



OLLSCOIL NA GAILLIMHE

UNIVERSITY OF GALWAY

Investigating the cell-type specific biology of schizophrenia using multi-omic data

By

Rebecca Mahoney, M.Sc.

A thesis submitted for the Degree of Doctor of Philosophy to the
School of Biological and Chemical Sciences,
University of Galway, Galway, Ireland.

Supervisor: Prof. Derek Morris
School of Biological and Chemical Sciences,
University of Galway, Galway, Ireland.

Co-Supervisor: Prof. Cathal Seoighe
School of Mathematical and Statistical Sciences,
University of Galway, Galway, Ireland.

Submitted January 2024

Contents

Declaration	vi
Statement of Contribution	vii
Funding	viii
Acknowledgements	ix
Abstract	xii
List of Tables	xiii
List of Supplementary Tables	xv
List of Figures	xvi
List of abbreviations used in this thesis	xix
1 Introduction	1
1.1 Schizophrenia as a Psychiatric Disorder	1
1.1.1 Clinical Description of Schizophrenia	1
1.1.2 Epidemiology of Schizophrenia	3
1.1.3 Environmental Risk Factors for Schizophrenia	3
1.1.4 Neuroimaging in Schizophrenia	4
1.1.5 Pathophysiology of Schizophrenia	5
1.1.6 Heritability of Schizophrenia	6
1.2 Genetic Studies of Schizophrenia	6
1.2.1 Candidate Gene and Linkage Studies	7
1.2.2 Common Variation in Schizophrenia	8
1.2.2.1 Genome-Wide Association Studies	8
1.2.2.2 Application of GWAS Results	10
1.2.2.2.1 Functional Annotation and Analysis of GWAS Findings	10

1.2.2.2.2	Trait Risk Prediction	12
1.2.3	Rare variation in Schizophrenia	13
1.2.3.1	Copy Number Variation	13
1.2.3.2	Exome Sequencing Studies	14
1.3	Functional Genomics Studies of Schizophrenia	17
1.3.1	Mouse Models of Schizophrenia	17
1.3.2	Postmortem Transcriptomic Studies in Schizophrenia	18
1.3.2.1	Bulk Transcriptomics	18
1.3.2.2	Cell-type Specific Analysis	21
1.3.3	Postmortem Studies of the Epigenome in Schizophrenia	24
1.3.3.1	Chromatin Accessibility Analysis	24
1.3.3.2	Analysis of the Methylome in Schizophrenia	25
1.3.3.3	Analysis of Histone Modifications and the 3D Genome in Schizophrenia	26
1.3.4	Issues with Postmortem Brain Studies	27
1.3.5	Integrative Multiomic Studies in Schizophrenia	28
1.4	Cell-type Deconvolution	29
1.4.1	Principles and Methods of Cell-type Deconvolution	30
1.4.2	Uses of Cell-type Deconvolution	32
1.4.2.1	Cell-type Deconvolution in Other Tissues	32
1.4.2.2	Cell-type Deconvolution in the Brain	33
1.5	Cell-type Expression Imputation	34
1.6	Limitations and Challenges in this Subject Area	35
1.6.1	Cell-type Specific Analysis in the Study of Schizophrenia	35
1.6.2	Analysis of Gene-sets Generated Using Functional Genomics Data	36
1.7	Aims and Objectives of this Thesis	36
1.7.1	Chapter 2	36
1.7.2	Chapter 3	36
1.7.3	Chapter 4	37
2	Using Cell-type Deconvolution to Explore the Role of Different Cell- types in Schizophrenia	39
2.1	Abstract	40
2.2	Introduction	41
2.3	Methods	45
2.3.1	Dataset Description and Processing	45
2.3.2	Cell-type Deconvolution	45
2.3.3	Polygenic Risk Scoring	46
2.3.4	Gene Expression Imputation	46

2.3.5	Differential Gene Expression Analysis	47
2.3.6	Gene-set Analysis	47
2.4	Results	49
2.4.1	Cell-type Deconvolution	49
2.4.2	Association of Cell-type Proportions with A PRS	51
2.4.3	Gene Expression Imputation	52
2.4.4	Differential Gene Expression Analysis	52
2.4.5	Gene-set Analysis	53
2.5	Discussion	63
3	Analysis of Genes Under the Control of <i>MEF2C</i> using Functional Ge-	
	nomics Data	66
3.1	Abstract	67
3.2	Introduction	68
3.3	Methods	72
3.3.1	Datasets	72
3.3.1.1	Gene-sets	72
3.3.1.2	Single-cell Data	72
3.3.2	Enrichment of <i>MEF2C</i> gene-sets in single-cell data	73
3.3.3	Enrichment of <i>MEF2C</i> Gene-sets in DEGs from Schizophrenia Single-	
	nuclei Analysis	74
3.3.4	Co-expression Network Construction	74
3.3.5	Co-expression Enrichment Analysis	75
3.4	Results	76
3.4.1	Cell-type Enrichment in the Developmental dataset	76
3.4.2	Cell-type Enrichment in the SCZ dataset	81
3.4.3	Enrichment of <i>MEF2C</i> Gene-sets in DEGs from	
	Schizophrenia Single-nuclei Analysis	88
3.4.4	Co-expression fold enrichment	89
3.5	Discussion	91
4	Investigation of the Chromatin Accessibility and Transcriptional Land-	
	scapes of Schizophrenia at the Cell-type Level	94
4.1	Preamble to Chapter 4	95
4.2	Abstract	96
4.3	Introduction	97
4.4	Methods	101
4.4.1	Description of Samples	101

4.4.2	Fluorescence Activated Nuclear Sorting (FANS) of Four Different Cell-types	101
4.4.3	RNA-seq Data Generation	101
4.4.4	ATAC-seq Data Generation	102
4.4.5	Alignment of Raw Sequencing Files	102
4.4.6	Peak Calling	103
4.4.7	Quality Control	103
4.4.8	Overlap of ATAC-seq peaks with Existing Annotation	104
4.4.9	Differential Chromatin Accessibility Analysis	104
4.4.10	Differential Gene Expression Analysis	105
4.4.11	Differential Transcript Analysis	105
4.4.12	Gene Set Enrichment Analysis	106
4.4.13	Gene-set Analysis of Differentially Expressed Genes	106
4.4.14	Disease Heritability of Open Chromatin Regions using LD-score Partitioned Heritability	107
4.4.15	Prediction of Enhancer-Gene Interactions	107
4.4.16	Identification of Genes Implicated by Dysregulated ATAC-seq Peaks	108
4.4.17	Cell-type Deconvolution and Gene Expression Imputation	108
	4.4.17.1 Using FANS Data as a reference	108
	4.4.17.2 Using Single-cell Data as a Reference	109
4.4.18	eQTL Detection in Expression Data Imputed from Bulk RNA-seq .	110
4.4.19	Colocalization Analysis	111
4.4.20	Integration of ATAC-seq and RNA-seq Data	111
	4.4.20.1 Canonical Correlation Analysis	111
	4.4.20.2 Multi-omic Intergation with SNFtool	111
4.5	Results	113
4.5.1	Dataset Description and Quality Control	113
4.5.2	Annotation of Open Chromatin Regions	114
4.5.3	Differential Accessibility Analysis	119
4.5.4	Gene-set Enrichment Analysis of Open Chromatin Regions	120
4.5.5	Disease Heritability of Open Chromatin Regions	123
4.5.6	Identification of Enhancer-Promoter Interactions and their Relationship with Differentially Accessible Open Chromatin Regions . .	127
4.5.7	Differential Gene Expression Analysis	130
4.5.8	Gene-set Enrichment Analysis of Differentially Expressed Genes . .	132
4.5.9	Differential Transcript Expression Analysis	135
4.5.10	Estimation of Cell-type Proportions using Cell-type Deconvolution .	138
4.5.11	Detection of Cell-type Specific eQTLs	140
4.5.12	Integration of ATAC-seq and RNA-seq Data	146

4.5.12.1	Canonical Correlation Analysis	146
4.5.12.2	Integration of Multi-omic Data with SNFtool	146
4.6	Discussion	148
5	Discussion	154
5.1	Summary of Main Findings	154
5.1.1	Chapter 2	154
5.1.2	Chapter 3	155
5.1.3	Chapter 4	155
5.2	The Use of Cell-type Deconvolution to Gain New Insights from Existing Bulk Expression Data	157
5.3	The Use of Multi-omic Datasets Derived from Postmortem Brain Tissue in the Study of Neuropsychiatric Disease	159
5.4	The Importance of Cell-type Specific Studies in Functional Genomics . . .	162
5.5	Concluding Remarks	165
	Bibliography	167
	Appendices	213
	List of Awards	213
	List of Presentations	214
	List of Oral Presentations	214
	List of Poster Presentations	214
	Manuscripts arising from this thesis	215

Declaration

I declare that this thesis has not been submitted as an exercise at this or any other university.

I declare that this thesis is entirely my own work, except where otherwise stated.

Signed:

Rebecca Mahoney

Statement of Contribution

I performed the analysis in Chapters 2 and 3 using publicly available data. All contributions to the study described in Chapter 4 is listed at the beginning of that chapter.

Funding

This research was funded by Science Foundation Ireland through the SFI Centre for Research Training in Genomics Data Science under Grant number 18/CRT/6214.

Acknowledgements

I would like to sincerely thank my primary supervisor, Prof. Derek Morris, for his support and guidance throughout my PhD. I learned a lot under your supervision. I would also like to extend my thanks to my co-supervisor Prof. Cathal Seoighe for your support and insights. Thank you for being so generous with your time throughout my PhD.

I would like to sincerely thank Prof. Panos Roussos for giving me the opportunity to collaborate with your group and for your support and guidance throughout our collaboration. I would like to express my thanks to Dr. Jaroslav Bendl for his guidance and expertise throughout this collaboration. I have truly learned a lot as a result of this experience. I would also like to extend my thanks to all members of the Centre for Disease Neurogenomics for welcoming me during my time there. In particular I would like to thank Dr. Gabriel Hoffman, Dr. Biao Zeng and Dr. Roman Kosoy for their advice on particular aspects of the analysis in Chapter 4 of this thesis.

I would also like to thank the members of Derek's lab, Cathal's lab and Cohort 1 of the CRT in Genomics Data Science. Thanks to all of you for making this a really enjoyable and supportive environment to work in.

Last but most certainly not least, I would like to thank my Mom, and the girls, Indie, Allie and Mollie for all their love, support and advice for which I am truly grateful. I would never have been able to do this without you.

Abstract

Schizophrenia (SCZ) is a complex neuropsychiatric disorder which affects approximately 1% of the population. It carries a high societal burden accounting for 1.7% of disease burden across all possible diseases or disorders despite its low global prevalence. It is clinically characterised by a combination of symptoms including psychotic episodes, cognitive deficits and social withdrawal. Yet, its full etiology is still unknown and is thought to be impacted by a mixture of genetic and environmental factors and begin in neurodevelopment. SCZ is highly heritable with the heritability of SCZ thought to be approximately 79%. Genetic studies such as genome-wide association studies (GWAS) and exome sequencing studies had uncovered a number of genes thought to convey increased risk to SCZ. However, the functional impact of genetic mutation in SCZ must also be considered. Functional genomics studies have aimed to bridge the gap between genomic variation and phenotype by characterising the underlying neurobiology of the disorder. Extensive changes have been observed at the transcriptomic and epigenomic levels in SCZ using postmortem brain tissue samples. Yet the vast majority of these studies use bulk tissue samples which only shows overall trends in a tissue rather than cell-type specific effects. Functional changes are thought to vary widely between cell-type with excitatory neurons primarily implicated in SCZ thus far.

This thesis comprises three studies which all aim to explore which cell-types are impacted in SCZ and characterise these cell-type specific changes. Chapter 2 utilised cell-type deconvolution to uncover new insights from published bulk gene expression data isolated from SCZ cases and controls. Chapter 3 investigated the cell-types affected by sets of differentially expressed genes derived from knockout models of *MEF2C*, a gene that has been implicated in a variety of neurodevelopmental disorders including SCZ and autism spectrum disorder (ASD). My final study which is outlined in Chapter 4 of this thesis explored the transcriptomic and chromatin accessibility landscapes in SCZ at the cell-type level.

Chapter 2 of this thesis used cell-type deconvolution and gene expression imputation methods with published bulk RNA-seq data from SCZ cases and controls in an effort to gain cell-specific insights from bulk data. We integrated the published bulk data with

single-nucleus data to estimate the proportions of ten cell-types in the bulk data. Our overall proportion estimates per cell-type were not consistent with previous studies and we observed a reduction in neuronal proportion estimates compared to previous studies. However, we did detect an increase in the proportion of excitatory neurons and inhibitory neurons in controls compared to SCZ cases. The increase in excitatory neuronal populations is consistent with previous work by the PsychENCODE consortium suggesting that despite discrepancies in our overall estimates we may be observing similar trends. We also used our estimated proportions to impute gene expression for each cell-type using the bulk expression data. We generated gene-sets of differentially expressed genes (DEGs) per cell-type which were used to conduct gene-set analysis. Biological processes that were found to be enriched for neuronal DEGs included neuronal development, neuron projection and NCAM1 interactions. Targeted analysis of the neuronal DEGs using the SynGO ontology resource implicated synaptic locations such as the postsynaptic membrane and endosome and the presynaptic vesicle with different locations implicated between cell-types. Despite these results, our estimated proportions do differ from previous deconvolution studies. This raised questions as to the robustness of our results but did suggest that this type of analysis may prove useful for gaining new insights from bulk data and was used to better effect in Chapter 4 of this thesis.

MEF2C is a transcription factor that has been implicated in a number of neurodevelopmental disorders including SCZ and ASD. Previous studies have generated gene-sets derived from DEGs detected in murine knockout models of *MEF2C*. We used these gene-sets in conjunction with data from snRNAseq studies from human postmortem brain tissue to investigate which cell-types *MEF2C* knockout is likely to affect. We primarily implicated neuronal cell-types which is consistent with previous studies that used mouse snRNAseq data but we also implicated oligodendrocyte precursor cells (OPCs). This highlights the benefit of using human brain data for this analysis. We also explored which cell-types would be implicated when the co-expression of the genes in each gene-set was considered. We detected a higher co-expression fold enrichment of these gene-sets in excitatory neurons. We also observed a difference in fold enrichment at different gestational timepoints using snRNA-seq isolated from the fetal brain.

My final study which is described in Chapter 4 of this thesis involved cell-type specific multi-omic analysis. In this study we have isolated nuclei from two neuronal (GABAergic and glutamatergic neurons) and two non-neuronal (Oligodendrocyte and Microglia/Astrocytes) cell-types from the dorsolateral prefrontal cortex (DLPFC) in SCZ cases and controls in order to characterise cell-type-specific changes at both the transcriptomic and the epigenomic levels using RNA-seq and ATAC-seq profiling. We have expanded the current repertoire of open chromatin regions (OCRs) in the human brain by detecting many

novel OCRs. We also explored disease-associated changes in chromatin accessibility and gene expression. The most extensive dysregulation in chromatin accessibility was found in oligodendrocytes followed by microglia/astrocytes. Gene-set enrichment (GSEA) analysis of these OCRs implicated processes such as glycosylation and apoptosis while common variants associated with SCZ were enriched in disease-associated GABAergic and oligodendrocyte OCRs. The most extensive disease-associated changes in gene expression were also observed in oligodendrocytes but we detected abundant disease-associated changes across all cell-types. In addition, the majority of DEGs were cell-type specific. Biological processes enriched for our neuronal DEG sets included regulation of the synaptic vesicle cycle and synaptic plasticity while non-neuronal cell-types implicated axon ensheathment and sterol biosynthesis, which indicated dysregulation related to the myelin sheath. Given the limited research on SCZ at the transcript level, our study aimed to address this gap by prioritising genes with cell-type-specific differentially expressed transcripts. This often revealed dysregulation of transcripts for genes that did not individually reach statistical significance including *CACNA1C*, a known SCZ risk gene. We also identified cell-type-specific enhancer–promoter interactions by integrating our OCRs with cell-specific Hi-C data. We then linked disease-associated chromatin accessibility changes with SCZ GWAS signal and predicted target genes using our enhancer–promoter interactions. Finally, we leveraged our cell-type specific gene expression profiles to enable a QTL analysis by using them to identify cell-type proportions in bulk RNA-seq data. We then used those proportions to impute expression for 848 bulk RNA-seq profiles which allowed for eQTL detection and colocalization analysis resulting in cell-type specific eQTLs which colocalized with SCZ risk loci.

Overall, this thesis uses multi-omic data to explore changes in gene expression and chromatin accessibility at the cell-type specific level in SCZ. By exploring these changes we aimed to identify new insights into the neurobiology of SCZ which is vital in order to one day provide new drug targets and treatment options.

List of Tables

1.1	Diagnostic criteria for schizophrenia	1
1.2	Features of bulk, cell-specific and single-cell sequencing studies	22
1.3	Description of the most commonly used deconvolution methods	32
2.1	Results from testing the cell-type proportions for differences between cases and controls using a two-sample t-test. P-values were corrected for multiple testing using the Bonferroni method.	51
2.2	Results from testing the association of cell-type proportions with a PRS using a linear model. P-values were corrected for multiple testing using the Bonferroni method.	52
2.3	Number of DEGs in the imputed expression matrices per cell-type	53
2.4	Top 5 GO terms and top 5 pathways for Excitatory Neurons	54
2.5	Top 5 GO terms and top 5 pathways for IN-MGE	54
2.6	Top 5 GO terms and top 5 pathways for IN-CGE	55
2.7	Top 5 GO terms and top 5 pathways for Oligodendrocytes	55
2.8	Top 5 GO terms and top 5 pathways for OPCs	56
2.9	Top 5 GO terms and top 5 pathways for Microglia	56
2.10	Top 5 GO terms and top 5 pathways for Astrocytes	57
2.11	Top 5 GO terms and top 5 pathways for Endothelial cells	57
2.12	Top 5 GO terms and top 5 pathways for Pericytes	58
2.13	Top 5 GO terms and top 5 pathways for VSMC	58
2.14	Results from MAGMA gene-set analysis of the DEGs per cell-type using GWAS data for SCZ, IQ and EA.	62
3.1	EWCE results using the KO_hom gene-set and the 15 cell-types from the developmental dataset	78
3.2	EWCE results using the KO_hom.down gene-set and the 15 cell-types from the developmental dataset	79
3.3	EWCE results using the KO_hom.up gene-set and the 15 cell-types from the developmental dataset	80

3.4	EWCE results using the KO_het gene-set and the 15 cell-types from the developmental dataset	81
3.5	EWCE results using the KO_hom gene-set and the 24 cell-types from the SCZ dataset	84
3.6	EWCE results using the KO_hom_down gene-set and the 24 cell-types from the SCZ dataset	85
3.7	EWCE results using the KO_hom_up gene-set and the 24 cell-types from the SCZ dataset.	86
3.8	EWCE results using the KO_het gene-set and the 24 cell-types from the SCZ dataset.	87
4.1	Results from MAGMA gene-set analysis of the DEGs per cell-type using GWAS data for SCZ, IQ and EA.	135
4.2	Number of genes containing a significant transcript per cell-type (BH corrected RE2C P-value < 0.05)	136

List of Supplementary Tables

The following supplementary tables are available at:

https://drive.google.com/drive/folders/11R3Ccix_ccf4pWhXY6AjqtoLdI0Qym86?usp=sharing

2.1 Table of DEGs per cell-type following differential gene expression analysis

2.2 Table of enriched GO terms from overrepresentation analysis

2.3 Table of enriched pathways from overrepresentation analysis

4.1 Table of demographic information for the cohort profiled in Chapter 4

4.2 Table of differentially accessible OCRs per cell-type

4.3 Table of E-P links predicted by ABC analysis

4.4 Table of differentially expressed gene per cell-type

4.5 Table of differentially expressed transcripts per cell-type

4.6 Table of genome-wide significant eQTLs per cell-type

List of Figures

- 1.1 Manhattan plot representing the latest schizophrenia GWAS 10
- 1.2 Graphical illustration depicting the convergence of schizophrenia risk genes
on common functions 16
- 1.3 Graphical illustration single-cell isolation techniques 23
- 1.4 Schematic of cell-type deconvolution 31

- 2.1 Boxplots showing the distribution of the proportion of each cell-type. . . . 49
- 2.2 Boxplots showing the difference in proportion between cases and controls
for four cell-types. 50
- 2.3 Sunburst plot depicting synaptic locations implicated by the DEGs from
excitatory neurons using the SynGO ontology resource. This plot features
the synapse at the center, pre- and post-synaptic locations in the first
ring, and child terms in subsequent rings. The color scheme in the legend
indicates the number of genes within each term. 59
- 2.4 Sunburst plot depicting synaptic locations implicated by the DEGs from
IN-MGE using the SynGO ontology resource. This plot features the synapse
at the center, pre- and post-synaptic locations in the first ring, and child
terms in subsequent rings. The color scheme in the legend indicates the
number of genes within each term. 60
- 2.5 Sunburst plot depicting synaptic locations implicated by the DEGs from
IN-CGE using the SynGO ontology resource. This plot features the synapse
at the center, pre- and post-synaptic locations in the first ring, and child
terms in subsequent rings. The color scheme in the legend indicates the
number of genes within each term. 61

- 3.1 Heatmap displaying the differing cell-type enrichments in the developmen-
tal dataset for each tested gene-set using the log-fold change statistic cal-
culated by the EWCE package. 77
- 3.2 Heatmap displaying the differing cell-type enrichments in the SCZ dataset
for each tested gene-set using the log-fold change statistic calculated by the
EWCE package. 82

3.3	Heatmap displaying the differing cell-type enrichments in the DEGs for each tested gene-set using the p-value calculated by a hypergeometric test.	88
3.4	Co-expression fold enrichment of <i>MEF2C</i> gene-sets in six brain cell types	89
3.5	Coexpression fold enrichment of <i>MEF2C</i> gene-sets in three brain cell types	90
4.1	Schematic outlining the overall study design	113
4.2	t-SNE clustering of samples using chromatin accessibility and gene expression data	114
4.3	Proportion of OCRs corresponding to different genomic annotations	115
4.4	Proportion of known and novel OCRs	115
4.5	Heatmaps visualising the overlap between the OCRs from this study and various reference studies using Jaccard index.	119
4.6	Heatmap displaying the total number of differentially accessible OCRs per cell-type	120
4.7	Heatmap visualising GSEA results for differentially accessible OCRs	122
4.8	Sunburst plots depicting synaptic locations implicated by the GABAergic OCRs.	123
4.9	Heatmap for p-value enrichment of OCRs in common variants	124
4.10	Heatmap for p-value enrichment of differentially accessible OCRs in common variants	125
4.11	Heritability coefficients for various traits for all OCRs in a given cell type	125
4.12	Heritability coefficients for SCZ for differentially accessible OCRs in a given cell type	126
4.13	Histogram of the number of OCR _{ABC} linked per gene	127
4.14	Histogram of the number of genes linked per OCR _{ABC}	128
4.15	Schematic outlining the strategy for linking risk variants associated with SCZ to differentially accessible OCRs and their respective causal genes in a cell-specific manner.	129
4.16	Genome browser visualisation of E-P link and ATAC-seq data for <i>ZNF281</i>	130
4.17	Heatmap summarizing the number of DEGs per cell type	131
4.18	Venn diagram showing overlap of DEGs between cell types	132
4.19	Heatmap visualisation of GSEA results for DEGs using MSigDB.	133
4.20	Sunburst plots depicting results of SynGO analysis	134
4.21	Graphical illustration of transcript expression for (a) <i>KMT5A</i> and (b) <i>CACNA1C</i>	137
4.22	Violin plot showing the distribution of estimated proportions per cell-type.	139
4.23	Spearman correlation coefficient between mean gene expression values of FANS (Ref.) data and imputed expression profiles (Imp.)	140
4.24	Bar chart showing the number of lead eQTLs per cell-type	141

