting to extend the analyses presented here at a larger scale. At the eQTL level, no single-cell brain study has performed a genome-wide *trans*-eQTL study. Although it could be argued that their reliability might be poor given that substantial power is required for discovery of *trans*-effects[4], it is still worth exploring given that eQTLs derived from single cell-types generally have stronger effects. The addition of *trans*-eQTLs may then be implemented in testing reverse causation of MR inferences, where it can be tested whether the combined genetic effects across the genome for a gene overlap with risk variants, thereby causing a bidirectional effect. Beyond this, identifying conditionally independent *cis*-eQTLs (where the effect of the top SNP is masked and added as a covariate) to increase the number of IVs in MR is also a potential next step and is common practice[296].

The eQTL interaction modelling also offers multiple research avenues. First and as mentioned earlier, performing a fully genome-wide interaction analysis is crucial in identifying the full set of genes which may or may not be masked by the presence of disease. It is also theoretically possible to test interactions with multiple potential covariates, such as pseudotime trajectory as done recently[110] and age as presented here. Given accurate measurements of pathology (such as *tau* or *alpha-synuclein)* for each patient, it should also be possible to test whether these dynamically regulate certain eQTL associations and how that affects downstream colocalisation patterns. In addition, understanding the combined effect of an interaction on all associated SNPs to a specific gene is important, because it is likely that an interaction affects each association differently (unless the SNPs are perfectly correlated). If these relationships can be clarified it could provide with a better understanding on how these affect dynamic colocalisations at region-level.

To strengthen the MR associations an important next step would be a more rigorous assessment of the epigenetic landscape surrounding the inferences in attempts to put forward well-supported causal variants. While many databases exist that could help refine their functions (as described in the discussion section of **Chapter 6**), the main leap forward would ideally consist of large-scale cell-type specific epigenetic assays as done by Nott et al.,[209]. In

fact, many of these assays have been applied at the single-cell level; for instance, the 10X multiome assay allows simultaneous assessment of open chromatin and gene expression in the same cell[191], and ChIP-based methods as well as chromatin conformation capture are also available at single-cell level[297–299]. The majority of these suffer from high levels of sparsity (as well as high cost) that is commonplace in single-cell approaches. Traditional bulk-level approaches are therefore still widely used, and are able to assay cell-type specificity when combined with isolation of cell-types through cell-sorting[209]. Combining these approaches together can elucidate genetic variant mechanisms as shown by our recent work, where a cohort of 34 individuals were jointly assessed by ATAC-seq, RNA-seq and Hi-C to discover "trimodal" QTLs[300]. In particular, the recent development of Micro Capture-C (MCC) has been shown to determine chromatin architecture at base-pair resolution, allowing clear identification of variant functions on chromatin organisation and their downstream impacts[63,64].

Functional validation is also a potential next step. At the gene-level, several methods can be applied to test whether targeting of the genes is possible in the desired cell-type. The use of antisense oligonucleotides (ASOs) has recently gained momentum as a potential therapeutic strategy in targeting specific genes to counter disease pathology, as shown recently by the FDA-approved ASO-mediated therapy for duchenne muscular dystrophy, spinal muscular atrophy and familial amyloid polyneuropathy[301,302]. Several ASOs have also been explored in other disease and show promise, such as those targeting alpha-synuclein pathology in PD[303] or in AD[304], although penetrating the blood-brain barrier is still an issue and must generally be administered spinally. These molecules are essentially short single-stranded synthetic oligodeoxynucleotide sequences that aim to modify final protein products by targeting mRNA molecules. Depending on the chemistry of the specific ASO, they can for example inhibit the translation of the mRNA by inhibiting the translation process itself or prompt degradation of the mRNA molecule through several pathways. Because they can also target splicing junctions, they can be particularly useful for disorders that are the result of specific splicing defects, where the ASO can restore normal splicing behaviour by introducing a frameshift. It would therefore be useful given their flexibility to them on the targets presented here, either *in vitro* using appropriate cellular systems (such as iPSC-derived glia[305]) before testing *in vivo* as done in mice[306].

To maximize chances of success functionally, targets should first be restricted to those with

the most evidence (i.e those replicated at protein level). Beyond the ones represented here, further work to expand the sets of proteins assayed at scale would help strengthen inferences made at single-cell eQTL level. Commercially available technologies assaying proteomics at scale are evolving rapidly. For example, the current instance of the *Olink* platform as used in the UKB-PPP study[207] only assayed ~3,000 proteins but is on its way to 5,400 proteins as demonstrated by a recent study[307]. A competing technology from *Somalogic (Somascan 11k)* aims to assay up to 11,000 proteins, and may improve target validation at a larger scale. The shift to proteomics in the field of target identification and its value in risk prediction was recently demonstrated by Carrasco-Zanini et al.,[308] which showed that the predictive ability of protein signatures outperformed predictive models derived from clinical information in 68 diverse diseases.