4.25	Heatmap for p-value enrichment of eQTLs in common variants associated with a variety of traits using LDSC	142
4.26	Upset plot showing the number of colocalized eQTLs per cell-type	143
4.27	Bar chart showing the number of eQTLs which colocalized with SCZ risk loci in the imputed, FANS and bulk data	144
4.28	Colocalization of signals from eQTL and SCZ GWAS for SEPT3 in Microglia/Astrocytes	145
4.29	Plot showing the shared correlation structures using canonical correlation vectors (CCA1 vs CCA2) between ATAC-seq and RNA-seq profiles per cell-type. Plot points annotated as gene refers to RNA-seq data while points denoted as OCR refer to ATAC-seq data.	146

List of abbreviations used in this thesis

Abbreviations	Description
ABC	Activity by Contact
AD	Alzheimer's disease
ALS	Amyotrophic lateral sclerosis
ASD	Autism spectrum disorder
BD	Bipolar disorder
BH	Benjamini-Hochberg
BIC	Bayesian Information Criterion
BOCA	Brain Open Chromatin Atlas
caQTL	Chromatin accessibility quantitative trait loci
CCA	Canonical Correlation Analysis
CGE	Caudal ganglionic eminence derived
cKO	Conditional Knockout
CMC	CommonMind Consortium
CNS	Central nervous system
CNV	Copy number variant
CPM	Counts per million
CTD	CellTypeDatasets
CTS	Cell-type specific
DEGs	Differentially expressed genes
DETs	Differentially expressed transcripts
DGE	Differential gene expression
DLPFC	Dorsolateral prefrontal cortex
DNA	Deoxyribonucleic acid
DSM	Diagnostic and statistical manual of mental disorders
DZ	Dizygotic
EA	Educational attainment
EB	Elution buffer
ENCODE	Encyclopedia of DNA Elements

Endo	Endothelial cells
EN	Excitatory Neurons
E-P	Enhancer-promoter
ER	Endoplasmic reticulum
eQTL	Expression quantitative trait loci
EWCE	Expression weighted cell type enrichment
FACS	Fluorescence activated cell sorting
FANS	Fluorescence activated nuclei sorting
FDR	False discovery rate
FMRP	Fragile X mental retardation protein
fQTL	Cell fraction quantitative trait loci
FRiP	Fraction of reads in peaks
GABA	Gamma-aminobutyric acid
GEO	Gene expression omnibus
GLU	Glutamatergic
GO	Gene Ontology
GSA	Gene-set analysis
GSEA	Gene-set enrichment analysis
GW	Gestation Week
GWAS	Genome wide association study
GTE _x	Genotype-Tissue Expression
G _x E	Genotype x environment effects
HWE	Hardy-Weinberg Equilibrium
IN-CGE	Caudal ganglionic eminence derived inhibitory neurons
IN-MGE	Medial ganglionic eminence derived inhibitory neurons
IN	Inhibitory neurons
IQ	Intelligence Quotient
iPSCs	Induced pluripotent stem cells
IPCs	Intermediate progenitor cells
isoQTL	Isoform quantitative trait loci
LD	Linkage disequilibrium
LDSC	Linkage disequilibrium score regression
LOF	Loss of function
MAF	Minor allele frequency
MCMC	Markov chain Monte Carlo
MEF2c	Myocyte-specific enhancer factor 2C
MGAS	Microglia/Astrocytes

MGE	Medial ganglionic eminence derived
MIA	Maternal immune activation
molQTL	Molecular quantitative trait loci
mQTL	Methylation quantitative trait loci
MT	Mitochondrial
MZ	Monozygotic
NMDA	N-methyl-D-aspartate
NPC	Neural progenitor cells
OCRs	Open chromatin regions
OLIG	Oligodendrocyte
OPCs	Oligodendrocyte precursor cells
PBC	Polymerase chain reaction bottle neck coefficient
PCR	Polymerase chain reaction
PGC	Psychiatric Genomics Consortium
PFC	Prefrontal cortex
PRS	Polygenic risk score
QC	Quality control
QTL	Quantitative trait loci
REMC	Roadmap Epigenomics Consortium
RG	Radial glia
RIN	RNA integrity number
RNA	Ribonucleic Acid
RSC	Relative strand cross-correlation coefficient
SC	Single-cell
SCHEMA	Schizophrenia Exome Sequencing Meta-Analysis
scRNAseq	Single-cell RNAseq
SCZ	Schizophrenia
SNCG	Gamma-synuclein
SNF	Similarity network fusion
SNP	Single nucleotide polymorphism
snRNAseq	Single-nucleus RNAseq
SST	Somatostatin
SV	Synaptic vesicle
TADs	Topologically associated domains
TCGA	The Cancer Genome Atlas
THS-seq	Transposome hypersensitive site sequencing
TMM	Trimmed Mean of M-values

TPM	Transcripts per million
TSS	Transcription start site
TWAS	Transcriptome wide association study
UPS	Ubiquitin protease system
VCF	Variant call format
VLMC	Vascular and Leptomeningeal Cell
VSMC	Vascular smooth muscle cells
WES	Whole exome sequencing
WGNCA	Weighted gene correlation network association
WGS	Whole genome sequencing

Chapter 1

Introduction

1.1 Schizophrenia as a Psychiatric Disorder

1.1.1 Clinical Description of Schizophrenia

Schizophrenia is a complex neuropsychiatric disorder which is both highly heritable and can be influenced by a wide variety of environmental factors (Kahn et al., 2015). It affects 20.9 million individuals globally and reduces life expectancy by 10-20 years on average (Charlson et al., 2018; Orrico-Sáet al., 2020). The complex nature of the disorder combined with its varying symptoms can make it difficult to treat. Affected individuals experience distinct symptoms but the combination of symptoms can vary and are the current basis for diagnosis (Kahn et al., 2015). Symptoms are generally divided into three distinct categories known as positive symptoms, negative symptoms and cognitive deficits and a shortened version of the official diagnostic criteria is available in Table 1.1.

Table 1.1: Diagnostic criteria for schizophrenia.

Adapted from the Diagnostic and statistical manual of mental disorders (DSM-5-TR), (American Psychiatric Association, 2022).

Criteria	Symptoms
Characteristic Symptoms	Two or more of the following: <ul style="list-style-type: none">- Delusions- Hallucinations- Disorganised speech- Grossly disorganised or catatonic behaviour- Negative symptoms
Social/Occupational Dysfunction	Decreased level of functioning in areas such as work, interpersonal relations, or self-care
Duration	Continuous signs of the disturbance persist for at least 6 months

Positive symptoms are behaviours and thoughts that are not usually observed in unaffected individuals. They include hallucinations and delusions that cause the affected individuals to lose touch with reality and are often the most recognisable or publicly known symptoms (Kahn et al., 2015). Hallucinations are perceptions created without the

presence of external stimuli resulting in affected individuals perceiving situations that are not actually occurring while delusions refer to beliefs which are held despite the absence of any evidence to support them. It is these symptoms that are managed using antipsychotic drugs which block dopamine receptors. Despite the discovery of anti-psychotics having occurred ~70 years ago, they remain the main course of treatment with development of other antipsychotic agents which act on other brain-related pathways such as the cholinergic and glutamatergic pathways proving unsuccessful (Correll et al., 2022a). While new treatments using muscarinic and trace amine-associated receptor 1 agonists have been successful in small trial studies, there remains a limited number of treatment options (Kantrowitz et al., 2023). The lack of progress in the development of antipsychotics also means that currently only the effects of positive symptoms can be ameliorated in affected individuals while approximately 10%-30% do not respond to any available treatment (Taylor and Jauhar, 2019).

Negative symptoms consist of social withdrawal and absence of motivation for daily activities and are a major contributor to poor treatment outcomes (Kahn et al., 2015). While positive symptoms are largely evident during acute phases of the disorder, negative symptoms are likely to persist throughout the life of an individual. They are also known as the first most commonly experienced symptom in schizophrenia, occurring in 73% during the “prodromal” phase of the disorder which occurs before the first psychotic episode (Correll and Schooler, 2020). They can be categorised into two major categories which are comprised of five distinct symptoms. Diminished symptoms include blunted affect which refers to reduced expression of emotion and alogia which refers to the reduction in the amount of words spoken. Negative symptoms include avolition which results in reduced participation in activities due to lack of motivation, asociality relating to reduced social interaction and anhedonia, the reduced experience of pleasure in an activity (Correll and Schooler, 2020).

Cognitive deficits include memory problems, learning difficulties and attention deficits and occur in 80% of individuals (McCutcheon et al., 2023). While distinct from negative symptoms, it is thought that some cognitive deficits may in turn lead to the development of negative symptoms. For example, a reduction in words spoken and reduced social interaction could develop as a result of memory problems and attention deficits (McCutcheon et al., 2023). While the initial focus was predominantly on the positive symptoms associated with schizophrenia, a move towards understanding the cognitive impairments has been observed (Kahn et al., 2015) as it is these symptoms that manifest first up to 10 years before the first psychotic episode and significantly limit the ability of individuals to live normal lives as unlike positive symptoms they cannot be treated. Overall, while treatment of positive symptoms can have varying degrees of success, the poor ability to

treat all associated symptoms results in only 13.5 % of individuals achieving full recovery criteria which includes social and clinical improvement with sustained improvement in one of these categories for at least a year (Huxley et al., 2021).

1.1.2 Epidemiology of Schizophrenia

Given the incredibly negative impact that schizophrenia has on affected individuals, efforts have been made to identify its incidence and causes. It currently affects 20.9 million people worldwide. In addition, up to 1% of the worlds population are believed to be at risk of schizophrenia. One study suggests that a given individual has a 0.7% risk of developing the disorder (Owen et al., 2016) while a detailed study in the Finnish population estimates a 0.87% risk suggesting geographical differences (Jääskeläinen et al., 2015). The latest study of point prevalence for schizophrenia takes its data from 2016 when 0.28% of the worlds population were believed to suffer from the disorder (Charlson et al., 2018). Disease burden refers to the impact of a health problem and can be measured by cost, mortality etc., and is quantified by measuring the number of years lost due to disability. Globally, schizophrenia accounts for 1.7% of disease burden across all possible diseases or disorders, which is quite large given the low global prevalence of the disorder (Charlson et al., 2018).

Disease onset typically occurs in late adolescence with the majority of cases diagnosed before the age of 45. Reports are conflicting regarding the difference in prevalence of schizophrenia between males and females with studies reporting a slightly increased risk in males or no differences (Ochoa et al., 2012; Charlson et al., 2018). Differences in age of onset depending on sex are observed with men developing schizophrenia earlier at 20-24 years of age in contrast to women between 25-29 years of age (Ochoa et al., 2012).

1.1.3 Environmental Risk Factors for Schizophrenia

Environmental factors such as drug use, early life experience and prenatal factors have been thought to be associated with schizophrenia (Kahn et al., 2015). Overall, environmental factors are thought to account for 15-40% of the overall risk of developing schizophrenia (Robinson and Bergen, 2021). Prenatal factors such as maternal immune activation (MIA), paternal age, birth complications and season of birth have been implicated in schizophrenia (Kahn et al., 2015). Some of these factors may also be influenced by genetic background such as paternal age which may be associated with schizophrenia due to epigenetic changes, increased *de novo* mutations or selection (Khachadourian et al., 2021). MIA is linked to schizophrenia risk by epidemiological studies which suggested that prenatal infection in combination with existing genetic risk factors can predispose individuals to schizophrenia (Choudhury and Lennox, 2021). MIA was first linked to schizophrenia risk through epidemiological studies. For instance, schizophrenia was

shown to have increased from a 1% prevalence to a 20% prevalence following a rubella pandemic in 1964 (Brown, 2011). While not all subsequent studies have resulted in such clear association, studies have also implicated immune system genes in schizophrenia. Increased pro-inflammatory cytokine levels of IL-1B and IL-8 are common in individuals with schizophrenia while anti-inflammatory cytokine levels such as those of IL-4 are reduced, which again suggests immune system involvement (Momtazmanesh et al., 2019). Early life adversities such as social isolation in conjunction with MIA is thought to further increase risk of schizophrenia (Khandaker et al., 2012). Other prenatal factors such as hypoxia at birth were both found to increase risk (Byrne et al., 2007; Davies et al., 2003). There also has been associations between season of birth and schizophrenia development with a stronger association in individuals born in the winter months (Kahn et al., 2015).

Some environmental factors that occur later in life are thought to be implicated in schizophrenia development. Early life adversity such as neglect or abuse during childhood is thought to increase risk with the severity of the event proportional to the increased risk (Varese et al., 2012). As schizophrenia is highly heritable and severe mental illness of a parent may make childhood adversity more likely to occur, it is difficult to discover the precise contribution of these factors. Studies have not found a correlation between polygenic risk score (PRS) for schizophrenia and early life adversity but others have found that in individuals with a high PRS for schizophrenia there is a link between psychosis and early life stress (Woolway et al., 2022).

Drug use, in particular cannabis use, has been associated with psychosis which is often observed in schizophrenia (Robinson and Bergen, 2021). Heavy use of cannabis is also associated with a subsequent increased risk of developing schizophrenia in individuals that abuse drugs at a younger age (Evins et al., 2013). In addition, a PRS for schizophrenia was associated with decreased grey matter in males who use cannabis (French et al., 2015). Finally, factors such as living in urban settings and migration are also thought to increase risk. Increased stress, socio-economic factors and social isolation are thought to be involved with the migration association while similar factors are thought to be at play in individuals who live in urban settings (Robinson and Bergen, 2021).

1.1.4 Neuroimaging in Schizophrenia

Structural neuroimaging has revealed altered brain volumes in individuals with schizophrenia (Kahn et al., 2015). Reductions in grey matter in cortical areas is common in schizophrenia as is a decrease in cortical thickness (Kuo and Pogue-Geile, 2019). Cortical volume is also seen to decrease with disease progression which is in line with the decreasing cognitive function observed (Cropley et al., 2017). Reduced volumes in the amygdala,

hippocampus and thalamus have also been reported (Howes et al., 2023). A decrease in structural and functional connectivity is also observed (Erdeniz et al., 2017). In addition, abnormal information processing is linked to schizophrenia with molecular neuroimaging studies showing altered activation in the cortex (Erdeniz et al., 2017). Molecular neuroimaging also shows altered dopamine uptake which is in line with the neuropathology of schizophrenia (Cumming et al., 2021).

1.1.5 Pathophysiology of Schizophrenia

Schizophrenia is a complex disorder with many environmental and genetic risk factors, many of which may act in tandem to cause neurological changes and subsequent symptoms. These factors are thought to influence neurodevelopment and alter brain function from an early stage. The dysregulation of a number of neurotransmitter systems including the dopamine, GABAergic and glutamatergic pathways is thought to be central to the pathophysiology of the disorder (McCutcheon et al., 2020). The original neurotransmitter hypothesis in schizophrenia involved the dopaminergic neurotransmitter system as dopamine signalling alterations are linked to the development of psychosis, which is a hallmark of schizophrenia (Kesby et al., 2018). The efficacy of traditional anti-psychotic drugs is due to their ability to bind D2 dopamine receptors in cortical tissue reducing psychotic symptoms such as hallucinations (Kesby et al., 2018). However while psychosis is a major part of schizophrenia it is not exclusive to the disorder and other symptoms of schizophrenia are not improved through the use of anti-psychotic agents suggesting that there are other pathways that produce schizophrenia related symptoms.

Other neurotransmitter systems such as the glutaminergic system have been investigated to try to identify alterations that can account for other symptoms in schizophrenia. This theory postulates that the reduced function of NMDA (N-methyl-D-aspartate) receptors in the prefrontal cortex can lead to positive symptoms through the increased activity of glutamate signalling (McCutcheon et al., 2020). It is thought that this may in turn over stimulate the dopamine pathway leading to hallucinations and delusions. Yet it is thought that over time glutamate levels are reduced in schizophrenia and postmortem studies have also identified a reduction in glutamatergic neurons (Wang et al., 2018; Kruse and Bustillo, 2022). In addition, antagonists of the NMDA receptor can lead to cognitive deficits similar to those seen in schizophrenia (Newcomer et al., 1999). The GABAergic system has also been implicated due to altered expression of the enzyme required for GABA synthesis and reduced numbers of GABAergic neurons observed in postmortem brain tissues (Jahangir et al., 2021). Again dopamine can affect the activity of GABA neurons which suggests all three systems may be implicated with complex interactions leading to observed symptoms.

1.1.6 Heritability of Schizophrenia

Schizophrenia, which has long been noted to be at increased incidence in some families, has been studied as a genetic disorder since the early 1900s (Bleuler and Jung, 1908). Initial theories posited that it was a monogenic disorder, however this theory was quickly replaced with the idea of a more complex inheritance pattern as observational studies did not support a monogenic basis (Henriksen et al., 2017). The first studies that investigated the impact of genetics on schizophrenia largely focused on twin studies to ascertain how heritable the disorder was. This involved investigating concordance rates for schizophrenia using twin studies. Concordance rates are defined incidence of both twins developing the same trait. The risk of developing schizophrenia is, as expected, highest for a monozygotic (MZ) twin of an affected individual at $\sim 50\%$, and risk decreases for first and second degree relatives of affected individuals (Chou et al., 2017; Poletti et al., 2020). Twin studies allow the broad-sense heritability, i.e., the proportion of variance that is due to genetic factors to be calculated. Heritability estimates based on sets of MZ and dizygotic (DZ) twins are useful for complex traits as MZ twins are taken to be almost genetically identical which implies that any phenotypic difference between MZ twins is caused by environmental effects. In addition, the variance contributed by additive and non-additive genetic effects can be discerned by comparing heritability estimates between DZ and MZ twins (Henriksen et al., 2017). Twin studies estimate the heritability of schizophrenia to be quite high at approximately 80% with rates consistent between historical and more recent studies (Hilker et al., 2018; Sullivan et al., 2003), which overall suggests a very strong genetic component to the disorder.

1.2 Genetic Studies of Schizophrenia

As a result of the strong evidence from the aforementioned observational and twin studies that schizophrenia was impacted by genetics, subsequent studies focused on molecular genetics. These studies attempted to identify genetic changes and biological causes that occur in individuals that develop schizophrenia (Wei and Hemmings, 2000; Levinson et al., 2003). Genetic studies of schizophrenia can largely be placed into three categories which include historical linkage and candidate gene studies from the pre-genomics era and then studies of common genetic variants and studies of rare genetic variants that are modern genomics-based approaches and have largely been made possible by the development of single nucleotide polymorphism (SNP) arrays and next-generation sequencing (Trubetskiy et al., 2022; Singh et al., 2022). These three broad categories will be discussed in the followings sections.

1.2.1 Candidate Gene and Linkage Studies

Initial genetic studies of schizophrenia used linkage analysis in an attempt to identify regions of DNA involved in schizophrenia development. Linkage analysis is a genetic mapping approach which aims to identify genes that are inherited with a particular trait by using pedigrees of high-risk families to identify co-segregation of a chromosomal marker (Teare and Santibañez Koref, 2014). Linkage analysis has been relatively successful in the study of Mendelian diseases such as cystic fibrosis but are less successful with regard to complex traits and are not as useful when the target trait is caused by a large number of common variants (Teare and Santibañez Koref, 2014). Linkage studies of schizophrenia have implicated a large number of genomic regions including regions on chromosomes 1,2,3,4,5,8 and 10 in particular through meta-analysis using the original studies (Ng et al., 2009; Levinson et al., 2003). These loci were believed to harbour risk variants but fine-mapping did not identify the susceptibility variants. Another downfall of linkage studies was that they were ineffective at discovering loci of small effects such as common SNPs which are now thought to convey a significant proportion of genetic risk for schizophrenia.

While there was limited success from the various linkage studies that were carried out they did enable subsequent research particularly candidate gene studies. Candidate gene studies aim to identify an association between a particular SNP at a gene of interest and a specific phenotype without considering genome-wide information (Duncan et al., 2019). Such genes were often chosen because they were located in regions identified via linkage studies (positional candidate genes) or because these genes affected particular biological processes or pathways thought to be involved in the phenotype of interest (functional candidate genes) (Henriksen et al., 2017). In the case of schizophrenia this often involved genes in the dopaminergic pathway (Psychiatric GWAS Consortium Coordinating Committee et al., 2009). Examples of candidate genes for schizophrenia include *DISC1*, *NOTCH4* and *COMT* (Wei and Hemmings, 2000; Millar et al., 2000; Johnson et al., 2017). Again, these studies have also only been partially successful due to underpowered analysis, lack of replication and limited knowledge of the underlying genetic basis of schizophrenia at the time these studies were performed. A 2017 study found that polymorphisms within candidate genes were no more likely to be associated with schizophrenia than non-candidate genes (Johnson et al., 2017). This study initially focused on the top 25 studied candidate genes and expanded to include a further set of 86 genes in a more extensive analysis. This found that these genes were no more enriched than random sets of control genes using gene-set analysis. While this study only considered common variants, and cannot rule out the possibility of significant enrichment had rare variants been considered, it still implies that candidate genes have provided little information in terms of uncovering the biological basis of schizophrenia. Furthermore, a 2019 study systemati-

cally identified all candidate gene studies completed since 2011 and combined them with the legacy SzGene database, which had compiled research in this area prior to that date, for a meta-analysis totaling 550 SNPs across 168 genes (Liu et al., 2019). They found that only 29 of these SNPs reached genome-wide significance and the majority had also been identified by modern genomic studies. This again suggests that they have made a limited overall contribution to schizophrenia research.

1.2.2 Common Variation in Schizophrenia

Following the completion of the Human Genome project in the early 21st century, the HapMap project extended analysis to common variation in the human population (International Human Genome Sequencing Consortium, 2001; The International HapMap Consortium, 2003). These studies in combination with the development of next-generation sequencing technologies enabled the study of common variation in schizophrenia to become an increasing focus in neuropsychiatric genetics. Genomic studies have become a powerful tool in not only discovering common variants associated with schizophrenia but also in subsequently ascertaining the risk that they convey, overtaking traditional molecular genetics techniques. In addition, the biological pathways that these genes associated with these variants effect have also been explored. This section will explore the findings from these studies and note the key advances that have been made using a variety of methods.

1.2.2.1 Genome-Wide Association Studies

Genome-wide association studies (GWAS) aim to identify an association between genotype and phenotype by analysing the allele frequency of common variants in cases and controls for binary traits such as disease status or for quantitative traits such as height (Uffelmann et al., 2021). They analyse SNPs but can be used to identify associations with other genomic variants such as copy number variants (CNVs) or structural variants, which can be indirectly genotyped using SNP data. A region containing a SNP or SNPs which shows association with the trait of interest is known as a risk locus. GWAS requires a large sample size in order to generate enough power to detect these loci and initial GWAS were hindered by their very low sample size (Sullivan et al., 2018). Large sample sizes are required for sufficient power in GWAS studies due to the large number of SNPs being tested which may not survive multiple testing correction in small samples. In order to be considered to be associated with a trait of interest, SNPs must reach genome-wide significance. Genome-wide significance often refers to p-values which are $\leq 5 \times 10^{-8}$ in a GWAS. This value has been ascertained by the use of a Bonferroni correction for all independent common SNPs in the genome (Dudbridge and Gusnanto, 2008). One of the first GWAS conducted for schizophrenia for example had just 320 schizophrenia

cases and 325 controls (Mah et al., 2006). Other early studies including (O'Donovan et al., 2008) identified few significant loci while others that had much larger sample sizes through combining samples from different cohorts began to find more genome-wide significant alleles involved in cognition and brain development confirming the polygenicity of the disorder (International Schizophrenia Consortium et al., 2009; Stefansson et al., 2009).

Today, the majority of findings from schizophrenia GWAS come from large-scale studies carried out by the Psychiatric Genomics Consortium (PGC) which combines many cohorts of cases and controls through a large international collaboration. The first of these studies PGC1 (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011), identified seven loci, only two of which had been previously implicated using 9,394 cases and 12,462 controls. The strongest finding suggested a previously unknown mechanism of dysregulation in schizophrenia involving a microRNA (*MIR137*) with a role in neuronal development. The second large scale analysis (PGC2) involved a sample size of 36,989 cases and 113,075 controls and resulted in a large increase in the number of genome-wide significant loci, which rose to 108 loci (128 SNPs) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). This confirmed the benefit and need for extremely large sample sizes in detecting associated loci and implicated biological processes such as synaptic transmission and calcium channels and the activity of glutamatergic neurons. It also reported the first significant association with a gene involved in the dopamine system. A meta-analysis involving the PGC data in 2018 further increased the number of loci to 145 involving 179 SNPs using 40,675 cases and 64,643 controls and replicated their findings using further samples (5,662 cases and 154,224 controls) (Pardiñas et al., 2018). This analysis replicated findings from the 2014 GWAS identifying 93 of the 108 loci previously identified with 50 novel loci reported.

The largest schizophrenia GWAS to date, PGC3 contains over twice the number of cases and controls than the previous PGC2 analysis (76,755 cases and 243,649 controls) demonstrating the effort to sufficiently power these studies to identify as many loci as possible (Trubetskoy et al., 2022). This study identified 287 distinct loci, replicating the association with all but one of the PGC2 loci and is illustrated in the Manhattan plot in Figure 1.1. In addition, this represents the most diverse schizophrenia GWAS that has been carried out. While it is still comprised of individuals of European descent it also includes a sizeable East Asian cohort along with smaller African and Latino cohorts but it is apparent that population specific loci will remain undiscovered until these samples are increased. One hundred and twenty of the genes associated at the significant loci were prioritised with high confidence using fine-mapping techniques and Mendelian randomisation. Biologically, most associated genes were involved in neuronal processes including differentiation and synaptic transmission. However, despite its successes, GWAS still

only explains 24% of the heritability of the disorder suggesting that a large proportion of heritability in schizophrenia remains missing (Trubetskoy et al., 2022; Dennison et al., 2020). It is thought that this missing heritability may be attributed to rare variation or epigenetic modifications rather than common variation. Additionally, despite having approximately double the sample size of the previous PGC2 GWAS, PGC3 only explains an additional 1.5% of the variance in disease heritability attributable to measured SNPs, suggesting that furthering our understanding the biology of schizophrenia will be limited by considering common variation alone.

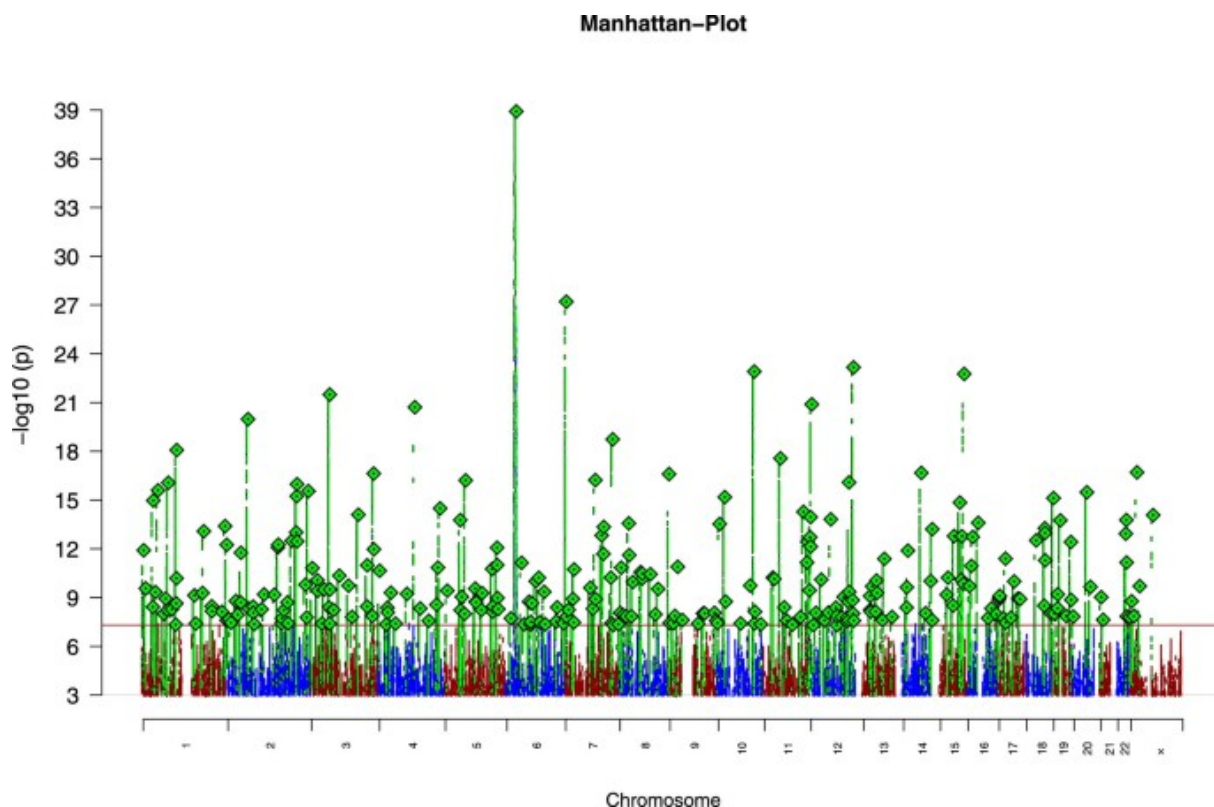


Figure 1.1: Manhattan plot representing the latest schizophrenia GWAS. This study identified 287 distinct loci associated with schizophrenia (Trubetskoy et al., 2022).

1.2.2.2 Application of GWAS Results

1.2.2.2.1 Functional Annotation and Analysis of GWAS Findings

Once associated SNPs have been identified through GWAS it is important to identify which variants are causal and what genes they are located in. As many SNPs pass significance thresholds they are often narrowed down by taking the SNP with the lowest p-value in a linkage disequilibrium block as the “lead” SNP. As not all “lead” SNPs are causal, fine-mapping techniques have become more widely used to identify causal SNPs.

Fine-mapping involves exploring the regions containing associated SNPs and identifying genes known to map to that region using Bayesian mapping methods such as FINEMAP (Schaid et al., 2018), (Benner et al., 2016). Once a causal SNP has been found there are a number of ways to map a SNP to a gene. If SNPs are located within a gene, positional mapping can be used by using coordinates within a reference genome. However, often a SNP is not in a translated area of a gene with only 2-3% of SNPs located in coding regions (Visscher et al., 2017). The majority are instead located in a regulatory element, which could be located hundreds of megabases away from the gene that it affects. In this case, it can be more difficult to define a casual gene. While it may be the gene located closest to the associated SNP is the risk gene, a recent study suggests that this may not always be the case (Morris et al., 2023). Molecular quantitative trait loci (molQTL) analysis can be used to identify causal genes by linking a SNP to a molecular phenotype such as gene expression, DNA methylation or chromatin accessibility using the associated data such as RNA-seq, Hi-C or ATAC-seq which will be discussed later in this chapter (Uffelmann et al., 2021).

Outside of mapping associated SNPs to genes, many studies aim to functionally annotate these in order to identify the biological pathways and processes that they affect. On an individual level genes which are involved in voltage-gated calcium channels, synaptic plasticity and glutamatergic transmission were implicated in the first large scale GWAS in 2014 (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). However, once mapped to genes, GWAS results are commonly organised into gene-sets for further analysis. Enrichment or over-representation analysis aims to identify pathways, biological processes, cell types or gene ontology (GO) terms where the genes present in the given gene-set are enriched. This enables us to link GWAS results which can include a large list of genes to molecular and biological functions as genes associated with certain phenotypes are expected to converge on similar pathways. Enrichment analysis can also be self-contained or competitive with the former identifying any association and the latter identifying enrichment in the tested gene-set compared to background gene-sets (de Leeuw et al., 2015).

Initial gene-set analysis from the 2014 PGC GWAS did not identify any enriched pathways with the associated genes after multiple testing correction but once variants were mapped to active enhancer sequences, associations were enriched at enhancers active in the brain but not in other tissues (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Variants from this GWAS were also enriched in genes that have been known to contain rare mutations associated with schizophrenia and were also enriched for genes expressed in neuronal cell types. The 2018 meta-analysis which identified 50 novel loci showed an enrichment for schizophrenia associations in loss-of-function (LOF) intol-

erant genes compared to other genes using the gene-set analysis tool MAGMA (Pardiñas et al., 2018). This is in contrast to other brain related and neuropsychiatric disorders which show at best a weak enrichment for the LOF gene-set. Of particular interest from this study was the fact that these deleterious alleles with small effects persisted in the population at a high frequency. This study suggests that background selection as a result of purifying selection in areas of low recombination reduces genetic diversity allowing alleles with small deleterious effects to rise in frequency. Further testing for enrichment of common variants in 6,677 publicly available gene-sets resulted in six gene-sets showing enrichment following multiple testing correction. This included the LOF intolerant gene-set followed by targets of Fragile X mental retardation protein (FMRP) and genes related to abnormal behaviour in mice, calcium ion transport, synaptic transmission and membrane depolarisation.

More recently, gene-set analysis using associated variants arising from the 2022 schizophrenia GWAS identified an enrichment in neuronal cell types particularly excitatory glutamatergic neurons from the cerebral cortex and hippocampus and inhibitory cortical interneurons using human and mouse single-cell data (Trubetskoy et al., 2022). These enrichments together with the lack of enrichment for glial cell types indicate that schizophrenia is linked to neuronal dysfunction but this dysfunction cannot be attributed to a single brain region. Gene-set analysis of prioritised genes from this study shows concentration of genes in pre- and post-synaptic locations and annotation points to disruption in functions such as calcium and chloride channels, synaptic transmission and differentiation. Overall, gene-set analysis arising from the major association studies converge on theory that points to a post-synaptic pathology occurring for the most part in neuronal cell types in a number of brain regions resulting from dysregulation in synaptic processes and ion channels. Gene-set analysis following GWAS is therefore invaluable in enabling us to discover biological implications of deleterious variation in risk genes. This informs functional genomic studies which aim to discover the impact of this variation and will be discussed in further sections of this chapter.

1.2.2.2.2 Trait Risk Prediction

Aside from using functional annotation to discern biological implications of common schizophrenia risk variants, the other main application of GWAS results is PRS. As many disorders are polygenic and the risk of their development can be dependent on a large number of genes, it is useful to calculate their cumulative risk which of course varies per individual. PRS can be calculated by a number of methods but requirements usually include raw genotype data and GWAS summary statistics of the desired phenotype. Most commonly, PRS are calculated using methods that use LD clumping to remove non-independent SNPs and a P value threshold is then applied to remove SNPs with weaker

associations (Choi et al., 2020). The remaining SNPs are then used to calculate the PRS by summing the minor alleles for each SNP and multiplying by the effect size in the GWAS, as is performed in the most frequently used tool PRSice-2 (Choi and O'Reilly, 2019). Other methods of calculating PRS include LDpred2 and Lassosum which use pseudo-validation approaches to give approximate validation of a score by adjusting shrinkage parameters using GWAS effect size (Ni et al., 2021; Pain et al., 2021). These methods have been shown to perform better than traditional clumping methods.

PRS was first developed and used for a 2009 schizophrenia GWAS study which showed that PRS was a predictor for schizophrenia status in a European cohort explaining around 3% of the variance. PRS scoring is now widely used to predict risk in a wide range of disorders and its potential clinical utility is hotly debated. In schizophrenia, PRS is still only weakly predictive of status with 7% of phenotypic variance explained when using the latest GWAS to predict schizophrenia status (Trubetskoy et al., 2022; Dennison et al., 2020). Despite this, it has been suggested that it may be useful for screening and prioritising samples for functional studies following GWAS (Dennison et al., 2020).

1.2.3 Rare variation in Schizophrenia

Given that common variation only explains 24% of overall disease liability in schizophrenia it is thought that rare variants may confer a substantial level of risk to a given individual. Rare variants are usually defined as variants that have a minor allele frequency (MAF) of less than 0.1% (Singh et al., 2022). Initial information regarding rare variation in schizophrenia has come from copy number variant (CNV) studies while more recent studies have focused on sequencing the exome which is the coding portion of the genome. Both of these approaches will be discussed in the following sections.

1.2.3.1 Copy Number Variation

CNVs were first implicated in schizophrenia through studies of 22q11 deletion syndrome as carriers of this deletion have a much higher risk of schizophrenia development with respect to the wider population (Singh et al., 2022). Other initial evidence of CNV involvement came from the enrichment of both *de novo* and rare deletions and duplications in schizophrenia cases relative to controls (Marshall et al., 2017). More recently, a 2016 study of 21,094 cases and 20,227 controls identified a global enrichment of CNVs in cases versus controls and this remained the case when the nine previously identified CNV loci were removed (Marshall et al., 2017). Additionally, it is thought that the eight loci that pass genome-wide significance explain 1.4% of disease liability and gene-set analysis of this dataset implicated genes involved in synaptic or neuronal function. One interesting finding from this study however is that just over half of the previously reported loci were

not significantly associated with schizophrenia which is likely either a result of underpowered studies resulting in chance associations or the inability to capture this association due to the rarity of these variants. This underlines the need to increase sample studies for CNV studies in order to robustly characterise their contribution to schizophrenia. While the previously discussed study was concerned with germline CNVs, a 2023 study also detected an enrichment of somatic CNVs in schizophrenia cases including mosaic deletions in *NRXN1*, a gene previously implicated in schizophrenia and neurodevelopmental disorders (Maury et al., 2023). Overall recent research continues to implicate CNVs as an important element contributing to schizophrenia risk.

1.2.3.2 Exome Sequencing Studies

Exome sequencing has become another approach that can be used to identify rare variants. It is largely employed to identify rare point mutations or small insertions or deletions. Individuals with schizophrenia have reduced fecundity and as such combined with natural selection it is thought that large-effect variants will be at low frequency in the population yet they may persist in the population due to newly arising *de novo* variants (Rees et al., 2011). Initial exome studies focusing on *de novo* variants identified an enrichment of these mutations in synapse related genes using a cohort of 623 trios (Fromer et al., 2014). Further studies with greatly increased sample sizes showed that *de novo* variants were highly enriched in genes that were under evolutionary constraint or highly expressed in the brain (Howrigan et al., 2020). Yet, these studies have failed to find genes that pass exome-wide significance due to low sample sizes (Howrigan et al., 2020).

While studies of *de novo* variants in the exome have been limited by small sample sizes, case-control studies have managed to study much larger cohorts. One of the first larger scale studies in 2014 utilised a cohort of 5,079 individuals (Purcell et al., 2014). The researchers prioritised a set of genes that were enriched for common variants associated with schizophrenia. They identified that this gene-set was enriched for rare mutations in case samples compared to controls. Association of single genes however was difficult due to the nature of rare mutations and sample size. However, gene-sets related to calcium channels were enriched for rare mutations which is consistent with results from studies of common variants discussed above as well as studies of cognitive impairment in mice and humans. In addition, genes previously found to be enriched for *de novo* mutations were similarly enriched for rare variants. A further study published in 2016 looked at both *de novo* mutations in addition to a case-control cohort revealing that genes with *de novo* mutations were also enriched for LOF variants in cases from the case-control cohort (Singh et al., 2016). Meta-analysis of these LOF variants in a case-control cohort identified an association between the *SETD1A* gene and schizophrenia. *SETD1A* is part of a

family of methyltransferases which were previously associated with Mendelian disorders resulting in intellectual disability and suggests epigenetic dysregulation is implicated in schizophrenia and until recently remained the only gene with ultra rare variation robustly associated with schizophrenia.

The most recent large-scale exome sequencing study which was published in 2022 aimed to combat the flaws of previous studies by initiating a collaborative effort to analyse and harmonise a large number of exome sequencing studies. The Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) consortium compiled a cohort of 24,248 schizophrenia cases and 97,322 controls which is one of the largest exome sequencing cohorts to date (Singh et al., 2022). In contrast to previous studies and as predicted, the increase in sample size led to the discovery of ten genes with ultra rare variation that pass exome-wide significance with a further 22 genes significant at a false discovery rate (FDR) rate $<5\%$. The implicated genes encompassed a wide variety of functions but did follow a similar theme to the findings from previous studies of rare and common variation including ion transport, neuronal migration and transcriptional regulation. Overall, while rare variant analysis has been ongoing for many years it is only recently that sample sizes have become large enough to identify genes in which these variants recur and should increase in number as samples sizes grow. Interestingly, the functions of these genes converge with the genes enriched for common variants suggesting that ion transport channels and synaptic function should be a focus for functional studies and treatment targets (Figure 1.2) (Nakamura and Takata, 2023).

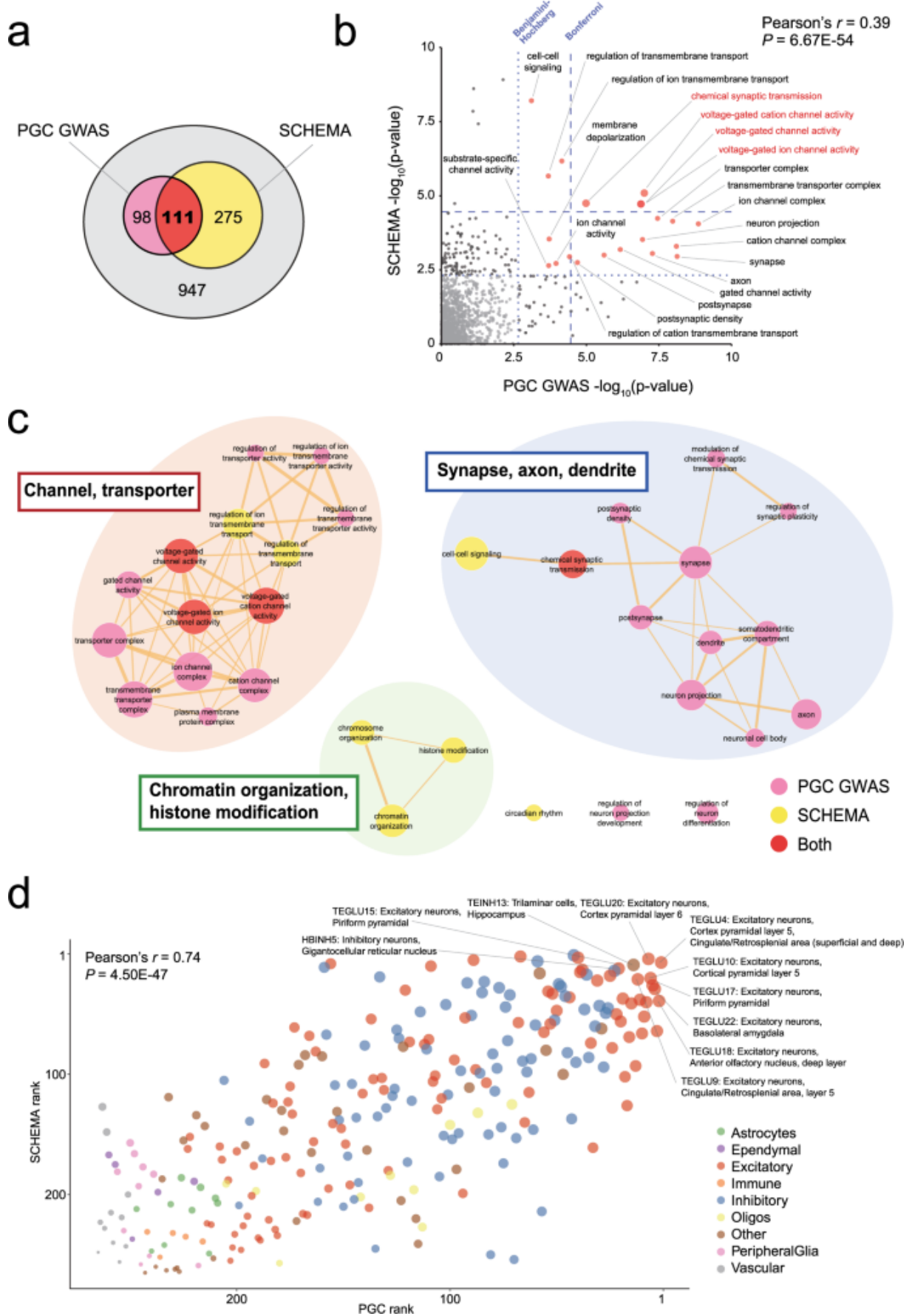


Figure 1.2: Graphical illustration depicting the convergence of schizophrenia risk genes on common functions.

a: Overlap of enriched GO terms for common variants (PGC GWAS) and rare variants (SCHEMA), **b:** Enrichment of GO terms from these studies showing considerable overlap, **c:** Networks displaying the connections between these GO terms showing inter-linked functions, **d:** Illustration of the cell type enrichment results from these studies showing significant overlap for excitatory neurons. Adapted from (Nakamura and Takata, 2023)

1.3 Functional Genomics Studies of Schizophrenia

While genetic and genomic studies have been invaluable in discovering aspects of the disorder such as heritability, risk genes and the degree to which these genes affect disease liability, much has yet to be elucidated with regard to disease biology. These studies have suggested possible pathways and mechanisms that may be affected by the aforementioned risk genes, yet the full extent of the biological dysregulation is poorly understood and uncharacterised. A wide variety of functional studies have attempted to uncover the underlying biological changes in gene expression and regulation in schizophrenia by using mouse models and more recently human postmortem brain tissue and brain organoids and induced pluripotent stem cells (iPSCs). This section will first recap the findings from historical mouse model studies before describing transcriptomic studies of postmortem brain tissue and will conclude with a discussion of studies that focus on dysregulation of the epigenome in schizophrenia and integrative studies that combine multiple types of omics data.

1.3.1 Mouse Models of Schizophrenia

Before the development of next-generation sequencing technologies and iPSCs, mouse models were used to investigate the neurobiology of schizophrenia. These aimed to recreate the phenomenology of schizophrenia usually through the knockout of a potential schizophrenia risk gene or recreation of an associated genetic or neurobiological defect (Lipska and Weinberger, 2000). As such, initial models focused on disrupting the dopamine system due to its association with schizophrenia yet they failed to represent the true clinical phenotype of the disorder. Further models aimed to disrupt the function of, for example, *DISC1* which was believed to be a highly penetrant risk gene for schizophrenia. This is now disputed as more recent genetic studies do not show significant support for this despite the initial cytogenetic and pedigree evidence (Sullivan, 2013). Nevertheless, these models represent an initial foray into possible functional changes that may occur and they discovered that LOF models of *DISC1* recapitulated similar brain morphology to that observed in schizophrenia with reduced cortical thickness in addition to cognitive changes (Clapcote et al., 2007; Jaaro-Peled, 2009). Other models have included generation of the 22q11 deletion syndrome microdeletions which also recreates similar brain morphology (Ellegood et al., 2014). Other potential risk genes such as *NRG-1* and *RELN* have also been perturbed in mouse models resulting in changes in dendritic spine density and increased neuronal packaging respectively which supports findings from GWAS that have also implicated similar processes (Chen et al., 2008; Krueger et al., 2006). While mouse models remain of interest, particularly in the context of drug development, they have come under some scrutiny as replication studies have had variable results and they do not always represent the true functional changes observed in the human disorder.

der. As a result, there has been increased focus on the use of brain tissue isolated from schizophrenia cases to uncover changes that occur in the human disorder.