7.6 Concluding remarks

The work presented in this thesis provides a powerful framework for target identification in neurological disease and phenotypic presentation, rooted in the use of gene expression combined with genetic information. The data shown here is one of the largest single-cell expression datasets in the post-mortem brain to date that is coupled with genotypic information at the individual level, and is aimed to become a resource for the wider research community for understanding brain health and disease.

In considering dynamic allelic effects I highlight an important consideration for therapeutic targeting; since treatment targets may be repurposed for risk mitigation (and vice-versa), it is crucial for these effects to be accounted for in therapeutic design. Notably, use of both control and disease tissue combined with careful understanding of exposure-outcome inferences is a potentially powerful approach for target identification, as highlighted by differences in genetic colocalisation patterns. In addition, use of accompanying molecular data (such as proteomics as shown here) has the potential to identify robust biomarkers predictive of clinical outcome and disease risk. Coupled with knowledge of cell-types in which these effects are measured (and their directionality), single-cell eQTL anchored causal inference promises to guide future therapeutic design for risk mitigation and preventative treatment.

References

1. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* **47**, 702–709 (2015).
2. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 1–21 (2021).
3. Aguet, F. *et al.* Molecular quantitative trait loci. *Nat Rev Methods Primers* **3**, 1–22 (2023).
4. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).
5. Vos, T. *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **396**, 1204–1222 (2020).
6. Steinmetz, J. D. *et al.* Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Neurology* **23**, 344–381 (2024).
7. Winkler, A. S. The growing burden of neurological disorders in low-income and middle-income countries: Priorities for policy making. *The Lancet Neurology* **19**, 200–202 (2020).
8. Rudnicka, E. *et al.* The World Health Organization (WHO) approach to healthy ageing. *Maturitas* **139**, 6–11 (2020).
9. Collaborators, G. 2019. M. D. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry* **9**, 137–150 (2022).

10. Brådvik, L. Suicide Risk and Mental Disorders. *Int J Environ Res Public Health* **15**, 2028 (2018).

11. Stein, D. J. *et al.* Integrating mental health with other non-communicable diseases. *BMJ* **364**, l295 (2019).

12. THE BRAINSTORM CONSORTIUM *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).

13. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* **47**, 20–33 (2016).

14. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, eaag1166 (2017).

15. Aarsland, D., Laake, K., Larsen, J. P. & Janvin, C. Donepezil for cognitive impairment in Parkinson's disease: A randomised controlled study. *Journal of Neurology, Neurosurgery & Psychiatry* **72**, 708–712 (2002).

16. Verger, A. *et al.* FDA approval of lecanemab: The real start of widespread amyloid PET use? — the EANM Neuroimaging Committee perspective. *European Journal of Nuclear Medicine and Molecular Imaging* **50**, 1553 (2023).

17. van Dyck Christopher H. *et al.* Lecanemab in Early Alzheimer's Disease. *New England Journal of Medicine* **388**, 9–21 (2023).

18. Sims, J. R. *et al.* Donanemab in Early Symptomatic Alzheimer Disease: The TRAILBLAZER-ALZ 2 Randomized Clinical Trial. *JAMA* **330**, 512–527 (2023).

19. Armstrong, M. J. & Okun, M. S. Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA* **323**, 548–560 (2020).

20. Alexander, W. Pharmacotherapy for Post-traumatic Stress Disorder In Combat Veterans. *P T* **37**, 32–38 (2012).

21. Franklin, M. E. & Foa, E. B. Treatment of Obsessive Compulsive Disorder. *Annual Review of Clinical Psychology* **7**, 229–243 (2011).

22. Shah, N., Grover, S. & Rao, G. P. Clinical Practice Guidelines for Management of Bipolar Disorder. *Indian J Psychiatry* **59**, S51–S66 (2017).

23. Bruijnzeel, D., Suryadevara, U. & Tandon, R. Antipsychotic treatment of schizophrenia: An update. *Asian Journal of Psychiatry* **11**, 3–7 (2014).

24. Paul, S. M. & Potter, W. Z. Finding new and better treatments for psychiatric disorders. *Neuropsychopharmacol.* **49**, 3–9 (2024).

25. McFarthing, K. *et al.* Parkinson's Disease Drug Therapies in the Clinical Trial Pipeline: 2023 Update. *Journal of Parkinson's Disease* **13**, 427–439 (2023).

26. Cummings, J. *et al.* Alzheimer's disease drug development pipeline: 2023. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **9**, e12385 (2023).

27. Bierhansl, L. *et al.* Thinking outside the box: Non-canonical targets in multiple sclerosis. *Nat Rev Drug Discov* **21**, 578–600 (2022).

28. Miller, A. H. & Raison, C. L. Burning down the house: Reinventing drug discovery in psychiatry for the development of targeted therapies. *Mol Psychiatry* **28**, 68–75 (2023).

29. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet* **15**, e1008489 (2019).

30. Ochoa, D. *et al.* Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nature Reviews Drug Discovery* **21**, 551–551 (2022).

31. Davies, N. M., Holmes, M. V. & Smith, G. D. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).

32. Burgess, S. *et al.* Using genetic association data to guide drug discovery and development: Review of methods and applications. *Am J Hum Genet* **110**, 195–214 (2023).

33. Hong, E. P. & Park, J. W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inf* **10**, 117–122 (2012).

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

35. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research* **51**, D977–D985 (2023).

36. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* **63**, 168–174 (2006).

37. Wirdefeldt, K., Gatz, M., Reynolds, C. A., Prescott, C. A. & Pedersen, N. L. Heritability of Parkinson disease in Swedish twins: A longitudinal study. *Neurobiol Aging* **32**, 1923.e1–1923.e8 (2011).

38. Hilker, R. *et al.* Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biol Psychiatry* **83**, 492–498 (2018).

39. Zhao, B. *et al.* Heritability of Regional Brain Volumes in Large-Scale Neuroimaging and Genetic Studies. *Cerebral Cortex* **29**, 2904–2914 (2019).

40. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* **54**, 412–436 (2022).

41. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

42. 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).

43. Howard, D. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. University of Edinburgh https://doi.org/10.7488/ds/2458 (2019).

44. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat Genet* **55**, 198–208 (2023).

45. Watson, H. J. *et al.* Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet* **51**, 1207–1214 (2019).

46. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet* **53**, 817–829 (2021).

47. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* **50**, 912–919 (2018).

48. Brouwer, R. M. *et al.* Genetic variants associated with longitudinal changes in brain structure across the lifespan. *Nat Neurosci* **25**, 421–432 (2022).

49. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).

50. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).

51. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

52. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).

53. Ding, K. & Kullo, I. J. Methods for the selection of tagging SNPs: A comparison of tagging efficiency and performance. *Eur J Hum Genet* **15**, 228–236 (2007).

54. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443–1448 (2016).

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

56. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**, 959–962 (2017).

57. Nakato, R. & Sakata, T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods* **187**, 44–53 (2021).

58. O'Geen, H., Echipare, L. & Farnham, P. J. Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. *Methods Mol Biol* **791**, 265–286 (2011).

59. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319**, 1791–1792 (2008).

60. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322**, 1845–1848 (2008).

61. Kawaji, H. *et al.* Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24**, 708–717 (2014).

62. Belton, J.-M. *et al.* Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 10.1016/j.ymeth.2012.05.001 (2012).

63. Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**, 125–129 (2021).

64. Downes, D. J. *et al.* Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. *Nat Genet* **53**, 1606–1615 (2021).

65. Hamley, J. C., Li, H., Denny, N., Downes, D. & Davies, J. O. J. Determining chromatin architecture with Micro Capture-C. *Nat Protoc* **18**, 1687–1711 (2023).

66. Akbarian, S. *et al.* The PsychENCODE project. *Nat Neurosci* **18**, 1707–1712 (2015).

67. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

68. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

69. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

70. Waldmann, P., Mészáros, G., Gredler, B., Fürst, C. & Sölkner, J. Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.* **4**, (2013).

71. Chen, W. *et al.* Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719–736 (2015).

72. Servin, B. & Stephens, M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLOS Genetics* **3**, e114 (2007).

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

74. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

75. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* **9**, 3268 (2018).

76. Smith, A. K. *et al.* Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* **15**, 145 (2014).

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

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

79. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

80. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: Function, expression and evolution. *Nat Rev Genet* **10**, 252–263 (2009).

81. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).

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

83. Keele, G. R. *et al.* Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation. *PLOS Genetics* **16**, e1008537 (2020).

84. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat Genet* **55**, 112–122 (2023).

85. Schwartz, D. Genetic Studies on Mutant Enzymes in Maize. III. Control of Gene Action in the Synthesis of Ph 7.5 Esterase. *Genetics* **47**, 1609–1615 (1962).

86. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**, 752–755 (2002).

87. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).

88. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).

89. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: The teenage years. *Nat Rev Genet* **20**, 631–656 (2019).

90. de Klein, N. *et al.* Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat Genet* **55**, 377–388 (2023).

91. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat Commun* **11**, 4307 (2020).

92. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* **89**, 3010–3014 (1992).

93. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382 (2009).

94. Guo, G. *et al.* Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Developmental Cell* **18**, 675–685 (2010).

95. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160–1167 (2011).

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

97. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041.

98. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).

99. Ding, R. *et al.* scQTLbase: An integrated human single-cell eQTL database. *Nucleic Acids Research* **52**, D1010–D1017 (2024).

100. Wills, Q. F. *et al.* Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* **31**, 748–752 (2013).

101. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet* **50**, 493–497 (2018).

102. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).

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

104. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453–457 (2015).

105. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773–782 (2019).