1.3.2 Postmortem Transcriptomic Studies in Schizophrenia

The use of postmortem tissue isolated from schizophrenia cases and controls has for the last 20 years been the primary focus of understanding the gene expression changes in the disorder. This section will first describe the initial microarray studies which profiled gene expression from bulk tissue dissections before discussing more recent profiling of the whole transcriptome using next-generation sequencing. It will conclude by highlighting the findings from the most recent studies which have endeavoured to identify changes in the transcriptome at the cell-type level.

1.3.2.1 Bulk Transcriptomics

The development of microarray technologies enabled affordable profiling of gene expression in a wide variety of disorders. It also provided a way of bridging the gap between genotype and phenotype as little was known about the way in which neurobiology was affected by the changes in genotype observed in schizophrenia cases (Mirnics and Pevsner, 2004; Mirnics et al., 2006). While microarray studies were limited as they could only profile a set of predefined genes, unlike RNA-seq studies which analyse the whole transcriptome, these initial studies still identified a number of pathways and processes still believed to be involved in schizophrenia such as presynaptic release, neuronal transcripts and metabolic pathways (Middleton et al., 2002; Mirnics et al., 2000). Recognising the limitations of only looking at subsets of genes and the output list of differentially expressed genes, another microarray study aimed to organise these genes into regional and functional networks in an attempt to provide a more holistic picture of the dysregulated expression (Roussos et al., 2012). In this analysis, they used weighted gene correlation network association (WGCNA) analysis to identify co-expression patterns of genes by grouping them into modules and identifying the most correlated genes within each module which are known as hubs (Langfelder and Horvath, 2008). These hub genes are then used to ascertain whether there is an association between the identified modules and variables of interest such as cell-type, phenotypes or brain region. In this study, (Roussos et al., 2012), five modules were found to be associated for schizophrenia with the most significant module implicating oligodendrocytes as it was enriched for oligodendrocyte expressed genes and contained genes previously shown to be dysregulated in schizophrenia. This provided early evidence of the benefit of integrating the results of these studies to provide a clearer picture of the transcriptomic dysregulation.

Following the development of RNA sequencing technologies, the ability to profile the tran-

scriptome dramatically improved. One of the main benefits is that RNA-seq can profile the entire transcriptome such that the expression of all genes present in the sample can be assayed. This also enables a hypothesis-free approach which can help overcome inherent biases present in historical studies. Initially, these studies were often underpowered and contained overlapping samples. One of the first large scale studies meta-analysed 153 individuals with schizophrenia and 153 matched controls to try and overcome the limitations of smaller scale studies (Mistry et al., 2013). They identified only one gene *OPCML* as differentially expressed and it was also implicated by GWAS, though at the time a much smaller number of risk genes were known through these methods. A larger overlap was observed in relation to other gene expression studies of schizophrenia. In addition genes involved in pathways associated with schizophrenia were also observed such as *CACNB3*, a gene involved in calcium signalling and others involved in postsynaptic signalling and the dopamine pathway. Yet, these early studies in combination with microarray studies often had inconsistent results particularly at the individual gene level. As a result, similar to GWAS, there has been a move towards large-scale consortium studies in an attempt to overcome these issues.

One of the first of these studies was by the CommonMind Consortium (CMC), who profiled gene expression in dorsolateral prefrontal cortex (DLPFC) tissue from 258 schizophrenia cases and 279 controls (Fromer et al., 2016). They identified changes in genes related to the GABA and NK3 receptors suggesting interneuron dysfunction and possible cognitive changes respectively, yet no significant pathways were observed. WGCNA analysis showed that four of 35 modules were associated with schizophrenia differentially expressed genes (DEGs). One of those four modules was associated with both the DEGs and genetic variants associated with schizophrenia. This module also displayed reduced “density” in cases which suggests that the regulation of these genes is disrupted in schizophrenia and functional annotation implicated processes including axon guidance and postsynaptic density. Interestingly, none of the schizophrenia risk genes identified by the GWAS at the time were found to be differentially expressed yet this was thought to be due to limited sample size. This study integrated genetic data with expression data to facilitate the detection of expression quantitative trait loci (eQTL), that are variants which explain variance in gene expression. Of the 108 schizophrenia risk loci reported at the time, 73 contained a significant eQTL yet this was expected by chance due to the sheer number of eQTLs tested. However, colocalization analysis of the eQTLs and schizophrenia risk loci identified 19 loci that contained a schizophrenia associated eQTL with eight implicating a single gene. Five of these (those which encoded a known protein) were replicated in another cohort. Of these five, two were involved in neurodevelopment while the others were involved in synaptic components and all are included in the PGC3 list of 120 prioritised genes for schizophrenia. This highlights the importance of integrating functional

and genetic data in order to understand the biology of a disorder (Trubetskoy et al., 2022).

Additional large-scale studies of the brain transcriptome in relation to neuropsychiatric disease has since been published by the PsychENCODE consortium (PsychENCODE Consortium et al., 2015). This consortium combined data from ENCODE (Encyclopedia of DNA Elements), CMC, GTex (Genotype-Tissue Expression) and Roadmap (Roadmap Epigenomics Mapping Consortium) in addition to samples sequenced for PsychENCODE (Wang et al., 2018). The ENCODE project has generated both gene expression and epigenomic data in order to define biochemical functions of the genome (The ENCODE Project Consortium, 2012) while GTex has compiled gene expression and genetic data in order to study the relationship between the transcriptome and genetic variation (Lonsdale et al., 2013). Roadmap has created epigenomic maps in order to better understand gene regulation using data that shows changes in methylation, histone modifications and other chromatin features in the genome (Bernstein et al., 2010). In terms of bulk transcriptomic data, there were 1,089 controls and 558 schizophrenia cases and their analysis provided a comprehensive look at not only gene level expression but transcript expression and differential splicing (Gandal et al., 2018). While differential expression analysis at gene level was concordant with previous studies, the extensive transcript analysis was a relatively new aspect to the analysis. It found that while DEGs were shared across neuropsychiatric disorders, differentially expressed transcripts (DETs) were more disease specific although similar pathways such as inflammatory response and transporter signalling were implicated in both analyses. By using expression data from individual cell types (which will be discussed in more detail in subsequent sections) they were also able to implicate cell-types relevant to disease. In addition, DETs whose gene was not differentially expressed appeared to be enriched in excitatory neuronal cell-types which have been previously associated with the disorder (Skene et al., 2018). It is true that much of the focus on transcriptome dysregulation is on protein-coding genes yet dysregulation was also observed in noncoding portions of the transcriptome and a class of noncoding RNAs (long noncoding RNAs) associated with neuropsychiatric disorders were found to be under greater selective constraint than non-associated noncoding RNAs (Gandal et al., 2018).

Differential splicing was also explored in order to complement the differential transcript expression analysis by characterising intron usage. Exon skipping was the most widely observed change followed by alternative 5' exon inclusion (Gandal et al., 2018). Enrichment of splicing changes was also observed in *RBFOX1* and *FMRP* which are both RNA binding proteins with targets implicated in schizophrenia. Pathway analysis of genes implicated in this analysis again pinpointed pathways involved in cytoskeleton and neuron development which suggests that overall large-scale studies of bulk gene expression point

towards a general picture of disruption in cell integrity, neuronal development and other synaptic processes. Yet, analysis of different aspects of the transcriptome such as splicing or transcript expression can elucidate further changes that are not observed by general analyses. This is also an inherent limitation of bulk analysis in general given that the dysregulation observed is an overall picture of the tissue chosen. Indeed, important changes at the cell-type level may be undetected due to differences in cell-type proportion with rarer cell-types not present in enough quantities to be picked up in bulk analyses. Furthermore, expression can be variable across brain regions, and even within regions, implying that both spatial and cell-specific analyses are required if we want to truly characterise the neurobiology of schizophrenia.

1.3.2.2 Cell-type Specific Analysis

As mentioned above, there has been an increasing focus on carrying out functional analysis in a cell-type specific manner. By focusing only on bulk tissue, the complexities and diversity of the transcriptome is not evident. As a result, transcriptomic analysis of specific cell-types is becoming increasingly common and can be carried out in two different ways depending on the technology used. In order to differentiate between them I will refer to the more modern single-cell (SC) studies which profile all cells in the sample as SC while older studies which profile more limited numbers and types of cell as cell-type specific (CTS). It must also be noted that the majority of cell-type specific studies of brain tissue only profile the nucleus and not the cytosol due to the difficulty of isolating whole neurons due to their structure (Lake et al., 2017; Ding et al., 2020). Yet, it has been demonstrated that this has minimal effect on the overall results and studies comparing single-cell and single-nucleus transcriptomes have shown high concordance (Slyper et al., 2020; Ding et al., 2020). In order to minimise confusion and in line with published literature, I will refer to these studies as SC despite only the nuclei been isolated.

Reverting back to the cell-type specific studies that have been carried out, both CTS and SC studies first require the isolation of single cells, which can be carried out using a wide variety of methods such as fluorescence activated cell sorting (FACS), which tags cells with a fluorescent marker allowing sorting of distinct populations, or micromanipulation, which uses microscope guided pipettes (Hwang et al., 2018). A variety of the available methods are displayed in Figure 1.3. FACS is now thought of as the preferred method as it produces highly purified populations and is useful for isolating cells that have low population numbers. Once the cells are isolated however, the CTS and SC studies differ, these differences are detailed in Table 1.2. One major advantage of SC technologies however is that they profile all available cell populations and as such are extremely high throughput which is advantageous as it provides a more representative view of the expression per

cell-type. Yet in studies of the transcriptome, CTS retains an advantage as it uses bulk sequencing technologies which can profile specific transcripts. SC technologies only capture part of a transcript, namely the 3' end, such that isoforms can not be profiled, this advantage of CTS studies is explored in Chapter 4 of this thesis.

Table 1.2: Features of bulk, cell-specific and single-cell sequencing studies.

Study Feature	Bulk Studies	Cell-type Specific Studies	Single-cell Studies
Sample size	Up to thousands of individuals	Usually less than 500 individuals	Usually less than 100 individuals
Number of cells	No individual cell types	Study dependent, often 10-100	Study dependent, up to millions of cells
Which cells are profiled?	None	Purified cells only	All cells in the sample are profiled
Is transcript information available?	Yes	Yes	Not currently
Main advantages	- Large sample size - Transcript information	- Cell-type information - Transcript information	- Cell-type information - Large numbers of cells
Main disadvantages	- No cell-type information	- Smaller sample sizes at both cell and individual level	- No transcript information

CTS studies have profiled a wide variety of cell types. Initial CTS studies profiled pyramidal cells as working memory, which is affected in schizophrenia, relies on the firing of pyramidal cells which end on dendritic spines which are reduced in schizophrenia (Arnstén et al., 2012). In terms of affected pathways the mitochondrial (MT) and ubiquitin protease (UPS) systems were the major functions implicated in this study. Both layer 3 and layer 5 cells were profiled with MT dysregulation most observed in layer 3 and UPS in layer 5 and not observed in bulk grey matter highlighting the need to research cell-specific effects. UPS activity is enriched at pre- and post-synaptic terminals which again suggests synaptic dysfunction and is concordant with studies previously discussed in this chapter (Arion et al., 2015). The majority of the DETs observed were under-expressed in cases and did not overlap with cases of schizoaffective disorder which were also profiled, suggesting differing neurobiology. This was confirmed by a followup study by the same group which also showed differing DEGs in schizophrenia, major depressive disorder and bipolar disorder (Arion et al., 2017). A further study which explored expression alteration in parvalbumin neurons again implicated the mitochondria, though perhaps more interesting was the finding that over 80% of DEGs were cell-type specific when compared

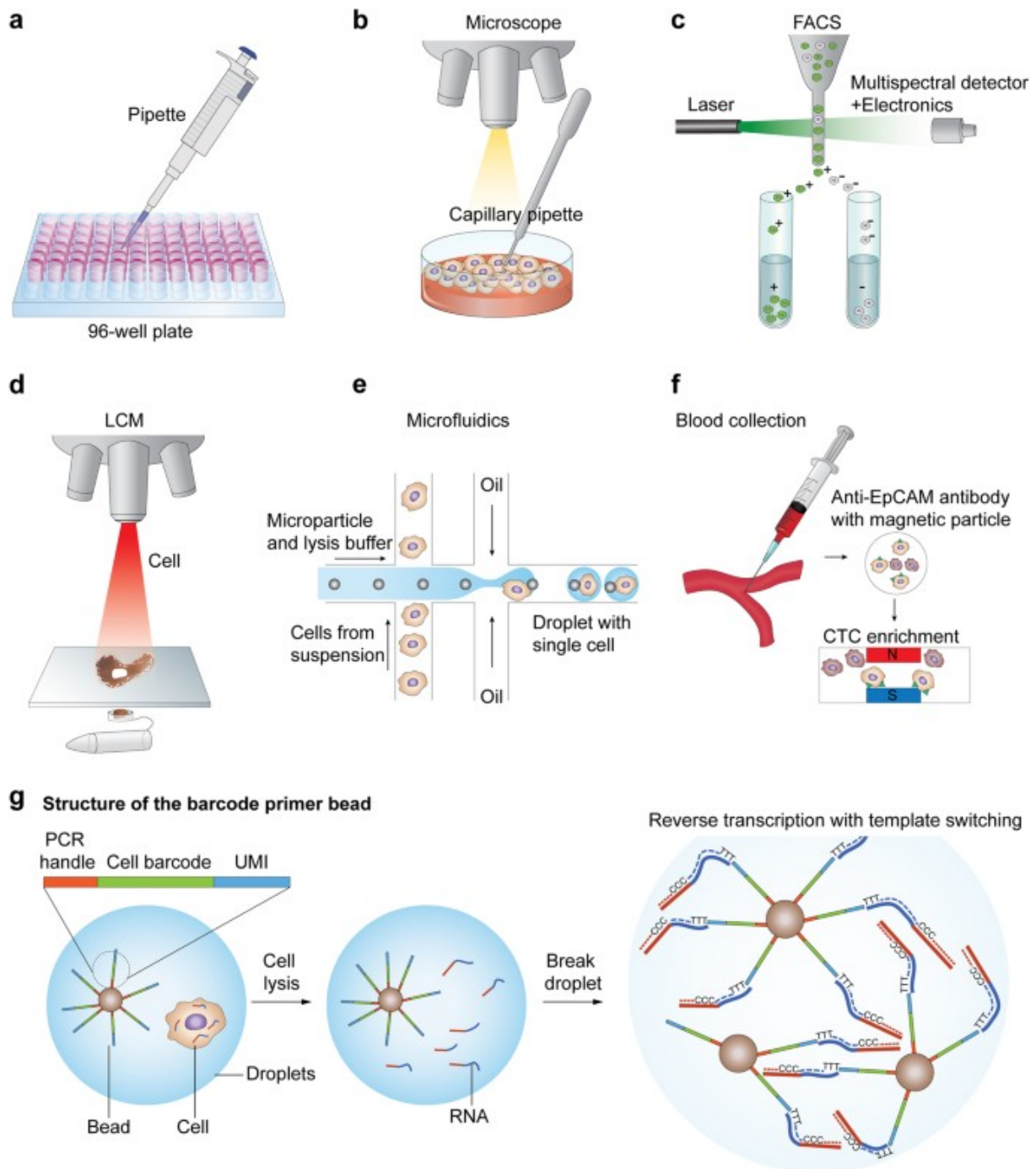


Figure 1.3: Graphical illustration single-cell isolation techniques.

Graphical representation of a number of methods used to capture individual cells which are then profiled using different omic technologies. It was adapted from (Hwang et al., 2018).

to previous results of the pyramidal cell study. This again showed the changes that would remain undetected in bulk studies (Enwright et al., 2018).

While the number of both CTS and SC studies that contain schizophrenia cases are limited due to difficulty in collecting large numbers of postmortem samples, studies which only profile cell-type specific expression in the brain of control samples can also be used to gain insights into the contribution of different cell-types to schizophrenia. The first SC transcriptome study of the human brain, published in 2015, identified 10 individual cell types (Darmanis et al., 2015). Subsequent studies have been able to identify more distinct cell populations given increased sample sizes (Lake et al., 2017). Not only that, they have been able to integrate this expression data with published GWAS in order to identify cell types that may be associated with the disorder. A 2018 study compiled several scRNAseq datasets from both mouse and human to conduct cell-type enrichment using common variants associated with schizophrenia (Skene et al., 2018). They found four cell types which were robustly associated with schizophrenia including hippocampal CA1 pyramidal cells, striatal medium spiny neurons, neocortical somatosensory pyramidal cells and cortical interneurons. Overall, the landscape of the cell-type specific transcriptome is one which continues to be discovered particularly as these studies become more feasible through the collection of further samples and the development of single-cell technologies.

1.3.3 Postmortem Studies of the Epigenome in Schizophrenia

Functional postmortem studies not only explore the transcriptome but also the epigenome both at the bulk level and the cell-specific level in order identify changes in genomic regulation. There are a number of epigenetic changes that can be profiled such as methylation, histone modifications and chromatin accessibility in addition to three-dimensional analysis.

1.3.3.1 Chromatin Accessibility Analysis

Several studies have profiled chromatin accessibility in the brain using both ATAC-seq and THS-seq (Transposome hypersensitive site sequencing) which is a variant of ATAC-seq (Buenrostro et al., 2013), (Sos et al., 2016). ATAC-seq profiles open chromatin regions (OCRs) using an Tn5 transposase enzyme which can cut the genome at these regions. Further details on ATAC-seq will be discussed in Chapter 4. A 2018 study profiled open chromatin in neurons and non-neurons from healthy individuals and found that differentially accessible neuronal OCRs were enriched for common variants associated with schizophrenia (Fullard et al., 2018). Another study from the same year carried out bulk ATAC-seq profiling in schizophrenia cases and controls also found that OCRs were enriched for common variants associated with schizophrenia, particularly so OCRs in conserved regions of the genome (Bryois et al., 2018). Yet they found only three differentially accessible regions between cases and controls, which is in contrast to other disorders (Bendl et al., 2022). However, it is possible that this was hindered by technical covariates as substantially more

OCRs were associated with postmortem interval. Single-cell profiling of open chromatin in control brain cells attempted to assess enrichment of schizophrenia risk variants in particular cell types yet failed to find a significant enrichment in any cell type in this early study (Lake et al., 2017). A further CTS study of ATAC-seq in controls across four cell types, GABAergic neurons, Glutamatergic neurons, oligodendrocytes and a mixture of microglia and astrocytes did find enrichment of common variants for schizophrenia in the OCRs of glutamatergic neurons (Hauberg et al., 2020). The per SNP heritability of schizophrenia risk variants overlapping cell type specific OCRs also showed an almost exclusive neuronal signal, again suggesting schizophrenia is a disorder of impaired neuronal function and is consistent with similar analysis of the transcriptome (Skene et al., 2018). This study, (Hauberg et al., 2020) forms a large basis for Chapter 4 of this thesis and will be discussed there.

1.3.3.2 Analysis of the Methylome in Schizophrenia

Studies have also endeavoured to identify changes in DNA methylation in schizophrenia and it is of particular importance as this epigenetic function regulates gene expression. Changes in gene expression, particularly in development, are of keen interest as schizophrenia is now thought of as a neurodevelopmental disorder which makes the underlying cause of these changes a vital area for study (Jakovcevski and Akbarian, 2012). Profiling of the DLPFC methylome in schizophrenia and fetal control brains showed an overlap with regions harbouring schizophrenia risk variants (Jaffe et al., 2016). CpG sites which were differentially methylated between cases and controls were enriched for genes involved in neurodifferentiation and development. These CpG sites also correlated with changes in neurodevelopment and coincided with prenatal to postnatal changes. This suggests that the epigenome is involved in the disrupted neurodevelopment in schizophrenia (Jaffe et al., 2016). CTS profiling of the methylome has provided further detailed information about neuronal and oligodendrocyte methylation in schizophrenia. Of particular interest was the differential methylation observed in oligodendrocytes which was more extensive than previously thought by other studies. This highlights the issue of bulk studies which may have inadvertently been biased towards neuronal populations (Mendizabal et al., 2019). Divergent methylation patterns were also observed in the two cell-types with schizophrenia associated differentially methylated sites being hypomethylated in neurons with the opposite effect observed in oligodendrocytes (Mendizabal et al., 2019). Future SC studies of the methylome in schizophrenia will be crucial in further elucidating changes in specific cell-types.

1.3.3.3 Analysis of Histone Modifications and the 3D Genome in Schizophrenia

In addition to methylation and chromatin accessibility, other epigenomic alterations that have been explored are histone modifications and their relationship to the three-dimensional structure of the genome. Histone modifications such as methylation and acetylation in neuronal and non-neuronal cell-types were profiled with acetylation marks showing much higher coverage across the neuronal genome (Girdhar et al., 2018). Similar to chromatin accessibility and CpG methylation, histone modifications were enriched for schizophrenia GWAS loci again highlighting the need to study the epigenome in tandem with genomic studies. Enrichment was almost exclusively observed in neurons which suggested that GWAS studies are characterising genomic changes predominantly in neuronal cell types and most of the variation in modifications could be attributed to that cell-type.

Three-dimensional organisation of the genome such as topologically associated domains (TADs) or chromosomal conformations may also be altered in disease implying a need to relate this to schizophrenia genomics. Hi-C technology has enabled this research and involves the detection of chromatin conformations allowing the three-dimensional structure of the genome to be studied (Oluwadare et al., 2019). A 2018 study identified three-dimensional changes in iPSCs using Hi-C in neurons and astrocyte-like cells (Rajaraman et al., 2018). They then leveraged this map of three-dimensional changes in order to link these changes to schizophrenia risk loci with neuronal chromosomal interactions more likely to be in sequences associated with schizophrenia risk. This enabled genes associated with these interaction changes to be listed on a cell-type specific basis (Rajaraman et al., 2018).

A 2022 study then integrated histone modification data from both purified neurons and bulk DLPFC isolated from schizophrenia cases and controls with Hi-C data in order to characterise the changes in these modifications in relation chromosomal conformation (Girdhar et al., 2022). Differentially modified H3K4me3 methylation marks in neuronal populations did not survive multiple testing suggesting that this mark is not consistently affected in schizophrenia in neurons yet H3K27ac acetylation marks were dysregulated between cases and controls. Alterations were also observed within the *SYNTAXIN 1A* locus which is a risk locus for schizophrenia and regulates the docking of the synaptic vesicle. Following a similar theme of other epigenomic studies, acetylation marks were enriched for common schizophrenia risk variants supporting the 2018 study discussed previously but most strikingly these enrichments were driven by hyper-acetylation marks (Girdhar et al., 2018). Furthermore, by clustering these acetylation marks into cis regulatory domains, this study integrated these domains with Hi-C data to show that these domains were more

likely to occur with TADs than any other random sequence (Girdhar et al., 2022). Thus they correlated risk-associated regulatory domains with chromosomal domains (Girdhar et al., 2022). Indeed, it is this type of integrative analysis that is needed going forward as many individual studies converge on similar results that associated neuronal epigenomic dysfunction with schizophrenia. It is only by expanding the types of cells profiled and integrating many types of epigenomic, genomic and transcriptional data that we will further our knowledge of the underlying biology.

1.3.4 Issues with Postmortem Brain Studies

Before focusing on the aforementioned multi-omic studies, I will first address a number of common issues concerning postmortem studies. Firstly, an historical issue for these studies is that they have limited sample sizes. This resulted in a number of conflicting results prior to the establishment of functional genomic consortia such as the CMC and PsychENCODE (Fromer et al., 2016; PsychENCODE Consortium et al., 2015). Since the establishment of these consortia, large-scale studies have become increasingly common and now converge on similar pathways and associations making findings more robust and reliable. Secondly, as mentioned previously, single-cell studies of the human brain are particularly difficult owing to the structure of neuronal cell-types which necessitates the isolation of nuclei rather than whole cells. While some changes are missed in the cytosolic transcriptome, single-nucleus studies do characterise a substantial number of features in brain cell types (Skene et al., 2018). In addition, the age of the individuals in most postmortem studies skews towards an older age profile (Wang et al., 2018; Hoffman et al., 2019). This presents the issue that the onset of illness for these individuals was many decades ago. Thus the functional genomics profile may reflect the effects of living with schizophrenia for decades (including the effect of medications) and not the cause of the illness. However, many studies try to account for this by including these as biological covariates in their analysis if they are found to be affecting the results.

Finally, perhaps the most contentious issue is the use of postmortem tissue itself. There remains debate as to how much postmortem tissue recapitulates the functional genomic activity of the living brain. Indeed, technical factors such as postmortem interval or cause of death have been tested for association with findings with differing results (Girdhar et al., 2018; Bryois et al., 2018). A recent study compared tissue isolated from DLPFC from living and postmortem tissue with expression differing between living and postmortem tissue for 80% of genes suggesting that postmortem tissue is not an accurate proxy for the living brain (Liharska et al., 2023). Yet, this study had a number of limitations including the lack of living tissue from any healthy individuals or living schizophrenia

cases in addition to the use of bulk tissue. It therefore remains to be seen whether the limitations of postmortem tissue is widespread across all studies or just limited to certain conditions.

1.3.5 Integrative Multiomic Studies in Schizophrenia

The integration of the vast omic resources that are now available presents a unique opportunity to understand disease pathophysiology. One of the most common ways to integrate omics data is through quantitative trait loci (QTL) analysis. Molecular QTLs such as expression QTLs (eQTL), methylation QTLs (mQTL) or chromatin accessibility QTLs (caQTL) can be discovered when their respective data types are integrated with genotype data for the same individuals profiled. These data can be used to investigate genomic loci that can explain variation in that molecular phenotype (Aguet et al., 2023). Risk loci discovered through GWAS studies (in this case schizophrenia GWAS) can then be integrated by identifying which QTLs colocalize with risk loci to identify causal genes from GWAS peaks. eQTL analysis in schizophrenia initially suffered from lack of power due to sample sizes. One of the first large-scale RNA-seq studies in schizophrenia identified eQTLs in 20 of the then 108 risk loci with 5 of these loci implicating only one putative causal gene (Fromer et al., 2016; Purcell et al., 2014). All five of these are now considered among the PGC’s high priority risk genes for schizophrenia (Trubetskoy et al., 2022). Other studies have generated large eQTL resources from a number of brain tissues by combining data from multiple cohorts, substantially increasing the number of detected eQTLs (Sieberts et al., 2020). Colocalization of these eQTLs also identified 17 genes as being putatively causal in schizophrenia. This study also replicated findings from previous studies that suggested the existence of few disease-specific eQTLs with many diseases showing overlap of significant eQTLs suggesting the need for more targeted analysis (Sieberts et al., 2020; Fromer et al., 2016; Ng et al., 2017). Other eQTL analysis have identified many eQTLs associated with schizophrenia with extensive difference in both the number of eQTLs and the regions of the genome in which they are detected between brain regions (Jaffe et al., 2018). In addition, many risk loci associated with more than one gene making the characterisation of these QTLs extremely difficult (Jaffe et al., 2018; Collado-Torres et al., 2019). The PsychENCODE project has been at the forefront of QTL efforts in neuropsychiatric disease including not only eQTL but also isoform QTL (isoQTL) and chromatin accessibility QTL (caQTL) which was more extensive than previous studies (Wang et al., 2018).

Characterising the cell specificity of these eQTLs also remains largely unexplored and given that bulk tissue is thought to mainly capture neuronal cell types there may potentially be many cell-specific eQTLs. These studies remain limited particularly in post-mortem tissue. A 2021 study profiled chromatin accessibility using primary human neural

progenitor cell lines in differentiated neuronal and progenitor cells which showed that the majority of caQTLs identified were cell-type specific (Liang et al., 2021) with approximately 30% associated with eQTLs.

Transcription wide association studies (TWAS) also aim to integrate eQTL and GWAS data. TWAS imputes expression data using external eQTL and individual level genotype data or summary statistics and then tests for association between the expression and the trait of interest (Gamazon et al., 2015). New susceptibility genes for schizophrenia have been suggested by TWAS with 35 of the 157 significant TWAS hits not overlapping with the risk loci known through GWAS at the time (Gusev et al., 2018). They also functionally validated their associations using Hi-C data which profiles chromatin interactions. This found that 43 of the 157 genes were associated with chromatin changes which implicates regulatory features that warrants further investigation. Subsequent TWAS also identified novel associations in addition to replicating a large number of the in the previous study (Collado-Torres et al., 2019).

Integrative analysis was undertaken by the PsychENCODE consortium to understand genetic regulation in disease and to improve risk prediction (Wang et al., 2018). Integration of Hi-C data with the QTL types described above enabled the construction of a regulatory network in order to connect non-coding GWAS loci to disease genes with 1111 schizophrenia associated genes identified, 321 of which are supported by more than one type of omic data. They shared functions of previously associated genes such as calcium channel and synaptic involvement but the increase in associated genes highlights the power of integrative analysis (Wang et al., 2018). Finally, a deep learning model was used to incorporate a wide variety of data including gene expression, QTLs, cell-type marker genes, co-expression modules and chromatin states. This allowed for increased prediction efficiency compared to previous models which shows the benefit of using functional and genomic data in tandem.

1.4 Cell-type Deconvolution

Cell-type deconvolution also known as cellular deconvolution aims to estimate the proportion of each cell-type present in a bulk sample (Avila Cobos et al., 2020). Most widely used to ascertain the proportions in bulk transcriptomic data, it can also be applied to other functional genomic data such as DNA methylation or chromatin accessibility data. This section will give an overview of the principles of cell-type deconvolution and some of the most popular methods and tools that are currently available. It will then focus on some of its major uses including its initial primary use in tumour composition as well as

some recent applications to brain expression data which is of most interest in this thesis.

1.4.1 Principles and Methods of Cell-type Deconvolution

Bulk tissue functional genomic studies begin with the isolation of a single tissue that is first homogenised and nucleic acids extracted. A sample is then sent for sequencing. Yet, this tissue is a diverse tapestry of different cell types. By performing bulk analysis we will only gain an insight into the overall trends in the molecular phenotype being assessed. (I will primarily focus on gene expression in this section unless otherwise stated). Gene expression however is incredibly diverse and varied across cell-types, which has been highlighted by cell-specific and single-cell studies (Lake et al., 2017; Darmanis et al., 2015). These cell-specific profiles therefore are represented in the bulk mixture yet we do not know how much of each cell type is present. It is true that cell-isolation techniques can be used to profile individual cell populations but this can be difficult and expensive indicating the need for computational techniques. Deconvolution can also be used to generate new results from historical bulk data. Deconvolution techniques often use prior knowledge of the representative gene expression for each cell-type to build a signature matrix containing expression data for each cell-type. This expression data can either be composed of known marker genes for a cell-type or use single-cell data which is common in more recent methods (Hunt et al., 2019; Wang et al., 2019; Newman et al., 2019). This is then used to decompose the mixture matrix to estimate the proportion of each cell-type for each sample in the mixture (Figure 1.4).

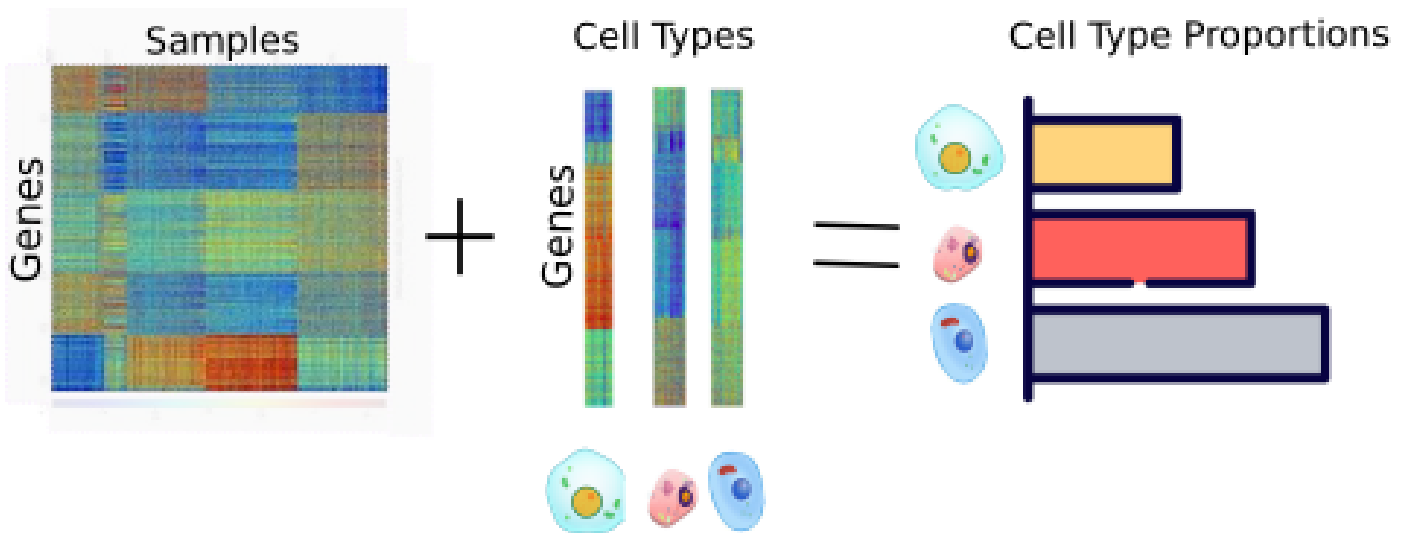


Figure 1.4: Schematic of cell-type deconvolution.

Graphical representation of a common approach to cell-type deconvolution. It shows a bulk gene expression matrix (left) with a number of samples with gene expression information for each sample. Cell-type specific gene expression profiles are then used to construct a signature matrix (centre) and the integration of both enables the estimation of the proportion of each cell-type in the mixture (right)

Different methods can be used to estimate the cell-type proportions. Some of the most widely used tools and their underlying methods are listed in Table 1.3. The vast majority are implemented through R packages and use non-negative matrix factorisation, support vector regression, linear mixed models among many other algorithms (Hunt et al., 2019; Wang et al., 2019; Newman et al., 2019). Other considerations within each method include the type of reference/signature matrix such as cell-specific markers, normalisation or data transformation. Given the large number of variables that have to be taken into account when calculating cell-type proportions, an extensive benchmarking study was carried out in 2020 (Avila Cobos et al., 2020). It was found that for the majority of methods, the normalisation and data transformation method did not greatly impact results. Of the deconvolution methods themselves, methods that used penalised regression methods to estimate proportions performed the worst (Avila Cobos et al., 2020). Additionally, the use of all available cell-type markers resulted in the most accurate proportions, presumably because it takes more of the expression profile into account. Furthermore, semi-supervised approaches which use all available information perform better than unsupervised approaches which do not use labelled training data (Avila Cobos et al., 2020). In this thesis we use dTangle (Hunt et al., 2019). The reason for the choice of this method and further details of the dTangle algorithm and implementation will be described in Chapters 2 and 4.

Table 1.3: Description of the most commonly used deconvolution methods.

Tool	Method	Reference Type	Citation
CibersortX	Support vector regression	Cell-type markers or Single-cell expression data	(Newman et al., 2019)
MuSIC	Weighted non-negative least squares	Single-cell expression data	(Wang et al., 2019)
dTangle	Linear mixing model	Cell-type markers or Single-cell expression data	(Hunt et al., 2019)
DeconRNASeq	Non-negative least squares regression	Cell-type markers	(Gong and Szustakowski, 2013)
Bisque	Non-negative least squares regression	Cell-type markers or Single-cell expression data	(Jew et al., 2020)

1.4.2 Uses of Cell-type Deconvolution

Cell-type deconvolution has historically been used in the study of cancer in order to estimate the number of tumor versus healthy cells (Avila Cobos et al., 2020). It has since been expanded to other areas of study within genomics. In particular, our interest is in its use in studies of the brain and neuropsychiatric disorders. As the number of these studies is more limited, I will first focus on the use of these methods in other tissues before focusing on what these studies have found in studies of the brain in order to demonstrate the utility of these methods.

1.4.2.1 Cell-type Deconvolution in Other Tissues

Deconvolution has been used to great effect in cancer research as mentioned above. Tumours are known to be incredibly heterogeneous and accounting for this heterogeneity can help predict tumour subtypes more accurately (Elloumi et al., 2011). A 2015 study used deconvolution to identify the abundance of different tumour-associated leukocytes present in tumour samples (Gentles et al., 2015). Once quantified, they integrated these proportions with survival statistics in order to assess the validity of using these cell populations for prognostic stratification and to identify candidate genes that could be investigated as therapeutic targets (Gentles et al., 2015). Other studies have profiled the abundance of the tumour microenvironment in bladder cancer finding that M2 macrophages were the most abundant tumour infiltrating immune cell and were associated with poor prognosis (Xue et al., 2019). This could be useful for novel immunotherapies which target these macrophage populations. Further studies have also used deconvolution to profile tumour infiltrating cell abundance in colon cancer and construct an immune risk score based on these relative abundances (Zhou et al., 2019). Similar strategies have also been used in

the study of acute myocarditis, an inflammatory disease which results in cardiac dysfunction (Kawada et al., 2021). Natural killer cells and T cells were the immune infiltrating cells with the highest observed proportions and this was consistent with immunohistochemistry, thus supporting the use of deconvolution methods to explore the landscape of immune infiltrating cells in disease.

1.4.2.2 Cell-type Deconvolution in the Brain

As this thesis is focused on the use of deconvolution in schizophrenia gene expression data, I will now focus on uses of this method in brain tissue. Firstly, deconvolution has been used as a technical covariate in brain gene expression studies. As mentioned, bulk studies take a sample of tissue and profile the expression within this sample. As such, there remains a question as to whether the differences in gene expression between cases and controls are true differences or a reflection of cell composition differences. A 2016 study tested this when carrying out a differential gene expression analysis between schizophrenia cases and controls (Fromer et al., 2016). They found that cell type proportions estimated using deconvolution did not explain significant differential expression. Had they found that the opposite was true, they then would have been able to include cell-type proportions as a covariate in their model in order to accurately identify DEGs. This is particularly useful when combining data from multiple cohorts where cell proportion may differ greatly due to a number of technical and biological reasons such as age, dissection or batch effects. A recent study from the CMC also looked at gene expression differences in schizophrenia by combining multiple cohorts (Hoffman et al., 2022). They estimated cell-type proportions for four cell-types (GABAergic neurons, glutamatergic neurons, oligodendrocytes, and microglia and astrocytes combined) using dTangle and found that factors such as age of death and sex explained some of the variation in these proportions. As a result, the cell-type proportions were used in the differential gene expression model in order to control for these differences and accurately define DEGs.

Another use of deconvolution is to ascertain whether proportions of certain cell-types differ between cases and controls. The 2016 study mentioned above did not observe any differences between cases and controls in the cell-types studied, a finding supported by another study from 2022 (Fromer et al., 2016; Hoffman et al., 2022). Analysis by the PsychENCODE consortium in 2018 expanded the number of cell-types analysed to 24 and found a decrease in an excitatory neuronal population (Ex5) associated with schizophrenia (Wang et al., 2018). Additionally, associations of cell-type proportions in Ex3 and Ex4 with age were also observed which were thought to be due to the gene *SST* which is involved in aging and neurotransmission. This study also identified QTLs associated with cell fractions (fQTLs) (Wang et al., 2018). Excitatory neurons Ex4 and

Ex5 were associated with the most fQTLs. While an eQTL was found to be associated with expression levels of the gene *FZD9* and this SNP was also associated with the proportion of Ex3 cells via an fQTL (Wang et al., 2018). This is of particular interest as *FZD9* is associated with fraction changes in Williams Syndrome which is a rare genetic disorder that among other things causes changes in cognition, a symptom shared by neuropsychiatric disorders (Chailangkarn et al., 2016). Overall, it is clear that estimating cell-type proportions in bulk brain tissue is important from both a technical and a disease perspective.

1.5 Cell-type Expression Imputation

An extension of cell-type deconvolution is cell-type expression imputation. In this case, the proportions that have been ascertained from deconvolution are integrated with bulk data to derive cell-type specific gene expression profiles given that bulk expression data is a mixture of cell-type gene expression that will vary based on the proportion of the cell-types present in that mixture (Wang et al., 2021). It is a novel way of gleaning extra information from historical bulk data and is used in Chapters 2 and 4 of this thesis.

Initial tools for expression imputation used proportions and bulk data to estimate population-average cell-type specific expression (Shen-Orr et al., 2010). Population level imputation involved the creation of an overall expression profile per cell-type but did not give sample-level information. This type of imputation was implemented by the tool csSAM and could be used to conduct cell-type specific differential expression analysis in the absence of single-cell or cell-specific data which makes it a valuable method (Shen-Orr et al., 2010). Further tools have been developed which can generate sample-level imputed gene expression rather than population-average profiles. There are two main tools for this type of analysis, the first of which, CIBERSORTx is an extension of the CIBERSORT deconvolution tool (Newman et al., 2019). It can impute both sample and population level profiles. The other tool and the one which is used in this thesis is known as bMIND and is improved version of an earlier implementation known as MIND (Wang et al., 2020, 2021). Extensive testing of bMIND on brain expression data for autism and Alzheimer’s disease has shown significant overlap with findings from unimputed data and was found to be more accurate than other methods (Wang et al., 2021). Overall, expression imputation tools can be used to generate cell-type specific profiles that may otherwise be unavailable, particularly at scale which can then be used to conduct analysis that would otherwise be impossible such as cell-type specific differential gene expression analysis with certain phenotypes.

1.6 Limitations and Challenges in this Subject Area

The area of neuropsychiatric genomics is fraught with many limitations and challenges. These include the small proportion of variance explained by common risk variants in schizophrenia, the difficulty of identifying rare variants due to small sample sizes in exome sequencing cohorts and the fact that many risk variants for schizophrenia are located in noncoding regions of the genome which makes it difficult to ascertain their functional impact. The two main limitations that I will discuss in relation to this thesis are cell-type specific analysis in schizophrenia and the analysis of gene-sets generated using functional genomics data.

1.6.1 Cell-type Specific Analysis in the Study of Schizophrenia

As evidenced by the discussion of brain postmortem studies of schizophrenia in this chapter, cell-type specific analysis are extremely limited with the majority of studies using bulk tissue. This is because of the difficulty of obtaining large numbers of postmortem samples from individuals diagnosed with neuropsychiatric disorders. In addition, post-mortem brain tissue is often of low quality with the RNA integrity number (RIN, a quality control metric used for RNA) often below eight indicating lower quality RNA for profiling gene expression. The profiling of whole cells is also unachievable due to the structure of neurons resulting in the majority of studies only profiling nuclei resulting in a lack of information regarding cytosolic gene expression. Developments in sequencing technologies have enabled new studies of cell-specific gene expression and other molecular phenotypes. However, these studies are rarely schizophrenia specific and have small samples sizes and do not detect transcript expression changes. It is known that understanding the cell-type specific changes in schizophrenia is of utmost importance with single-cell brain studies showing vastly difference expression profiles (Darmanis et al., 2015; Wang et al., 2018). Therefore, by only utilising bulk tissue we may be missing the vast majority of disease-specific changes. Furthermore, most studies only investigate one type of genomic data and the integration of different molecular phenotypes are needed. We were interested to characterise cell-specific changes in schizophrenia to add to the current knowledge of its neurobiology. We therefore sought to leverage historical bulk data and deconvolution and imputation methods to gain new cell-specific insights using published data. We also generated and analysed a new cell-specific dataset to identify changes in both the transcriptomic and chromatin accessibility landscapes in schizophrenia.

1.6.2 Analysis of Gene-sets Generated Using Functional Genomics Data

Vast numbers of gene-sets relating to a variety of phenotypes are now available as a result of GWAS and functional genomic studies from both human and mouse studies. Characterisation of these gene-sets in terms of the phenotype of interest is of great importance. In keeping with the focus on cell-type specificity in this thesis, we were interested in characterising the cell-type specificity of genes dysregulated as a result of *MEF2C* knockout in mouse models. We were interested in *MEF2C* as it is noted as a key risk gene in schizophrenia. Previous studies have investigated the cell-type enrichment of these gene-sets yet have had to rely on mouse data due to the lack of schizophrenia specific data. We were also interested in characterising co-expression enrichment patterns given that this remains largely unexplored.

1.7 Aims and Objectives of this Thesis

1.7.1 Chapter 2

The first study of this thesis aimed to leverage the bulk transcriptomic data compiled by the CMC to create cell-type specific expression profiles by first using cell-type deconvolution to ascertain the proportions of each cell-group in the population and then using these proportions to impute sample-level gene expression for each individual. This study aimed to i) investigate whether these proportions differed between schizophrenia cases and controls or were associated with a schizophrenia-based PRS, ii) conduct cell-specific differential expression and downstream analysis using these imputed profiles to provide further information on the neurobiology of schizophrenia.

1.7.2 Chapter 3

This analysis aimed to explore the effects and cell type specificity of genes which were dysregulated as a result of *MEF2C* perturbation and were first identified by conditional knockouts of *MEF2C* (Harrington et al., 2016, 2020). This study aimed to i) identify which human cell types these gene-sets are enriched in using open-access scRNA-seq data in schizophrenia cases and controls, ii) use scRNA-seq data from the developing brain to identify whether these genes are enriched in certain human cell types at particular time points iii) use these gene-sets to assess co-expression enrichment patterns in specific cell-types using scRNA-seq data.

1.7.3 Chapter 4

This study aimed to profile cell-type specific changes in gene and transcript expression and chromatin accessibility in schizophrenia. This study profiled both of these molecular phenotypes in schizophrenia cases (n=50) and controls (n=50) for four cell types: GABAergic neurons, glutamatergic neurons, oligodendrocytes and microglia/astrocytes. For each individual, one of each cell-type was profiled resulting in an RNA-seq dataset and ATAC-seq dataset of 400 cells each. Following data generation, we characterised gene expression differences between cases and controls using differential gene expression analysis for each of the four cell types. This study also aimed to identify the extent to which these DEGs overlapped between cell types in order to pinpoint which DEGs were contributing to cell-specific dysregulation. We then aimed to characterise the functions that these DEGs could have impacted using pathway and gene ontology analysis which used both general gene-sets and a targeted approach using synaptic gene-sets from SynGO (Koopmans et al., 2019). This study aimed to leverage the novelty of the cell-type specific transcript expression profiles to pinpoint differential transcript expression which is not captured by current single-cell technologies. We sought to accomplish this through the use of a new meta-analysis tool, remaCor (<https://gabrielhoffman.github.io/remaCor/>) further increase our power to identify dysregulated transcripts (Han and Eskin, 2011; Han et al., 2016; Lee et al., 2017). Pathway analysis was chosen to gain an insight into the functional processes affected by our transcript-level results. using both general gene-sets and a targeted approach using synaptic gene-sets from SynGO.