106. Li, B. *et al.* Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol* **17**, 174 (2016).

107. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).

108. Bryois, J. *et al.* Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat Neurosci* **25**, 1104–1112 (2022).

109. Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* **606**, 120–128 (2022).

110. Soskic, B. *et al.* Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. *Nat Genet* **54**, 817–826 (2022).

111. Murphy, A. E. & Skene, N. G. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat Commun* **13**, 7851 (2022).

112. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat Commun* **11**, 810 (2020).

113. Xu, Z. *et al.* High-throughput single nucleus total RNA sequencing of formalin-fixed paraffin-embedded tissues by snRandom-seq. *Nat Commun* **14**, 2734 (2023).

114. Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Res* **46**, e133 (2018).

115. LaFramboise, T. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research* **37**, 4181–4193 (2009).

116. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype Imputation. *Annu Rev Genomics Hum Genet* **10**, 387–406 (2009).

117. Chat, V., Ferguson, R., Morales, L. & Kirchhoff, T. Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies. *Frontiers in Genetics* **12**, (2022).

118. Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat Genet* **49**, 1747–1751 (2017).

119. Degnan, J. H. *et al.* Genomics and genome-wide association studies: An integrative approach to expression QTL mapping. *Genomics* **92**, 129–133 (2008).

120. Zhao, Y. *et al.* TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine* **19**, 269 (2021).

121. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* **53**, 1290–1299 (2021).

122. Yamamoto, R. *et al.* Tissue-specific impacts of aging and genetics on gene expression patterns in humans. *Nat Commun* **13**, 5803 (2022).

123. Dachet, F. *et al.* Selective time-dependent changes in activity and cell-specific gene expression in human postmortem brain. *Sci Rep* **11**, 6078 (2021).

124. Powell, J. E. *et al.* Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. *PLOS Genetics* **9**, e1003502 (2013).

125. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500–507 (2012).

126. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biology* **20**, 228 (2019).

127. Moffatt, M. F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).

128. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42**, 295–302 (2010).

129. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**, 139–144 (1999).

130. Nica, A. C. *et al.* Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLOS Genetics* **6**, e1000895 (2010).

131. Kraft, P. Curses—Winner's and Otherwise—in Genetic Epidemiology. *Epidemiology* **19**, 649 (2008).

132. He, X. *et al.* Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. *The American Journal of Human Genetics* **92**, 667–680 (2013).

133. Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum Mol Genet* **21**, 2815–2824 (2012).

134. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10**, e1004383 (2014).

135. Guo, H. *et al.* Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum Mol Genet* **24**, 3305–3313 (2015).

136. Hauberg, M. E. *et al.* Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression. *The American Journal of Human Genetics* **100**, 885–894 (2017).

137. Kia, D. A. *et al.* Identification of Candidate Parkinson Disease Genes by Integrating Genome-Wide Association Study, Expression, and Epigenetic Data Sets. *JAMA Neurology* **78**, 464–472 (2021).

138. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).

139. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat Commun* **12**, 764 (2021).

140. Yu, X. *et al.* A suggested shared aetiology of dementia - a colocalization study. *Neurobiology of Aging* **117**, 71–82 (2022).

141. Wang, Y. *et al.* Remdesivir in adults with severe COVID-19: A randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet* **395**, 1569–1578 (2020).

142. Hammond, J. *et al.* Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *N Engl J Med* **386**, 1397–1408 (2022).

143. Sheiner, L. B. & Rubin, D. B. Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics* **57**, 6–15 (1995).

144. Peto, R., Doll, R., Buckley, J. D. & Sporn, M. B. Can dietary beta-carotene materially reduce human cancer rates? *Nature* **290**, 201–208 (1981).

145. Hennekens, C. H. *et al.* Lack of Effect of Long-Term Supplementation with Beta Carotene on the Incidence of Malignant Neoplasms and Cardiovascular Disease. *N Engl J Med* **334**, 1145–1149 (1996).

146. Hooper, L., Ness, A. R. & Smith, G. D. Antioxidant strategy for cardiovascular disease. *The Lancet* **357**, 1705 (2001).

147. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology* **32**, 1–22 (2003).

148. Timpson, N. J. *et al.* C-reactive protein and its role in metabolic syndrome: Mendelian randomisation study. *The Lancet* **366**, 1954–1959 (2005).

149. Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *The American Journal of Human Genetics* **109**, 767–782 (2022).

150. Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nat Hum Behav* **1**, 1–13 (2017).

151. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

152. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* **50**, 693–698 (2018).

153. Wang, L., Gao, B., Fan, Y., Xue, F. & Zhou, X. Mendelian randomization under the omnigenic architecture. *Brief Bioinform* **22**, bbab322 (2021).

154. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* **91**, 444–455 (1996).

155. Hingorani, A. & Humphries, S. Nature's randomised trials. *The Lancet* **366**, 1906–1908 (2005).

156. Burgess, S., Thompson, S. G. & CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol* **40**, 755–764 (2011).

157. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet Epidemiol* **37**, 658–665 (2013).