In terms of the chromatin accessibility, this study aimed to i) identify open chromatin regions in the samples and ascertain the number of novel peaks and the overlap with previous annotations, ii) characterise differences in chromatin accessibility between cases and controls using differential chromatin accessibility analysis for each of the four cell types. Following this, this study aimed to i) characterise the functions that these differentially accessible regions could have impacted using pathway and gene ontology analysis using both general gene-sets and a targeted approach using synaptic gene-sets from SynGO, ii) identify whether these regions are enriched for common variants associated with schizophrenia using LD score regression.

This study also aimed to use canonical correlation analysis and similarity network fusion to break our samples into schizophrenia molecular subtypes. Molecular subtypes are subsets of cases that show similar molecular dysregulation, in this case at the gene-expression level. These were created in order to reanalyse the data on a subtype specific basis. Finally, this study aimed to use deconvolution and expression imputation in two ways. We first aimed to use single-cell data as a reference to find proportions of cell subtypes in

our four cell-types and then impute the expression of these subtypes. We also sought to use our four cell-types as a reference to find proportions of cell-types in bulk CMC gene expression data and then impute the expression of these four cell types using the bulk data. Once the latter analysis was complete, we aimed to detect cell-specific eQTLs in this cohort using matched genotypes and following eQTL detection this study aimed to identify causal SNPs using a colocalization analysis.

Chapter 2

Using Cell-type Deconvolution
to Explore the Role of
Different Cell-types in
Schizophrenia

2.1 Abstract

Schizophrenia (SCZ) is a severe psychiatric disorder which is thought to affect 1% of the population. Functional genomics studies that have integrated bulk expression data with cell-type specific information have indicated that transcriptomic changes vary greatly between cell-types. Yet, cell-type specific studies of SCZ remain limited due to technical challenges, low sample sizes and sequencing costs. We aimed to integrate bulk transcriptomic data derived from SCZ cases and controls with single-nucleus data to gain cell-type specific insights into this disorder. We did so by using cell-type deconvolution to estimate cell-type proportions. We observed differences in cell-type proportions between cases and controls for a number of cell-types including excitatory neurons. We also found that proportions of microglia are associated with a polygenic score for SCZ. We used these proportions to carry out gene expression imputation to generate cell-type specific expression profiles. This enabled us to conduct a cell-type specific differential expression analysis. An overrepresentation analysis of these differentially expressed genes implicated a variety of disease-relevant processes. However, our estimated proportions do differ from previous deconvolution studies which raises questions as to the robustness of our results. Despite this we do observe similar SCZ-relevant functions and trends identified in previous studies. This suggests that this type of analysis may prove useful for gaining new insights from bulk data in future work.

2.2 Introduction

Schizophrenia is a neuropsychiatric disorder which affects approximately 1% of the population (Kahn et al., 2015). It is among the top 15 causes of disability and affects 20.9 million people globally (GBD 2016 Disease and Injury Incidence and Prevalence Collaborators, 2017; Charlson et al., 2018) There are three major categories of symptoms associated with SCZ. Most notably these include positive symptoms such as delusions and hallucinations but extend to negative symptoms that consist of a decline in normal behaviours related to motivation or expression or emotion (Kahn et al., 2015; Correll and Schooler, 2020). Cognitive deficits are also a hallmark of SCZ and include disorientation, poor attention and issues with abstract thinking (McCutcheon et al., 2023).

SCZ is a complex multifactorial disorder and a wide variety of both genetic and environmental factors are believed to influence its development. In terms of environmental associations, drug use, prenatal stress, maternal immune activation, viral infection and season of birth have been shown to increase risk of SCZ (Vaucher et al., 2018; Choudhury and Lennox, 2021; Khandaker et al., 2012; Kahn et al., 2015). From a genetic perspective, twin studies have estimated heritability of SCZ to be approximately 79% (Hilker et al., 2018). However, SCZ has been found to be a polygenic disorder with no single genetic cause. Detection of common variants associated with SCZ has occurred primarily through genome-wide association studies (GWAS). The current and most extensive GWAS identified 342 SNPs across 287 distinct loci in the genome implicating hundreds of genes in the disorder (Trubetskoy et al., 2022). Furthermore, common variants currently only explain 24% of SNP heritability which suggests that other genetic factors also play a role in the development of SCZ. Rare variation in SCZ proved initially difficult to detect given the small sample sizes of exome sequencing cohorts at that time (Purcell et al., 2014). However, further studies with increased sample size have managed to detect rare variants which are exome-wide significant. The most recent large-scale rare variant analysis from the SCHEMA consortium (Singh et al., 2022) identified 10 ultra-rare variants which passed the exome-wide significance level. Downstream analysis of both rare and common variants has aimed to discern functional consequences of this variation. Functionally, synaptic processes in addition to calcium channels have been implicated by both common and rare genetic variants associated with SCZ risk (Trubetskoy et al., 2022; Singh et al., 2022; Pardiñas et al., 2018). Rare and common variants both also point to the dysregulation of postsynaptic functions yet recent results from SCZ GWAS (Trubetskoy et al., 2022) also suggests that a presynaptic pathology may also be involved in SCZ.

Purely genetic studies can only provide a small insight into the etiology of a condition and as such there has been a increased move towards functional genomic studies. Functional

genomics aims to bridge the gap between phenotype and genetic variation by assessing changes in dynamic processes such as gene expression and gene regulation. Initial functional genomics studies in SCZ primarily profiled gene expression in order to understand the neurobiology of the disorder. Early studies explored changes in SCZ mouse models that aimed to recreate the phenotypic observations in SCZ using knockouts of putative SCZ risk genes or the generation of a model that recapitulated a genetic or neurobiological defect associated with SCZ (Lipska and Weinberger, 2000). Examples include models of the *DISC1* gene, 22q11 deletion syndrome (in which SCZ is highly prevalent) and the reelin gene which have had varying levels of success (Uliana et al., 2022). *DISC1* in particular has been under much scrutiny in recent years because despite its success in recapitulating brain morphology and symptomology observed in schizophrenia (Clapcote et al., 2007), its disease relevance is under question. More recent genomic studies such as GWAS do not support its role in SCZ (Trubetskoy et al., 2022; Sullivan, 2013; Pardiñas et al., 2018). Indeed, its discovery was based on cytogenetic and pedigree findings and while its perturbation does produce a related phenotype, it is possible that it does not reflect genetic variation of schizophrenia in the wider population. Additionally, mouse models will not truly reflect the human condition despite phenotypic similarities. As such, there has been a move towards functional studies of postmortem human brain tissue that may better reflect disease associated changes.

Transcriptomic studies of postmortem brain tissue in SCZ has led to a number of findings in recent years. Initial smaller scale studies of gene expression in SCZ used microarray technologies and were more limited due to the number of genes profiled in comparison to modern RNA-seq studies. Yet they are still consistent with subsequent GWAS and more recent functional studies with synaptic dysfunction, the mitochondrial and ubiquitin protease systems and metabolic pathways emerging as relevant to SCZ (Middleton et al., 2002; Mirnics et al., 2000, 2006). Identifying particular genes that showed differential expression did however prove to be more difficult with smaller-scale studies not providing sufficient power for DEG detection (Mistry et al., 2013). The first large-scale study was carried out by the CommonMind consortium (CMC) in 2016 and overcame a number of this issues (Fromer et al., 2016). While they did not find any significantly enriched pathways from their differentially expressed genes (DEGs), they implicated GABA and neurokinin receptors which suggested dysregulation of interneurons. Even larger efforts have followed this, most notably via the PsychENCODE study which contained more than double the number of SCZ cases in the original CMC study (Wang et al., 2018; Gandal et al., 2018). This study not only included gene expression data but also epigenomic data and was outlined in the introduction chapter of this thesis. In terms of gene expression, which is the focus of this chapter, this study provided extensive characterisation of the transcriptome at gene, transcript and splicing level (Gandal et al., 2018). Enriched path-

ways were consistent with results described previously but they also implicated cell-types that are believed to be important in SCZ including excitatory neuronal types and oligodendrocytes. Other findings from this study also pointed towards cell-type effects with co-expression modules and eQTL results pointing towards cell-type specific changes in the context of SCZ. As such there has been much focus on cell-type specific transcriptomic studies in SCZ.

While there has been a vastly increased effort to carry out cell-type specific studies in SCZ, these have been limited as it was previously difficult to carry out such studies. Difficulties included cell-isolation, low throughput technologies and small individual level sample sizes of postmortem tissue cohorts, an issue which has also impacted bulk studies. Initial studies focused on excitatory neuronal subtypes as the glutamatergic system has been consistently implicated in SCZ as antagonists of the NMDA receptor can mimic the cognitive deficits observed in schizophrenia (Newcomer et al., 1999; McCutcheon et al., 2020). Findings from studies on pyramidal cells, a major excitatory neuron type in the brain, found regional expression differences in this cell-type (Arion et al., 2015). DEGs from pyramidal cells in layer 3 mainly implicated the mitochondrion while layer 5 pyramidal cells implicated the ubiquitin protease system. If even cells of the same type show regional differences, we would expect different cell-types to show even more specific alterations. This is indeed true in SCZ. A subsequent cell-specific study profiling inhibitory neurons compared their DEGs to those identified in pyramidal neurons and found that approximately 80% of the DEGs were cell-type specific. As such bulk studies only capture an overall picture of transcriptomic changes and cell-specific studies are needed to truly understand the neurobiology of SCZ.

Cell-type specific or single-cell studies are however expensive to carry out and until recently were extremely limited in the literature. In addition, there is now a substantial amount of bulk transcriptomic data whose sample size is much larger than cell-specific studies. Cell-type deconvolution techniques which estimate cell-type proportions in bulk data provide an opportunity to leverage this bulk data to gain new insights into cell-type specific dysregulation in SCZ. The underlying principle of cell-type deconvolution was explained in detail in Chapter One of this thesis but, in brief, most tools work as follows. They use prior knowledge (usually derived from cell-specific or single cell studies) to build a signature matrix containing expression data for each cell-type (Newman et al., 2019; Hunt et al., 2019). This reference data is then used to estimate proportions for each cell-type for every sample in the bulk expression matrix. These proportions can be useful in two main ways. Firstly, they can be used to assess whether the proportions differ between cases and controls. This is important because if they differ, we must consider if the gene expression differences in bulk data are due to disease specific effects or just

differing proportions of cells in a sample. Secondly, estimated cell-type proportions can be combined with expression imputation techniques such as bMIND (Wang et al., 2021) to infer cell-type expression and leverage bulk data to identify cell-type specific changes.

This study aimed to build on the work carried out by the PsychENCODE consortium (Wang et al., 2018; Gandal et al., 2018) by estimating the cell-type proportions in bulk RNA-seq expression data ($n=494$ controls, $n=354$ SCZ cases) using newer snRNA-seq as a reference to produce a more representative estimate. We then assessed whether these proportions were associated with a polygenic risk score (PRS) for SCZ by using matched genotype data for the individuals present in the bulk RNA-seq data. We then tried to leverage these proportions and the published RNA-seq data-set to gain new insight of cell-type specific changes in SCZ. This was achieved through the use of the expression imputation tool bMIND. We then aimed to run a differential gene expression analysis on these imputed matrices to see if we could identify cell-specific changes. In addition, we ran gene-set enrichment analysis to infer functional processes implicated by the genes from each cell-type. We also assessed if DEGs overlap with common risk variants associated with SCZ and other related traits using GWAS summary statistics. Overall, this study aimed to gain new cell-specific insights into SCZ from published bulk tissue functional genomics data.

2.3 Methods

2.3.1 Dataset Description and Processing

Gene expression data from bulk RNA-seq ($n=494$ controls, $n=354$ SCZ cases) were obtained from the CMC, which is a large-scale resource for functional genomics studies of neuropsychiatric disorders and has been described previously in this thesis (Hoffman et al., 2019). Our dataset is restricted to those with matching genotype data for the purposes of downstream analysis. Reads were trimmed using the adapter trimming tool, Trimmomatic (v0.36) (Bolger et al., 2014) and assessed for quality using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/). They were then mapped to the GRCh38 version of the human reference genome using the alignment tool STAR (v2.5.3a) (Dobin et al., 2013). The quantification tool RSEM (v1.3.0) (Li and Dewey, 2011) was used to quantify reads at transcript level, which were then summarised at the gene level for the final raw count matrix.

2.3.2 Cell-type Deconvolution

We sought to estimate the proportion of cell-types present in the bulk RNA-seq data described above using the cell type deconvolution tool dTangle (Hunt et al., 2019). dTangle estimates cell-type proportions using a linear mixed model and was chosen for this purpose as it has been shown to have among the most accurate tools for deconvolution in brain expression data (Sutton et al., 2022). Cell-type deconvolution requires a bulk expression matrix in which the proportion of each cell-type per sample is estimated, which is the CMC data described above. It also requires a reference data-set that provides cell-type specific expression data or cell-type specific markers which are used as a reference to estimate the proportions in the bulk data. To this end, we used a single-nucleus RNA-seq (snRNA-seq) data-set from a multi-omic study that profiled single-nucleus gene expression in the neocortex (Zhu et al., 2023). We filtered this data-set to only include adult samples. It contained ten distinct cell-types (excitatory neurons, medial ganglionic eminence derived inhibitory neurons (IN-MGE), caudal ganglionic eminence derived inhibitory neurons (IN-CGE), oligodendrocyte precursor cells (OPCs), astrocytes, oligodendrocytes, microglia, endothelial cells, pericytes and vascular smooth muscle cells (VSMC)). We filtered both the bulk and snRNA-seq expression matrices in order to remove lowly expressed genes, i.e. those which had a Count Per Million (CPM) < 1 in 30% of samples. We converted the snRNA-seq data “pseudobulk” by summing the expression values across all cells in a cell-type. Both count matrices were normalized using the Trimmed Mean of M-values (TMM) method using the limma R package and transformed to $\log(\text{CPM}+1)$. We then subsetted these matrices to include only genes that were common between them and combined them into one matrix which is protocol for dTangle. The `normalizeBetweenArrays()`

function within limma was then used to normalise expression intensities between the bulk and snRNA-seq data. We selected marker genes for the snRNA-seq data through the use of the cell-type marker genes identified in the original study (Zhu et al., 2023). We then estimated the proportions of each cell-type in the bulk matrix using the dtangle() function. This function requires as input: i) our combined FANS cell-type and reference expression matrix, ii) the list of marker genes per cell-type, iii) a vector containing information on which samples of the combined matrix are to be considered reference profiles and iv) the data-type which was “rna-seq”.

2.3.3 Polygenic Risk Scoring

In order to compute PRS for each individual in our bulk expression matrix we used the PRS software, PRSice-2 (Choi and O’Reilly, 2019). GWAS summary statistics from the latest SCZ GWAS were used as the “base input” to PRSice. While, the “target dataset” used was matching genotype data from the CMC individuals included in our bulk expression data, which is publicly available (Hoffman et al., 2019). The genotype data was converted from variant call format (VCF) to PLINK format using the PLINK software (V1.9) (Chang et al., 2015) (<https://www.cog-genomics.org/plink/>). Standard quality control measures were carried out using PLINK. We removed SNPs that were missing in at least 10% of individuals (`-geno 0.1`) and no individuals were removed for excessive missingness (`-mind 0.1`). The minor allele frequency (MAF) threshold used was 0.05 (`-maf 0.05`) and Hardy-Weinberg Equilibrium (HWE) outliers were also removed (`-hwe 1e-6`). PRSice uses a thresholding and clumping method to calculate a PRS for each individual. It sums up the number of minor alleles for each SNP and multiplies this by the GWAS effect size. Both were supplied to PRSice using default settings and the resulting PRS was used to assess whether there was an association between the proportions calculated using cell-type deconvolution and PRS.

2.3.4 Gene Expression Imputation

The proportions estimated by dTangle were used as input to impute cell-type specific gene expression profiles using bMIND (Wang et al., 2021). bMIND can estimate per-sample cell-type specific gene expression profiles. It takes a bulk expression matrix and an estimate of the proportion of each cell-type present in each bulk sample as input. It uses the Markov chain Monte Carlo (MCMC) model to impute this gene expression (Wang et al., 2021). This created ten cell-type specific count matrices ($n=494$ controls, $n=354$ SCZ cases), (Excitatory neurons, IN-MGE, IN-CGE, OPCs, Astrocytes, Oligodendrocytes, Microglia, Endothelial cells, Pericytes, VSMC).

2.3.5 Differential Gene Expression Analysis

We conducted a differential gene expression analysis of the imputed count matrices to identify disease associated expression changes. Each imputed count matrix was filtered to include only genes where the CPM was ≥ 1 in at least 20% of the samples. We used the Bayesian information criterion to select covariates which resulted in sex, GC coverage and the unmapped read percentage being included in the model for certain cell types. The count matrices for each cell-type were normalised using the TMM method and then transformed to $\log_2(\text{CPM}+1)$ using the `voom()` function in the `limma` R package (Ritchie et al., 2015). In order to identify the DEGs, we used the `dream()` function from the `variancePartition` R package (Hoffman and Schadt, 2016; Hoffman and Roussos, 2021) and applied a multiple testing correction ($\text{FDR} < 5\%$).

2.3.6 Gene-set Analysis

We used `ConsensusPathDB` to carry out an overrepresentation analysis of the DEGs detected for each cell-type using the genes expressed in the cell-type tested as the background. `ConsensusPathDB` is a browser-based database that integrates information from various pathway and gene-ontology databases. Each set of DEGs was uploaded to the `ConsensusPathDB` website where `ConsensusPathDB` calculated a P value for each pathway/GO term available in its database using a hypergeometric test. Hypergeometric tests are used to investigate whether the ratio of the number of genes in a set that map to a particular pathway or GO term being tested is significantly greater than those that do not map to that pathway/term.

We also employed a targeted approach for the three neuronal DEG gene-sets. This was carried out using the synaptic gene ontology resource `SynGO` (Koopmans et al., 2019). We uploaded each neuronal gene-set to `SynGO` individually to identify possible synaptic locations implicated by our gene-sets.

We also used the gene-set analysis tool `MAGMA` to test enrichment of the DEGs per cell type in GWAS summary statistics (de Leeuw et al., 2015) (<http://ctg.cncr.nl/software/magma>). We used summary statistics from published GWAS on SCZ (Trubetskoy et al., 2022) ($n=76,755$ SCZ cases, $n=243,649$ controls), intelligence (IQ) (Savage et al., 2018) ($n=269,867$) and educational attainment (EA) (Lee et al., 2018) ($n=766,345$). We mapped SNPs with available GWAS results to genes using GRCh37/hg19 coordinates. Gene P-values are calculated for each set of GWAS summary statistics and the genes were padded by 35kb upstream and 10kb downstream in order to include regulatory region variants. We estimated linkage disequilibrium (LD) using the European panel of 1000 Genomes Project. The gene based analysis in `MAGMA` is based on a multiple linear principal components

regression model. This model accounts for LD between SNPs in each gene, the number of SNPs in each gene, the GWAS sample size and the inverse of the mean minor allele count of variants per gene. We corrected for multiple testing using Benjamini-Hochberg correction to compute Q values.

2.4 Results

2.4.1 Cell-type Deconvolution

We took bulk RNA-seq data generated by the CMC (Hoffman et al., 2019) in addition to scRNA-seq data and processed it as described in Methods. Using both datasets we estimated the proportions of the ten cell-types in each sample present in the bulk expression matrix. These proportions are illustrated using boxplots in Figure 2.1. Excitatory neurons had a mean proportion of 0.151, while the IN-MGE and IN-CGE neurons had mean proportions of 0.047 and 0.055 respectively. Oligodendrocytes were estimated to have an average proportion of 0.111 while OPCs has a mean proportion of 0.032. Microglia had an average proportion of 0.062 while endothelial cells, pericytes and VSMCs has proportions of 0.202, 0.154 and 0.114 respectively.

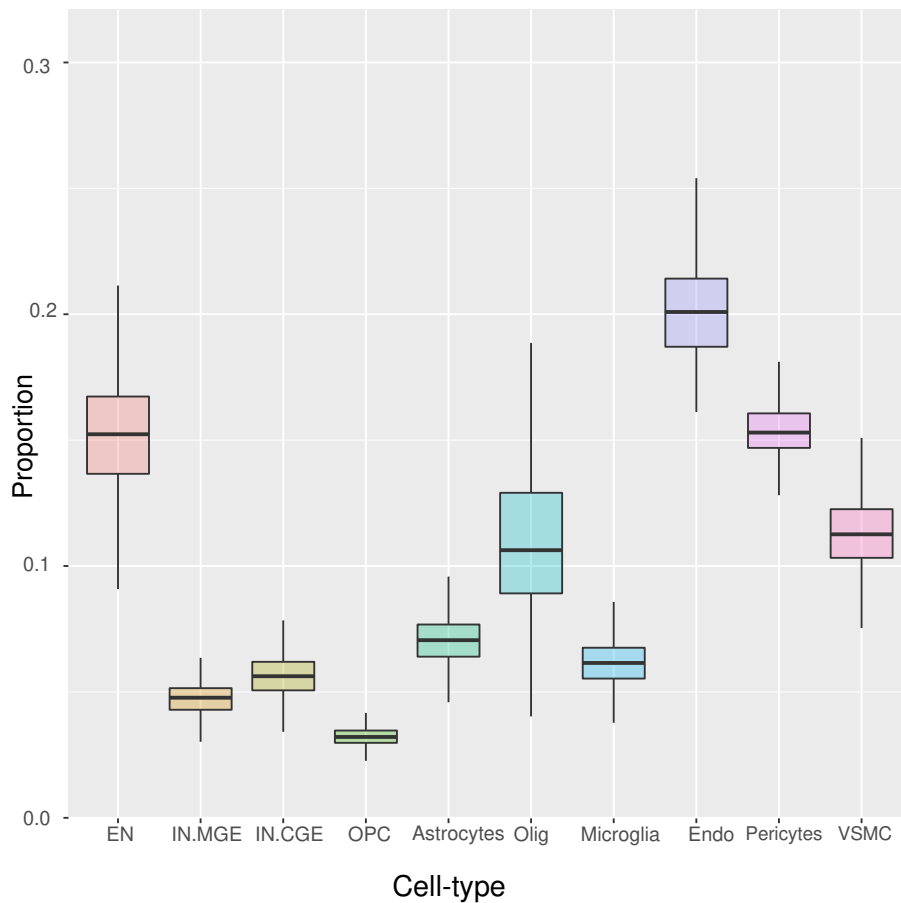


Figure 2.1: Boxplots showing the distribution of the proportion of each cell-type across the samples in the bulk expression matrix. The proportions of the following cell-types were estimated, excitatory neurons (EN), medial ganglionic eminence derived inhibitory neurons (IN-MGE), caudal ganglionic eminence derived inhibitory neurons (IN-CGE), oligodendrocyte precursor cells (OPC), astrocytes, oligodendrocytes (Olig), microglia, endothelial cells (Endo), pericytes and vascular smooth muscle cells (VSMC)

We also assessed whether these proportions differed between cases and controls. Excitatory neurons ($P=0.003$), IN-CGE ($P=0.036$), astrocytes ($P=0.016$) and VSMCs ($P=0.00003$) were found to differ between cases and controls (Table 2.1). Excitatory neurons and IN-CGE were increased in controls versus SCZ cases, while astrocytes and VSMC were increased in SCZ cases compared to controls (Figure 2.2). However only excitatory neurons and VSMCs survived multiple testing correction.

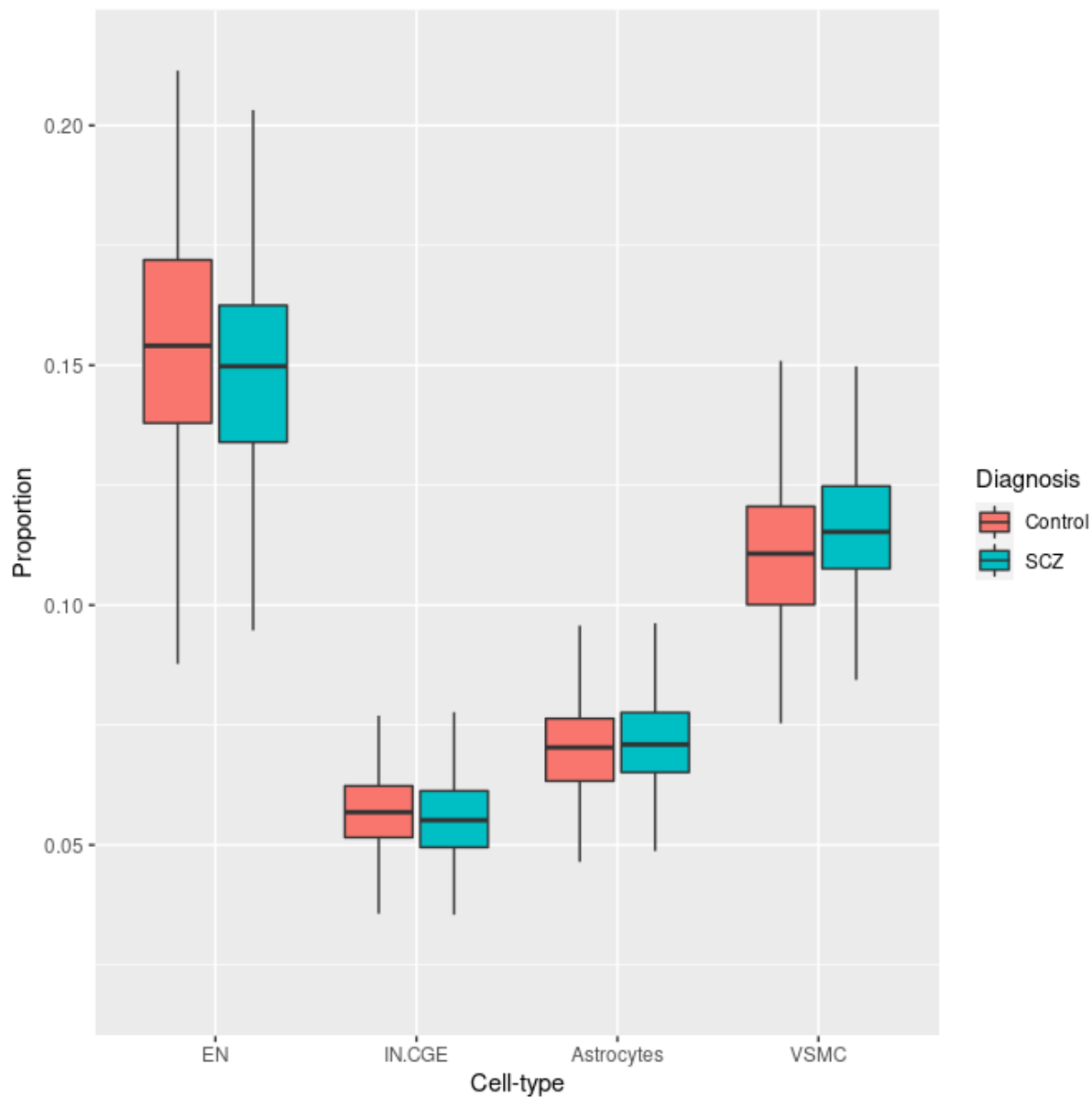


Figure 2.2: Boxplots showing the difference in proportion between cases (SCZ) and controls for four cell-types which showed differences based on diagnosis. The following cell-types are included excitatory neurons (EN), caudal ganglionic eminence derived inhibitory neurons (IN.CGE), astrocytes and vascular smooth muscle cells (VSMC)

Table 2.1: Results from testing the cell-type proportions for differences between cases and controls using a two-sample t-test. P-values were corrected for multiple testing using the Bonferroni method.

Cell-type	P-Value	Adjusted P-Value
Excitatory neurons	0.003	0.03
IN-CGE	0.036	0.36
IN-MGE	0.172	1
Oligodendrocyte	0.79	1
OPCs	0.308	1
Astrocytes	0.016	0.16
Microglia	0.569	1
Pericytes	0.22	1
Endothelial cells	0.645	1
VSMC	0.00003	0.003

2.4.2 Association of Cell-type Proportions with A PRS

We sought to determine if any of the cell-type proportions were associated with a PRS for SCZ. PRS were calculated for each of the individuals in the bulk expression matrix using matched genotype data and GWAS summary statistics from the latest SCZ GWAS (Trubetskoy et al., 2022). Following calculation of the PRS, we assessed for each cell-type whether the estimated proportion using a linear model. We observed an association between decreased microglia proportions and increased PRS ($P=0.0001$). However we did not identify an association between any of the other cell-type proportions and a PRS for SCZ (Table 2.2).

Table 2.2: Results from testing the association of cell-type proportions with a PRS using a linear model. P-values were corrected for multiple testing using the Bonferroni method.

Cell-type	P-Value	Adjusted P-Value
Excitatory neurons	0.906	1
IN-CGE	0.352	1
IN-MGE	0.868	1
Oligodendrocytes	0.812	1
OPCs	0.428	1
Astrocytes	0.942	1
Microglia	0.0001	0.001
Pericytes	0.051	0.51
Endothelial cells	0.0263	0.263
VSMC	0.052	0.52

2.4.3 Gene Expression Imputation

bMIND imputed individual-level expression data for each cell type (Excitatory neurons, IN-MGE, IN-CGE, OPCs, Astrocytes, Oligodendrocytes, Microglia, Endothelial cells, Pericytes, VSMC). We extracted this information from the bMIND results and constructed a count matrix for each cell-type that was used for further analysis.

2.4.4 Differential Gene Expression Analysis

Using the count matrices, we carried out a DGE analysis to identify disease-associated expression changes per cell-type. We included covariates such as GC coverage, unmapped read percentage, sex and institution ID as covariates using the BIC as described in methods. The full list of DEGs per cell-type are listed in Supplementary Table 2.1 (https://drive.google.com/drive/folders/11R3Ccix_ccf4pWhXY6AjqtoLdI0Qym86?usp=sharing).

Table 2.3: Number of DEGs in the imputed expression matrices per cell-type

Cell Type	Number of DEGs (FDR < 0.05)
Excitatory neurons	1808
IN-CGE	2928
IN-MGE	772
Oligodendrocyte	1132
OPCs	4052
Astrocytes	4049
Microglia	1404
Pericytes	3772
Endothelial cells	3292
VSMC	1016

2.4.5 Gene-set Analysis

Gene ontology (GO) and pathway analysis was used to infer the pathways and functions that may be impacted by the DEGs listed above. It revealed a variety of implicated functions per cell-type. The excitatory neuronal DEGs were enriched for genes involved in neuronal processes. The most significant GO terms included “nervous system development” ($Q = 1.94e-08$), “neuron part” ($Q = 7.35e-09$), and “neuron projection” ($Q = 4.28e-07$) (Table 2.4). While the most significant pathways included signal transduction ($Q = 1.73e-06$) and potassium channels ($Q = 0.0362$). MGE DEGs implicated “neural cell adhesion molecule (NCAM1) interactions” ($Q=0.000804$) and “potassium ion transmembrane transport” ($Q=0.00658$) (Table 2.5). While CGE DEGs implicated the “Calcium signalling pathway” ($Q=0.0216$), “Nervous system development” ($Q=6.04e-11$) and “dendrite tree” ($Q=2.49e-09$) (Table 2.6). The non-neuronal cell-types also implicated a variety of pathways and functions. “Regulation of Signalling” ($Q=5.13e-07$) and “Potassium Channels” ($Q=0.0131$) were among the top terms for oligodendrocytes and OPCs respectively (Tables 2.7-2.8). Microglia and astrocytes were enriched for genes associated with “Cell projection” ($Q=5.64e-08$) and “Signal Transduction” ($Q=2.12e-05$) (Tables 2.9-2.10). Finally, endothelial cells, pericytes and VSMCs implicated “Collagen biosynthesis and modifying enzymes” ($Q=0.00424$), “Extracellular matrix organization” ($Q=0.00315$), and “Signal transduction” ($Q=0.0042$) (Tables 2.11-2.13). The GO and pathway enrichment results are available in full in Supplementary Tables 2.2 and 2.3 (https://drive.google.com/drive/folders/11R3Ccix_ccf4pWhXY6Ajqt0LdI0Qym86?usp=sharing).

Table 2.4: Top 5 GO terms and top 5 pathways for Excitatory Neurons

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0007399 nervous system development	2351	274 (11.7%)	1.57e-11	1.94e-08
GO:0097458 neuron part	1748	214 (12.2%)	5.44e-11	7.35e-09
GO:0036477 somatodendritic compartment	828	118 (14.3%)	3.38e-10	5.95e-08
GO:0043005 neuron projection	1326	166 (12.5%)	2.06e-09	4.28e-07
GO:1901700 response to oxygen-containing compound	1592	190 (11.9%)	5.87e-09	3.65e-06
Signal Transduction	2432	272 (11.2%)	7.06e-10	1.73e-06
Signaling by GPCR	706	94 (13.4%)	5.15e-07	0.00063
Nuclear Events (kinase and transcription factor activation)	61	17 (27.9%)	3.86e-06	0.00272
GPCR downstream signalling	633	83 (13.2%)	4.46e-06	0.00272
Signaling by Receptor Tyrosine Kinases	464	64 (13.8%)	1.28e-05	0.00624

Table 2.5: Top 5 GO terms and top 5 pathways for IN-MGE

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0031224 intrinsic component of membrane	5568	242 (4.4%)	1.85e-09	1.72e-07
GO:0016021 integral component of membrane	5401	236 (4.4%)	2.08e-09	2.64e-07
GO:0031012 extracellular matrix	566	44 (7.8%)	3.6e-08	1.67e-06
GO:0062023 collagen-containing extracellular matrix	426	36 (8.5%)	8.8e-08	5.59e-06
GO:0006812 cation transport	1143	65 (5.7%)	2.71e-06	0.00193
NCAM1 interactions	37	9 (24.3%)	7.09e-07	0.000804
Extracellular matrix organization	287	23 (8.0%)	8.48e-06	0.00389
Collagen formation	91	12 (13.2%)	1.03e-05	0.00389
Collagen biosynthesis and modifying enzymes	67	10 (14.9%)	1.87e-05	0.00531
NCAM signaling for neurite out-growth	59	9 (15.3%)	4.11e-05	0.00932

Table 2.6: Top 5 GO terms and top 5 pathways for IN-CGE

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0005737 cytoplasm	11920	1761 (14.8%)	1.8e-14	3.95e-12
GO:0007399 nervous system development	2351	434 (18.5%)	4.26e-14	6.04e-11
GO:0044424 intracellular part	14850	2121 (14.3%)	1.14e-13	1.39e-11
GO:0044444 cytoplasmic part	10011	1507 (15.1%)	1.46e-13	1.6e-11
GO:0005622 intracellular	14857	2121 (14.3%)	1.59e-13	1.39e-11
Glycosaminoglycan metabolism	119	35 (29.4%)	2.55e-06	0.00798
Ciliopathies	183	45 (24.6%)	2.03e-05	0.0216
Calcium signalling pathway - Homo sapiens (human)	240	55 (23.0%)	2.07e-05	0.0216
Signal Transduction	2432	383 (15.8%)	3e-05	0.0235
Nuclear Events (kinase and transcription factor activation)	61	20 (32.8%)	6.57e-05	0.0412

Table 2.7: Top 5 GO terms and top 5 pathways for Oligodendrocytes

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0032879 regulation of localization	2686	200 (7.5%)	9.58e-11	4.98e-08
GO:0051239 regulation of multicellular organismal process	3158	224 (7.1%)	4.78e-10	1.24e-07
GO:0051270 regulation of cellular component movement	1013	93 (9.2%)	1.6e-09	1.59e-06
GO:0023051 regulation of signaling	3492	239 (6.9%)	3.06e-09	5.13e-07
GO:0040012 regulation of locomotion	1001	91 (9.1%)	3.95e-09	5.13e-07
TYROBP causal network in microglia	60	14 (23.3%)	1.01e-06	0.00193
Signal Transduction	2432	166 (6.8%)	4.28e-06	0.00408
Glycosaminoglycan metabolism	119	18 (15.1%)	2.25e-05	0.0143
Synthesis of IP3 and IP4 in the cytosol	29	8 (27.6%)	6.04e-05	0.0288
Metabolism	1954	131 (6.7%)	0.000129	0.0491

Table 2.8: Top 5 GO terms and top 5 pathways for OPCs

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0051239 regulation of multicellular organismal process	3158	754 (23.9%)	4.99e-22	3.73e-19
GO:0050793 regulation of developmental process	2592	612 (23.6%)	1.65e-16	6.15e-14
GO:0007399 nervous system development	2351	554 (23.6%)	1.14e-14	1.74e-11
GO:0048731 system development	4796	1023 (21.4%)	5.5e-14	1.37e-11
GO:0023051 regulation of signaling	3492	773 (22.2%)	9.93e-14	1.85e-11
Signal Transduction	2432	538 (22.1%)	2.94e-10	9.88e-07
Glycosaminoglycan metabolism	119	44 (37.0%)	4.14e-07	0.000695
RHOD GTPase cycle	51	23 (45.1%)	5.32e-06	0.00596
Extracellular matrix organization	287	80 (27.9%)	1.08e-05	0.0091
Potassium Channels	103	36 (35.0%)	1.95e-05	0.0131

Table 2.9: Top 5 GO terms and top 5 pathways for Microglia

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0051239 regulation of multicellular organismal process	3158	278 (8.8%)	1.74e-10	9.67e-08
GO:0042995 cell projection	2278	212 (9.3%)	4.75e-10	5.65e-08
GO:0120025 plasma membrane bounded cell projection	2176	204 (9.4%)	5.87e-10	9.74e-08
GO:0031344 regulation of cell projection organization	702	84 (12.0%)	4.85e-09	5.04e-06
GO:0031012 extracellular matrix	566	72 (12.7%)	4.89e-09	2.91e-07
Neuronal System	391	49 (12.5%)	1.07e-06	0.00207
Signal Transduction	2432	194 (8.0%)	1.57e-05	0.0153
Collagen biosynthesis and modifying enzymes	67	14 (20.9%)	4.04e-05	0.023
Collagen chain trimerization	44	11 (25.0%)	4.73e-05	0.023
Extracellular matrix organization	287	35 (12.2%)	6.48e-05	0.0252

Table 2.10: Top 5 GO terms and top 5 pathways for Astrocytes

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0005737 cytoplasm	11920	2313 (19.4%)	1.71e-17	4.2e-15
GO:0044444 cytoplasmic part	10011	1977 (19.8%)	2.57e-16	3.16e-14
GO:0051239 regulation of multicellular organismal process	3158	710 (22.5%)	1.05e-14	7.73e-12
GO:0098588 bounding membrane of organelle	2165	503 (23.3%)	8.56e-13	7.02e-11
GO:0044424 intracellular part	14850	2764 (18.7%)	2.19e-12	2.89e-10
Glycosaminoglycan metabolism	119	52 (43.7%)	1.89e-11	6.17e-08
Signal Transduction	2432	520 (21.4%)	1.3e-08	2.12e-05
TYROBP causal network in microglia	60	28 (46.7%)	1.55e-07	0.000169
RHO GTPase cycle	442	117 (26.5%)	8.92e-07	0.000661
Integrin	124	44 (35.5%)	1.01e-06	0.000661

Table 2.11: Top 5 GO terms and top 5 pathways for Endothelial cells

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0005737 cytoplasm	11920	1960 (16.5%)	7.11e-20	1.65e-17
GO:0044444 cytoplasmic part	10011	1675 (16.8%)	1.16e-17	1.34e-15
GO:0051239 regulation of multicellular organismal process	3158	612 (19.4%)	1.04e-15	7.35e-13
GO:0005515 protein binding	14003	2217 (15.9%)	3.19e-15	2.23e-13
GO:0007399 nervous system development	2351	475 (20.2%)	3.67e-15	5.3e-12
Extracellular matrix organization	287	74 (25.8%)	6.22e-07	0.00201
Signal Transduction	2432	433 (17.8%)	2.64e-06	0.00424
Collagen biosynthesis and modifying enzymes	67	25 (37.3%)	4.57e-06	0.00424
Collagen chain trimerization	44	19 (43.2%)	5.26e-06	0.00424
Collagen formation	91	30 (33.0%)	9.94e-06	0.00641

Table 2.12: Top 5 GO terms and top 5 pathways for Pericytes

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0005737 cytoplasm	11920	2120 (17.8%)	7.32e-15	1.66e-12
GO:0051239 regulation of multicellular organismal process	3158	653 (20.7%)	2.05e-13	1.52e-10
GO:0007399 nervous system development	2351	504 (21.5%)	6.91e-13	1.04e-09
GO:0005515 protein binding	14003	2424 (17.4%)	8.3e-13	6.06e-11
GO:0044444 cytoplasmic part	10011	1802 (18.0%)	9.08e-13	1.03e-10
Signal Transduction	2432	475 (19.6%)	1.39e-07	0.000315
Extracellular matrix organization	287	80 (27.9%)	1.91e-07	0.000315
Glycosaminoglycan metabolism	119	39 (32.8%)	4.4e-06	0.00484
EGFR1	455	107 (23.6%)	1.35e-05	0.0105
miR-509-3p alteration of YAP1-ECM axis	18	11 (61.1%)	1.83e-05	0.0105

Table 2.13: Top 5 GO terms and top 5 pathways for VSMC

Gene Ontology Term/Pathway	Set Size	Candidates	P-value	Q-value
GO:0048731 system development	4796	291 (6.1%)	2e-09	9.8e-07
GO:0051239 regulation of multicellular process	3158	206 (6.5%)	5.26e-09	1.29e-06
GO:0051241 negative regulation of multicellular process	1181	96 (8.1%)	8.43e-09	8.26e-06
GO:0007275 multicellular organism development	5363	314 (5.9%)	1.64e-08	2.75e-06
GO:0043167 ion binding	6337	357 (5.7%)	4.46e-08	2.37e-06
Effects of PIP2 hydrolysis	27	8 (29.6%)	1.44e-05	0.0218
Glycerolipid metabolism - Homo sapiens (human)	61	10 (16.4%)	0.000304	0.148
White fat cell differentiation	32	7 (21.9%)	0.000402	0.148
Triacylglyceride synthesis	24	6 (25.0%)	0.000488	0.148
RHO GTPase cycle	442	35 (7.9%)	0.000588	0.148

We also performed a targeted analysis for the three neuronal gene-sets using the synaptic ontology resource, SynGO (Koopmans et al., 2019). This identified various synaptic locations which were implicated by the respective gene-sets. The DEGs from excitatory neurons implicated a variety of locations including the synapse, presynaptic membrane, vesicle and active zone (Figure 2.3). Postsynaptic locations included the postsynaptic membrane and endosome. In the IN-MGE derived neuron DEGs, presynaptic locations were more abundant including the presynaptic vesicle, presynaptic dense core vesicles and the cytoskeleton (Figure 2.4). IN-CGE derived DEGs implicated both pre and postsynaptic locations including presynaptic vesicle, dense core vesicles, cytosol and membrane 2.5. Postsynaptic locations included the endosome, membrane and cytosol.

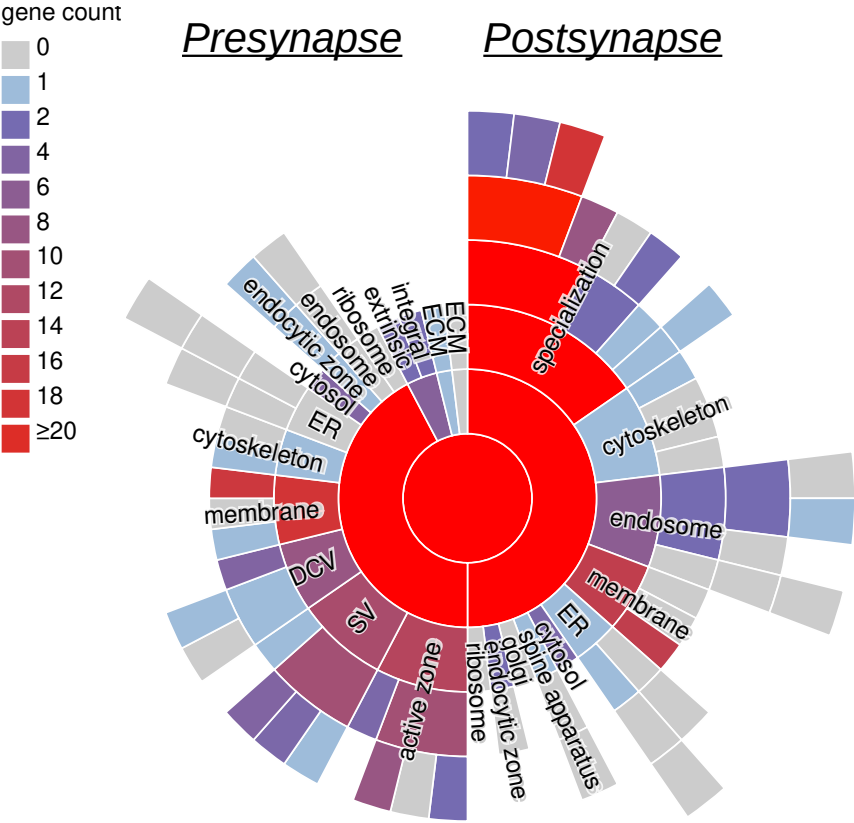


Figure 2.3: Sunburst plot depicting synaptic locations implicated by the DEGs from excitatory neurons using the SynGO ontology resource. This plot features the synapse at the center, pre- and post-synaptic locations in the first ring, and child terms in subsequent rings. The color scheme in the legend indicates the number of genes within each term.