158. Pierce, B. L. & Burgess, S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology* **178**, 1177–1184 (2013).

159. Baranova, A., Cao, H. & Zhang, F. Causal effect of COVID-19 on Alzheimer's disease: A Mendelian randomization study. *J Med Virol* 10.1002/jmv.28107 (2022) doi:10.1002/jmv.28107.

160. Seyedsalehi, A. *et al.* Educational attainment, structural brain reserve and Alzheimer's disease: A Mendelian randomization analysis. *Brain* **146**, 2059–2074 (2023).

161. Andrews, S. J., Goate, A. & Anstey, K. J. Association between alcohol consumption and Alzheimer's disease: A Mendelian randomization study. *Alzheimers Dement* **16**, 345–353 (2020).

162. Ding, P., Gorenflo, M. P., Zhu, X. & Xu, R. Aspirin Use and Risk of Alzheimer's Disease: A 2-Sample Mendelian Randomization Study. *J Alzheimers Dis* **92**, 989–1000 (2023).

163. Wingo, A. P. *et al.* Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nat Genet* **53**, 143–146 (2021).

164. Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet* **34**, 154–156 (2003).

165. Sabatine Marc S. *et al.* Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *New England Journal of Medicine* **376**, 1713–1722 (2017).

166. Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* **26**, 2333 (2017).

167. Boehm, F. J. & Zhou, X. Statistical methods for Mendelian randomization in genome-wide association studies: A review. *Comput Struct Biotechnol J* **20**, 2338–2351 (2022).

168. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–525 (2015).

169. Qi, G. & Chatterjee, N. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nat Commun* **10**, 1941 (2019).

170. Zhao, J. *et al.* Bayesian weighted Mendelian randomization for causal inference based on summary statistics. *Bioinformatics* **36**, 1501–1508 (2020).

171. Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations: Update for summer 202. (2023) doi:10.12688/wellcomeopenres.15555.3.

172. Porcu, E. *et al.* Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nat Commun* **12**, 5647 (2021).

173. Burgess, S., Swanson, S. A. & Labrecque, J. A. Are Mendelian randomization investigations immune from bias due to reverse causation? *Eur J Epidemiol* **36**, 253–257 (2021).

174. Gaziano, L. *et al.* Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. *Nat Med* **27**, 668–676 (2021).

175. Zhang, H. *et al.* Advances in developing ACE2 derivatives against SARS-CoV-2. *The Lancet Microbe* **4**, e369–e378 (2023).

176. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).

177. Reis Gilmar *et al.* Early Treatment with Pegylated Interferon Lambda for Covid-19. *New England Journal of Medicine* **388**, 518–528 (2023).

178. Smith, A. M. *et al.* Diverse human astrocyte and microglial transcriptional responses to Alzheimer's pathology. *Acta Neuropathol* **143**, 75–91 (2022).

179. Tsartsalis, S. *et al.* A single nuclear transcriptomic characterisation of mechanisms responsible for impaired angiogenesis and blood-brain barrier function in Alzheimer's disease. *Nat Commun* **15**, 2243 (2024).

180. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).

181. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).

182. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

183. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

184. Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Research* **51**, D933–D941 (2023).

185. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

186. Feleke, R. *et al.* Cross-platform transcriptional profiling identifies common and distinct molecular pathologies in Lewy body diseases. *Acta Neuropathol* **142**, 449–474 (2021).

187. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLOS ONE* **13**, e0209648 (2018).

188. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat Med* **26**, 792–802 (2020).

189. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89–94 (2018).

190. Lun, A. T. L. *et al.* EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* **20**, 63 (2019).

191. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).

192. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329–337.e4 (2019).

193. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).

194. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol* **21**, 57 (2020).

195. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

196. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).

197. Cavalcante, R. G. & Sartor, M. A. Annotatr: Genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).

198. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. **lmerTest** Package: Tests in Linear Mixed Effects Models. *J. Stat. Soft.* **82**, (2017).

199. Vuong, Q. H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57**, 307–333 (1989).

200. Lai, Y. A statistical method for the conservative adjustment of false discovery rate (q-value). *BMC Bioinformatics* **18**, 69 (2017).

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

202. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. 2020.08.10.244293 https://www.biorxiv.org/content/10.1101/2020.08.10.244293v1 (2020) doi:10.1101/2020.08.10.244293.

203. Meyer, D., C. Aguiar, V. R., Bitarello, B. D., C. Brandt, D. Y. & Nunes, K. A genomic perspective on HLA evolution. *Immunogenetics* **70**, 5–27 (2018).

204. Bowles, K. R. *et al.* 17q21.31 sub-haplotypes underlying H1-associated risk for Parkinson's disease are associated with LRRC37A/2 expression in astrocytes. *Molecular Neurodegeneration* **17**, 48 (2022).

205. Yavorska, O. O. & Burgess, S. MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology* **46**, 1734–1739 (2017).

206. Robins, C. *et al.* Genetic control of the human brain proteome. *Am J Hum Genet* **108**, 400–410 (2021).

207. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).

208. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res* **45**, D139–D144 (2017).

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

210. Goldenberg, M. M. Multiple Sclerosis Review. *P T* **37**, 175–184 (2012).

211. Poewe, W. *et al.* Parkinson disease. *Nat Rev Dis Primers* **3**, 1–21 (2017).

212. Zhao, L. *et al.* Age-Related Differences in Brain Morphology and the Modifiers in Middle-Aged and Older Adults. *Cerebral Cortex* **29**, 4169–4193 (2019).

213. Morrison, J. H. & Baxter, M. G. The ageing cortical synapse: Hallmarks and implications for cognitive decline. *Nat Rev Neurosci* **13**, 240–250 (2012).

214. Allen, W. E., Blosser, T. R., Sullivan, Z. A., Dulac, C. & Zhuang, X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* **186**, 194–208.e18 (2023).

215. Dutta, D. *et al.* Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood. *Nat Commun* **13**, 4323 (2022).

216. Cuomo, A. S. E. *et al.* Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biology* **22**, 188 (2021).

217. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci Data* **4**, 170112 (2017).

218. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

219. Hauberg, M. E. *et al.* Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat Commun* **11**, 5581 (2020).

220. Sieberts, S. K. *et al.* Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Sci Data* **7**, 340 (2020).

221. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).

222. Xiao, Q. *et al.* Immunosuppression causes dynamic changes in expression QTLs in psoriatic skin. *Nat Commun* **14**, 6268 (2023).

223. Deczkowska, A. *et al.* Disease-Associated Microglia: A Universal Immune Sensor of Neurodegeneration. *Cell* **173**, 1073–1081 (2018).

224. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *Journal of Cell Biology* **217**, 459–472 (2017).

225. Depp, C. *et al.* Myelin dysfunction drives amyloid- deposition in models of Alzheimer's disease. *Nature* **618**, 349–357 (2023).

226. Smajić, S. *et al.* Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state. *Brain* **145**, 964–978 (2022).

227. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).

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

229. Ahadi, S. *et al.* Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nat Med* **26**, 83–90 (2020).

230. Shen, X. *et al.* Nonlinear dynamics of multi-omics profiles during human aging. *Nat Aging* 1–16 (2024) doi:10.1038/s43587-024-00692-2.

231. Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat Commun* **9**, 4285 (2018).

232. D'Antonio, M. *et al.* Fine mapping spatiotemporal mechanisms of genetic variants underlying cardiac traits and disease. *Nat Commun* **14**, 1–18 (2023).

233. Lamontagne, M. *et al.* Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. *Human Molecular Genetics* **27**, 1819–1829 (2018).

234. Etheridge, A. S. *et al.* A New Liver Expression Quantitative Trait Locus Map From 1,183 Individuals Provides Evidence for Novel Expression Quantitative Trait Loci of Drug Response, Metabolic, and Sex-Biased Phenotypes. *Clin Pharmacol Ther* **107**, 1383–1393 (2020).

235. Wu, Y. *et al.* Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution. *Hum Mol Genet* **28**, 4161–4172 (2019).

236. Orozco, L. D. *et al.* Integration of eQTL and a Single-Cell Atlas in the Human Eye Identifies Causal Genes for Age-Related Macular Degeneration. *Cell Reports* **30**, 1246–1259.e6 (2020).

237. Chen, B. Y. *et al.* ColocQuiaL: A QTL-GWAS colocalization pipeline. *Bioinformatics* **38**, 4409–4411 (2022).

238. Schilder, B. M., Humphrey, J. & Raj, T. echolocatoR: An automated end-to-end statistical and functional genomic fine-mapping pipeline. *Bioinformatics* **38**, 536–539 (2022).

239. Natri, H. M. *et al.* Cell-type-specific and disease-associated expression quantitative trait loci in the human lung. *Nat Genet* **56**, 595–604 (2024).

240. Jayaswamy, P. K. *et al.* Implicative role of epidermal growth factor receptor and its associated signaling partners in the pathogenesis of Alzheimer's disease. *Ageing Research Reviews* **83**, 101791 (2023).

241. Xiong, X. *et al.* Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422–4437.e21 (2023).

242. Ando, K. *et al.* Picalm reduction exacerbates tau pathology in a murine tauopathy model. *Acta Neuropathol* **139**, 773–789 (2020).

243. Jochemsen, H. M. *et al.* The association of angiotensin-converting enzyme with biomarkers for Alzheimer's disease. *Alzheimers Res Ther* **6**, 27 (2014).

244. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat Commun* **12**, 1610 (2021).

245. Chauquet, S. *et al.* Association of Antihypertensive Drug Target Genes With Psychiatric Disorders. *JAMA Psychiatry* **78**, 1–9 (2021).

246. Huang, P. *et al.* Deep white matter hyperintensity is associated with the dilation of perivascular space. *J Cereb Blood Flow Metab* **41**, 2370–2380 (2021).

247. Huo, Y. *et al.* Deep white matter hyperintensity is spatially correlated to MRI-visible perivascular spaces in cerebral small vessel disease on 7 Tesla MRI. *Stroke Vasc Neurol* **8**, (2023).

248. Wightman, D. P. *et al.* The genetic overlap between Alzheimer's disease, amyotrophic lateral sclerosis, Lewy body dementia, and Parkinson's disease. *Neurobiology of Aging* **127**, 99–112 (2023).

249. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Science Advances* **6**, eaba2083 (2020).

250. Rocheleau, G. *et al.* A tissue-level phenome-wide network map of colocalized genes and phenotypes in the UK Biobank. *Commun Biol* **5**, 1–9 (2022).

251. Pan, S. *et al.* Brain Catalog: A comprehensive resource for the genetic landscape of brain-related traits. *Nucleic Acids Research* **51**, D835–D844 (2023).

252. Sanderson, E. *et al.* Mendelian randomization. *Nat Rev Methods Primers* **2**, 1–21 (2022).

253. Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L. & Davey Smith, G. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nat Commun* **11**, 185 (2020).

254. Smith-Byrne, K. *et al.* Identifying therapeutic targets for cancer among 2074 circulating proteins and risk of nine cancers. *Nat Commun* **15**, 3621 (2024).

255. Wang, X. *et al.* Integrating genome-wide association study and expression quantitative trait loci data identifies *NEGR1* as a causal risk gene of major depression disorder. *Journal of Affective Disorders* **265**, 679–686 (2020).

256. Singh, K. *et al.* Neural cell adhesion molecule Negr1 deficiency in mouse results in structural brain endophenotypes and behavioral deviations related to psychiatric disorders. *Sci Rep* **9**, 5457 (2019).

257. Iacoangeli, A. *et al.* SCFD1 expression quantitative trait loci in amyotrophic lateral sclerosis are differentially expressed. *Brain Communications* **3**, fcab236 (2021).

258. Cullell, N. *et al.* ICA1L Is Associated with Small Vessel Disease: A Proteome-Wide Association Study in Small Vessel Stroke and Intracerebral Haemorrhage. *Int J Mol Sci* **23**, 3161 (2022).

259. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* **2**, 1135–1148 (2020).

260. Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B. & Hopewell, J. C. Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genetic Epidemiology* **41**, 714–725 (2017).

261. Loda, E. & Balabanov, R. Interferon regulatory factor 1 regulation of oligodendrocyte injury and inflammatory demyelination. *Rev Neurosci* **23**, 145–152 (2012).

262. Ren, Z. *et al.* IRF-1 signaling in central nervous system glial cells regulates inflammatory demyelination. *Journal of Neuroimmunology* **233**, 147–159 (2011).

263. Choi, H.-J., Jeong, Y. J., Kim, J. & Hoe, H.-S. EGFR is a potential dual molecular target for cancer and Alzheimer's disease. *Front Pharmacol* **14**, 1238639 (2023).

264. Mansour, H. M., Fawzy, H. M., El-Khatib, A. S. & Khattab, M. M. Repurposed anticancer epidermal growth factor receptor inhibitors: Mechanisms of neuroprotective effects in Alzheimer's disease. *Neural Regen Res* **17**, 1913–1918 (2022).

265. Farrow, S. L. *et al.* Identification of 27 allele-specific regulatory variants in Parkinson's disease using a massively parallel reporter assay. *npj Parkinsons Dis.* **10**, 1–14 (2024).

266. Fujita, M. *et al.* Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat Genet* 1–10 (2024) doi:10.1038/s41588-024-01685-y.

267. Molecular interactions of complement receptors on B lymphocytes: A CR1/CR2 complex distinct from the CR2/CD19 complex. *J Exp Med* **173**, 1083–1089 (1991).

268. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947 (2020).

269. Chen, X. *et al.* A brain cell atlas integrating single-cell transcriptomes across human brain regions. *Nat Med* **30**, 2679–2691 (2024).

270. Madsen, P. M. *et al.* Oligodendrocytes modulate the immune-inflammatory response in EAE via TNFR2 signaling. *Brain Behav Immun* **84**, 132–146 (2020).

271. Boccazzi, M., Raffaele, S. & Fumagalli, M. Not only myelination: The immune-inflammatory functions of oligodendrocytes. *Neural Regen Res* **17**, 2661–2663 (2022).

272. Shah, A., Kishore, U. & Shastri, A. Complement System in Alzheimer's Disease. *Int J Mol Sci* **22**, 13647 (2021).

273. Van Os, J. & Jones, P. B. Neuroticism as a risk factor for schizophrenia. *Psychol Med* **31**, 1129–1134 (2001).

274. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

275. Moloney, E. B., Moskites, A., Ferrari, E. J., Isacson, O. & Hallett, P. J. The glycoprotein GPNMB is selectively elevated in the substantia nigra of Parkinson's disease patients and increases after lysosomal stress. *Neurobiology of Disease* **120**, 1–11 (2018).

276. Diaz-Ortiz, M. E. *et al.* GPNMB confers risk for Parkinson's disease through interaction with -synuclein. *Science* **377**, eabk0637 (2022).

277. Marzi, S. J. *et al.* A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex. *Nat Neurosci* **21**, 1618–1627 (2018).

278. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).

279. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).

280. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

281. Simone, M. D. *et al.* Comparative Analysis of Commercial Single-Cell RNA Sequencing Technologies. 2024.06.18.599579 https://www.biorxiv.org/content/10.1101/2024.06.18.599579v1 (2024) doi:10.1101/2024.06.18.599579.

282. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends in Genetics* **37**, 109–124 (2021).

283. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet* **55**, 1866–1875 (2023).

284. Cooper, Y. A. *et al.* Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* **377**, eabi8654 (2022).

285. Zhao, S. *et al.* A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat Genet* **55**, 346–354 (2023).

286. Morris, J. A. *et al.* Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* **380**, eadh7699 (2023).

287. Ma, X. R. *et al.* TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. *Nature* **603**, 124–130 (2022).

288. Zuber, V. *et al.* Multi-response Mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes. *Am J Hum Genet* **110**, 1177–1199 (2023).

289. Zhao, J. H. *et al.* Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. *Nat Immunol* **24**, 1540–1551 (2023).

290. VanderWeele, T. J. Mediation Analysis: A Practitioner's Guide. *Annu. Rev. Public Health* **37**, 17–32 (2016).

291. Cashin, A. G., McAuley, J. H., VanderWeele, T. J. & Lee, H. Understanding how health interventions or exposures produce their effects using mediation analysis. *BMJ* **382**, e071757 (2023).

292. Freeman, D. *et al.* The effects of improving sleep on mental health (OASIS): A randomised controlled trial with mediation analysis. *The Lancet Psychiatry* **4**, 749–758 (2017).

293. Carter, A. R. *et al.* Mendelian randomisation for mediation analysis: Current methods and challenges for implementation. *Eur J Epidemiol* **36**, 465–478 (2021).

294. 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).

295. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nat Rev Genet* **17**, 129–145 (2016).

296. Dobbyn, A. *et al.* Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. *The American Journal of Human Genetics* **102**, 1169–1184 (2018).

297. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet* **51**, 1060–1066 (2019).

298. Liu, Z. *et al.* Linking genome structures to functions by simultaneous single-cell Hi-C and RNA-seq. *Science* **380**, 1070–1076 (2023).

299. Hughes, J. R. & Davies, J. O. J. Single-cell technologies meet Hi-C. *Nat Genet* **56**, 1542–1543 (2024).

300. Ray-Jones, H. *et al.* Shared and distinct molecular effects of regulatory genetic variants provide insight into mechanisms of distal enhancer-promoter communication. 2023.08.04.551251 https://www.biorxiv.org/content/10.1101/2023.08.04.551251v3 (2024) doi:10.1101/2023.08.04.551251.

301. Rinaldi, C. & Wood, M. J. A. Antisense oligonucleotides: The next frontier for treatment of neurological disorders. *Nat Rev Neurol* **14**, 9–21 (2018).

302. Scoles, D. R., Minikel, E. V. & Pulst, S. M. Antisense oligonucleotides. *Neurology Genetics* **5**, e323 (2019).

303. Cole, T. A. *et al.* A-Synuclein antisense oligonucleotides as a disease-modifying therapy for Parkinson's disease. *JCI Insight* **6**, e135633.

304. Mummery, C. J. *et al.* Tau-targeting antisense oligonucleotide MAPTRx in mild Alzheimer's disease: A phase 1b, randomized, placebo-controlled trial. *Nat Med* **29**, 1437–1447 (2023).

305. Stöberl, N., Maguire, E., Salis, E., Shaw, B. & Hall-Roberts, H. Human iPSC-derived glia models for the study of neuroinflammation. *Journal of Neuroinflammation* **20**, 231 (2023).

306. Tasfaout, H. *et al.* Antisense oligonucleotide-mediated Dnm2 knockdown prevents and reverts myotubular myopathy in mice. *Nat Commun* **8**, 15661 (2017).

307. Rooney, M. R. *et al.* Plasma proteomic comparisons change as coverage expands for SomaLogic and Olink. 2024.07.11.24310161 https://www.medrxiv.org/content/10.1101/2024.07.11.24310161v1 (2024) doi:10.1101/2024.07.11.24310161.

308. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat Med* **30**, 2489–2498 (2024).