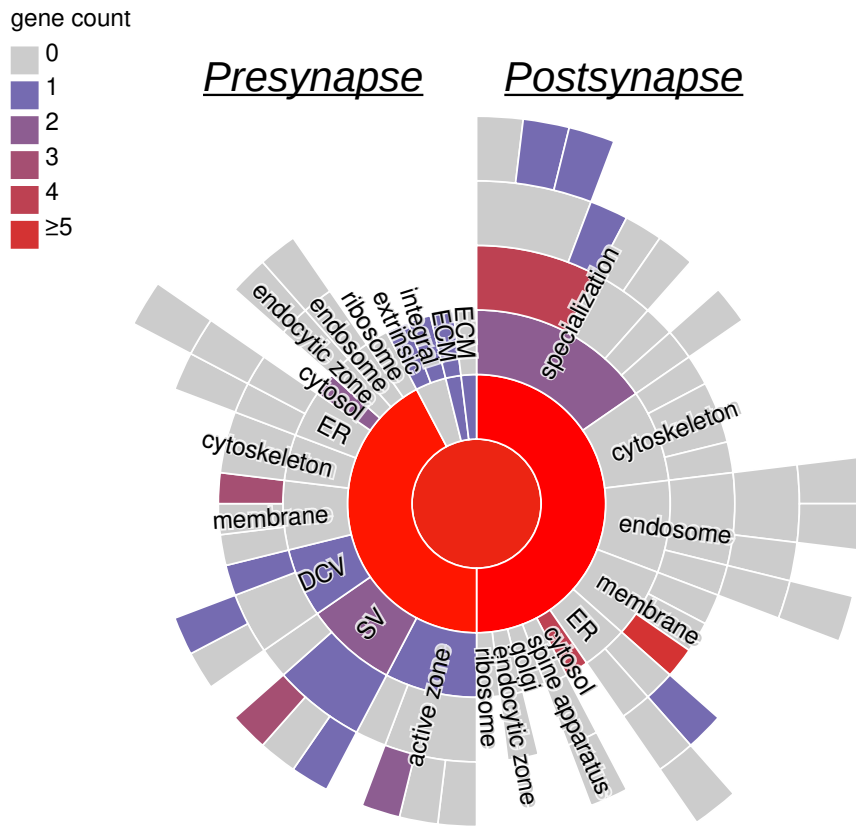


Figure 2.4: Sunburst plot depicting synaptic locations implicated by the DEGs from IN-MGE using the SynGO ontology resource. This plot features the synapse at the center, pre- and post-synaptic locations in the first ring, and child terms in subsequent rings. The color scheme in the legend indicates the number of genes within each term.

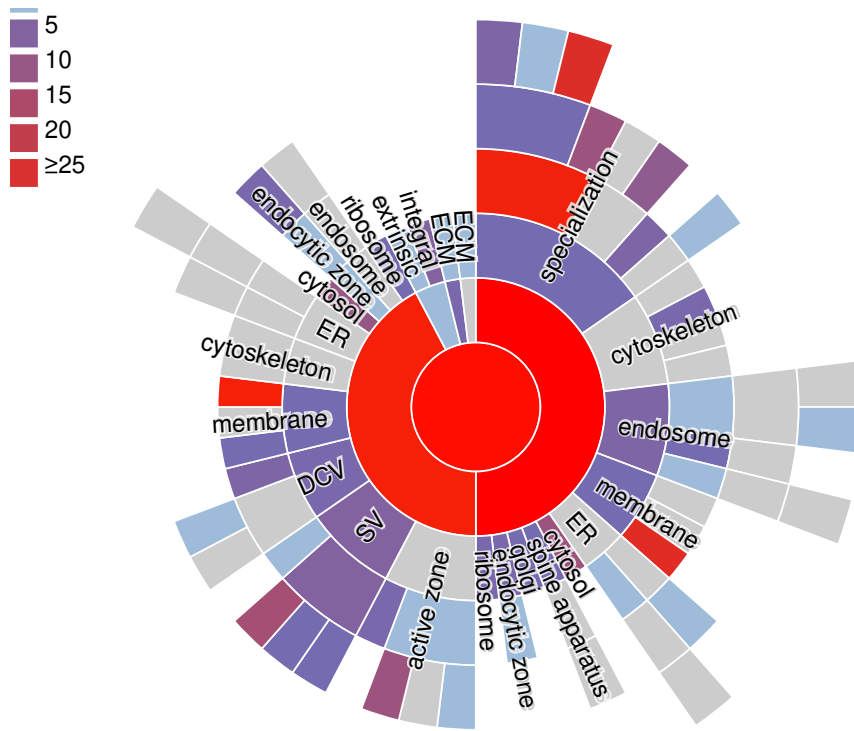


Figure 2.5: Sunburst plot depicting synaptic locations implicated by the DEGs from IN-CGE using the SynGO ontology resource. This plot features the synapse at the center, pre- and post-synaptic locations in the first ring, and child terms in subsequent rings. The color scheme in the legend indicates the number of genes within each term.

We used MAGMA gene-set analysis to identify if any of the DEG sets were enriched for genes containing common variants associated with SCZ, and the cognitive phenotypes of IQ and educational attainment (EA). We used summary statistics for EA and IQ as SCZ is associated with cognitive deficits and these GWAS contained large sample sizes which increases power to detect enrichment. No DEG set was enriched for variants associated with EA or IQ. IN-CGE ($P=0.031$), astrocytes ($P=0.001$), pericytes ($P=0.005$), endothelial cells ($P=0.026$) and OPCs ($P=0.048$) were significantly enriched for genes containing SNPs associated with SCZ (Table 2.14). However only astrocytes remained significant following multiple testing correction ($Q=0.045$).

Table 2.14: Results from MAGMA gene-set analysis of the DEGs per cell-type using GWAS data for SCZ, IQ and EA. Q-values were calculated using the Benjamini-Hochberg method for multiple testing correction.

Cell Type	Phenotype	Beta	P-value	Q-value
Excitatory Neurons	SCZ	0.043	0.569	0.767
IN-MGE	SCZ	-0.041	0.827	0.885
IN-CGE	SCZ	0.043	0.031	0.231
Microglia	SCZ	0.044	0.083	0.414
Astrocytes	SCZ	0.06	0.001	0.045
Pericytes	SCZ	0.054	0.005	0.077
VSMC	SCZ	0.041	0.139	0.534
Endothelial Cells	SCZ	-0.041	0.026	0.231
Oligodendrocytes	SCZ	0.014	0.341	0.612
OPCs	SCZ	0.034	0.048	0.29
Excitatory Neurons	EA	0.02	0.403	0.637
IN-MGE	EA	-0.044	0.824	0.885
IN-CGE	EA	0.006	0.265	0.549
Microglia	EA	-0.04	0.874	0.885
Astrocytes	EA	0.013	0.275	0.549
Pericytes	EA	0.022261	0.162	0.534
VSMC	EA	-0.032	0.782	0.885
Endothelial Cells	EA	0.023	0.168	0.534
Oligodendrocytes	EA	-0.0459	0.885	0.885
OPCs	EA	0.02	0.182	0.534
Excitatory Neurons	IQ	-0.006	0.588	0.767
IN-MGE	IQ	-0.002	0.518	0.74
IN-CGE	IQ	0.015	0.255	0.549
Microglia	IQ	-0.014	0.677	0.846
Astrocytes	IQ	0.017	0.196	0.534
Pericytes	IQ	0.008	0.347	0.611
VSMC	IQ	0.0008	0.49	0.735
Endothelial Cells	IQ	0.006	0.382	0.636
Oligodendrocytes	IQ	-0.041	0.882	0.885
OPCs	IQ	0.015	0.224	0.549

2.5 Discussion

Functional genomics data can provide a valuable insight into the biology of a complex disorder like SCZ. Much of the existing datasets, particularly at the time this study was designed, are however comprised of bulk data. Bulk data is indeed useful for interrogating the neurobiology of this disorder as it can identify potentially affected processes and pathways. However, it does not give us a full picture into the dysregulation observed in the disorder and instead shows the overall trends in gene expression in the tissue as a whole. Alterations in gene expression in particular genes may not be detected if they are only expressed in cell with low population numbers. As such, it is important that cell-specific studies are carried out as they provide a much greater insight into transcriptomic landscape. While bulk datasets lack this level of detail, they do have much greater sample sizes which can give greater power to identify changes.

We attempted to leverage this advantage to identify cell specific changes in SCZ. We first estimated cell-type proportions for each individual in the bulk CMC dataset. This dataset had previously been used for this type of analysis as part of the PsychENCODE study (Wang et al., 2018). However we aimed to improve this analysis by using a more recent snRNA-seq data which had a greater sample size, although there is less subtype diversity. Our idea was that this would provide a more representative picture of the cell-type specific gene expression and thus give more accurate cell-type proportions. We do observe different proportions, particularly for the neuronal cell populations. We find that on average each sample has a combined neuronal cell proportion of approximately 0.2 while the PsychENCODE analysis has an average neuronal proportion of 0.6. We considered a number of reasons as to why this could have occurred. Firstly, cell-type deconvolution is known to be a very sensitive technique. Proportion estimates can vary greatly depending on the tool used (Sutton et al., 2022). dTangle (Hunt et al., 2019), the tool we used in this analysis, was chosen as it is reported to be among the most accurate tools for use with postmortem brain tissue data (Sutton et al., 2022) and has been used successfully in previous studies (Hauberg et al., 2020; Hoffman et al., 2022). Another possible reason for our results could be attributed to the snRNAseq data used. As proportion estimation is dependent on the reference data, it is possible the marker genes chosen based on this dataset may be affecting the estimation.

Despite our questions regarding our estimated proportions, we carried out our analysis plan to identify if they show similar trends to past results despite the discrepancies. We first assessed them for differences between cases and controls and observed a difference in proportion for excitatory neurons, astrocytes and an inhibitory neuronal subtype which was also observed by the previous PsychENCODE analysis (Wang et al., 2018). This

shows that despite the differences, we are observing a similar overall trend in relation to the proportions. Interestingly, the first snRNA-seq study which uses data from both SCZ cases and controls does not observe any differences between proportions between the groups (Ruzicka et al., 2022). This is in contrast to both this analysis and the PsychENCODE study. However, bulk studies take an overall tissue sample rather than profiling specific cell-types and may be biased towards more abundant cell-types thus affecting results. Further data generation and experimental work would be required to understand whether this is the case.

PRS has been used to great effect in the study of complex disease. In terms of SCZ, it is associated with a number of phenotypes including cognitive deficits and positive symptoms (Legge et al., 2021; Jonas et al., 2019). It is also associated with alterations in white and grey matter structure (Stauffer et al., 2021). As such, we investigated whether changes in cell-type proportion was associated with SCZ PRS. We only observed an association with PRS for microglial proportions but future analysis with improved proportion estimates or larger GWAS sample size could change this. However given the issues with proportion estimation this may be a spurious association.

We aimed to leverage the estimated proportions to impute cell-type specific gene expression. We constructed count matrices for each individual in the bulk expression data and used these profiles to conduct a DGE analysis. We observed differing numbers of DGEs per cell-type. A number of these in terms of gene-set size are comparable a snRNA-seq study of SCZ (Ruzicka et al., 2022) which also found 1000-2000 DEGs per cell-type. However, some of the imputed profiles including OPCs showed substantially larger numbers of DEGs. OPCs in particular are not a very abundant cell-type in the adult human brain and it is possible that their low proportion estimate which is used to generate the expression matrix is affecting these numbers.

We then performed a gene-set analysis to identify on a cell-specific level which functions and processes were affected. We found that despite our reservations regarding our estimated proportions, we observe a variety of implicated pathways and ontology terms that you would expect to see in these cell-types. Excitatory neurons which are believed to be highly involved in SCZ (Gandal et al., 2018; Skene et al., 2018; Ruzicka et al., 2022) implicated processes such as neuronal development and neuron projection. Indeed neuron projection directly links back to the dopamine hypothesis of SCZ as a lack of glutamatergic projection in the DLPFC is linked to altered release of dopamine in the midbrain leading to deficits in working memory (Tanaka, 2006). Inhibitory subtypes implicated distinct processes to excitatory neurons including NCAM1 interactions. NCAM1 is a synaptic adhesion molecule and knockouts show phenotypes related to SCZ, although not enough

to be a SCZ animal model (Albrecht and Stork, 2012; Sytnyk et al., 2017). A recent study has also shown that autoantibodies against NCAM1 isolated from individuals with SCZ can cause SCZ-like behaviour and synapse dysfunction in mice (Shiwaku et al., 2022). This analysis suggests that expression alterations in genes related to NCAM1 may be related to inhibitory neurons. Non-neuronal cell-types also displayed disease-associated changes in expression, some of which overlapped with neuronal cell-types such as signal transduction which is known to be aberrant in SCZ (McGuire et al., 2017).

Our targeted gene-set approach on the neuronal DEGs using SynGO implicated a wide variety of postsynaptic locations using the excitatory and IN-CGE and IN-MGE DEGs. Postsynaptic locations are consistent with enrichment testing from SCZ GWAS (Trubetskoy et al., 2022) which has suggested a postsynaptic pathology. Both excitatory and IN-CGE DEGs however also implicated presynaptic locations. Recent prioritisation of SCZ risk genes and subsequent gene-set analysis has also implicated presynaptic locations (Trubetskoy et al., 2022). These results suggest that this may vary between cell-types. Finally, we also investigated whether our gene-sets were enriched for genes containing common variants associated with SCZ. Five cell-types were enriched at nominal significance but only the astrocyte enrichment survived correction for multiple testing. This analysis implicated both neuronal and non-neuronal cell-types suggesting multiple cell-types may contribute to SCZ development but the aforementioned issues with our DEGs and proportions may be affecting these results.

Overall, this analysis aimed to leverage published bulk expression data to gain new insights into the neurobiology of SCZ. We achieved this to some extent by associating microglial proportions with PRS and implicating disease-associated functions at a cell-type specific level. However, our conclusions do have the caveat that our estimated proportions were not as expected for reasons that I have previously outlined. This could be improved in future work by estimating these proportions using other reference data to compare these results. Despite this issue, this analysis does show how we can use older datasets to provide cell-type specific insights which could be used in future to complement other functional genomics studies of SCZ.

Chapter 3

Analysis of Genes Under the
Control of *MEF2C* using
Functional Genomics Data

3.1 Abstract

Myocyte enhancer factor 2 (*MEF2C*) is a transcription factor thought to play a key role in neurodevelopment. It has been implicated in a number of neurodevelopmental disorders including schizophrenia (SCZ) and autism spectrum disorder (ASD). Murine knockout models of *MEF2C* have enabled the creation of gene-sets comprised of genes likely to be under the control of *MEF2C*. We investigated these gene-sets using data derived from single-nucleus sequencing studies of the human brain in order to identify cell-types that may be affected by these knockouts and included data from fetal brain in order to implicate particular timepoints given the role of *MEF2C* in neurodevelopment. We identified a number of neuronal cell-types at both fetal and adult stages that were enriched for genes under the control of *MEF2C*. We also identified that oligodendrocyte precursor cells were enriched for genes in our gene-set which was not observed by previous studies which only used cell-type specific data from mouse brain. We also investigated co-expression fold enrichment which was found to be highest in interneurons. This analysis was more limited than our cell-type enrichment analysis due to the number of cell-types in the dataset used and future work could explore co-expression patterns in a more diverse set of cell-types.

3.2 Introduction

Myocyte enhancer factor 2 C (*MEF2C*) is a transcription factor that is a member of the MEF2 family of transcription factors (Zhang and Zhao, 2022a). The MEF2 family is noted to have a key role in development and continues to mediate alterations in the epigenome throughout the lifespan of an individual (Flavell et al., 2008). It is comprised of *MEF2A*, *MEF2B*, *MEF2C* and *MEF2D*. *MEF2C* is located on chromosome 5 and has been implicated in brain development and a variety of neuropsychiatric and brain-related disorders such as schizophrenia, major depressive disorder and Alzheimer’s disease (AD) (Assali et al., 2019; Purcell et al., 2014; Hyde et al., 2016; Lambert et al., 2013). Of all MEF2 family members, it is first to be expressed during development which indicates a key role in the developing embryo and in the context of neuropsychiatric disorders and the developing brain (Assali et al., 2019). More specifically, it inactivates target genes in the absence of stimulation such as depolarisation or synaptic stimulation (Gu et al., 2018). This stimulation leads to dephosphorylation which releases MEF2 bound histone deacetylases resulting in activation of MEF2 target genes resulting in the promotion of activities such as axon growth, dendritic formation and remodelling, synaptic development and neuronal excitability (Ma and Telese, 2015; Li et al., 2008)

Recent genomic studies such as genome-wide association studies (GWAS) have further implicated a role for *MEF2C* in a variety of neurodevelopmental disorders. GWAS has identified an association for *MEF2C* with schizophrenia (Trubetskoy et al., 2022), IQ (Savage et al., 2018) and educational attainment (EA) (Okbay et al., 2022). Furthermore, functional genomic studies such as single-cell sequencing has also identified transcriptional perturbations of *MEF2C* in single-cell expression data generated from individuals with schizophrenia (Ruzicka et al., 2022). As such exploring the genes regulated by *MEF2C* and the cell types in which it does so is crucial to further our understanding of its possible mechanisms and roles in these disorders.

Due to the role of *MEF2C* in brain development and related activities, numerous studies have sought to further understand its biological role in cognition and the impact of its deletion or mutation on brain function. Knockout studies can complement genomic studies by identifying genes affected by *MEF2C* dysregulation in animal models. These studies have primarily employed the use of knockout mice to investigate structural abnormalities and gene expression changes. Homozygous knockouts of *MEF2C* are embryonic lethal as a result of the role of *MEF2C* in cardiac development, which results in cardiac defects before brain development at embryonic day 9.5 (Lin et al., 1997). As a result, mouse models of *MEF2C* have focused on heterozygous or conditional knockouts. Heterozygous knockouts of *MEF2C* display reduced viability at 44% in comparison to wild-type mice and

display abnormal neuronal and synaptic functionality including an increase in excitatory synaptic transmission in the hippocampus (Tu et al., 2017). They also display autistic-like behavioural phenotypes with repetitive head motions and decreased spatial learning showing that there appears to be a range of cognitive deficits associated with *MEF2C* heterozygous knockout. Conditional knockouts of *MEF2C* have also been explored with a range of studies conditionally knocking out *MEF2C* at either different developmental timepoints or in particular tissues. Conditional knockout in neural progenitor cells results in abnormal compaction of neurons, reduced overall brain size in addition to anxious behaviours and decreased cognitive function (Li et al., 2008). Central nervous system (CNS) deletion of *MEF2C* resulted in a deficit in hippocampal dependent learning due to the suppression of excitatory synapses (Barbosa et al., 2008). Postnatal conditional deletion of *MEF2C* significantly increased dendritic spine numbers resulting in deficits in motor coordination (Adachi et al., 2016). However the cognitive and learning deficits associated with similar deletions in embryonic models were not observed which indicates that while *MEF2C* has a continuing role in postnatal brain activity this is distinct to its role in embryonic brain development.

(Harrington et al., 2016) sought to characterise the role of *MEF2C* in early neuronal development by creating a conditional knockout (cKO) of differentiated excitatory neurons as *MEF2C* is highly expressed in early brain development particularly in differentiated neurons (Lyons et al., 1995; Leifer et al., 1997). This resulted in a decrease in excitatory and an increase in inhibitory synaptic transmission in addition to behavioural changes such as altered social behaviours, learning deficits and hyper activity which is consistent with previous studies which described autistic-like behaviours following *MEF2C* knockouts (Li et al., 2008; Barbosa et al., 2008; Tu et al., 2017). Differential gene expression analysis between knockout and wild-type mice also showed that differentially expressed genes (DEGs) were enriched for neuron expressed genes. DEGs that showed up-regulation in knockout mice were enriched for genes involved in differentiation of neuronal cells suggesting that *MEF2C* may act as a repressor for these genes and functions.

A subsequent study by our group (Cosgrove et al., 2021) further investigated these genes to assess whether they were associated with phenotypes such as schizophrenia, IQ, and EA given that *MEF2C* knockouts display deficits in cognitive ability and GWAS of these phenotypes report single SNP associations at the *MEF2C* locus. They converted the DEGs identified in mice from (Harrington et al., 2016) to their human orthologs and tested these for enrichment in the aforementioned phenotypes using the recent GWAS summary statistics. They identified significant associations for all three phenotypes but not for autism spectrum disorder (ASD) which is to be expected given the low contribution of common variants to this disorder. Interestingly, when the DEGs were separated in groups

of down and upregulated genes, they showed differing enrichments with down-regulated genes showing stronger association with EA, schizophrenia and IQ. Following this, gene ontology (GO) analysis was then used to ascertain the functionality of these DEG. The down-regulated DEGs were involved in synaptic transmission while up-regulated DEGs were involved in neuronal differentiation. Another focus of this study was the identification of cell types that were enriched for genes affected by the *MEF2C* cKO. This analysis was performed by using single-cell RNAseq (scRNAseq) data from the mouse brain rather than human data due to data availability. Cell-types enriched for these DEGs primarily included excitatory neuron subtypes in addition to medium spiny neurons and inhibitory neurons to a lesser extent. This is consistent with cell types previously implicated in genomic studies using single-cell expression data (Ruzicka et al., 2022; Skene et al., 2018; Savage et al., 2018). Strikingly, little enrichment was observed for non-neuronal cell-types despite association of astrocyte expression with genetic variation in schizophrenia (Ruzicka et al., 2022). This suggests reanalysis with human-specific data may be useful to identify further enriched cell-types.

A further knockout model was generated by (Harrington et al., 2020) in order to mimic *MEF2C* microdeletion syndrome, a recently discovered neurodevelopmental disorder (Le Meur et al., 2010). It is characterised by microdeletions on chromosome 5 q14.3 a region that includes *MEF2C* and are thought to affect the DNA binding ability of *MEF2C* (Harrington et al., 2020). This heterozygous knockout is thought to better characterise the *MEF2C* mutational profile observed in humans in comparison to the previous model described above (Harrington et al., 2016, 2020). It was successful in reproducing the behavioural phenotype observed in this syndrome such as altered social interaction and hyperactivity and also resulted in a large number of genes showing differential expression between knockout and wild-type. These DEGs were associated with excitatory neurons and microglia and ASD which is consistent with previous knockout studies (Harrington et al., 2016). Cell-type specific differences in the expression pattern of these DEGs was also observed with DEGs linked to excitatory neurons being predominantly downregulated while DEGs linked to microglia display were more often upregulated. Furthermore, it was observed that mutation in excitatory neurons was sufficient to induce all phenotypic changes except social deficits while mutation in microglia reproduced all effects except anxiety-like behaviours. This again highlights the need to explore changes at the cell-type specific level as many cell-types contribute to the observed phenotypic changes.

This gene-set from (Harrington et al., 2020) was also assessed for association with other human phenotypes by our group (Fahey et al., 2023). Similar to (Cosgrove et al., 2021), all DEGs were converted to human orthologs. In contrast to the previous study, the full set of DEGs was not found to be enriched for common variants associated with schizophre-

nia, EA or IQ. However, there was an association with IQ and EA when the gene-set was restricted to upregulated genes, which included genes involved in synaptic function. Furthermore, genes that were believed to have elevated expression in cortical neurons were identified using supporting evidence from single-cell data. These genes were found to be enriched for common variants associated with EA and IQ. As this enrichment was not observed in the overall gene-set, it again highlights the importance of considering cell-type differences when characterising the biological changes as a result of *MEF2C* knockout.

Overall, my analysis here aims to explore the effects and cell type specificity of genes found to be dysregulated as a result of *MEF2C* perturbation, which were first identified by conditional knockouts of *MEF2C* (Harrington et al., 2016, 2020). Using snRNAseq data from the developing brain, we aim to identify whether these genes are enriched in certain human cell types at particular time points. We also aim to identify which adult human cell types these gene-sets are enriched in using open-access single-cell RNAseq data from schizophrenia cases and controls. Finally, these gene-sets will be used to assess co-expression enrichment patterns on a cell-type basis using scRNAseq data.

3.3 Methods

3.3.1 Datasets

3.3.1.1 Gene-sets

An initial set of DEGs were identified by (Harrington et al., 2016). This study utilised *MEF2C* conditional knockout mice to explore the effects of insufficient MEF2C protein levels on neurodevelopment. This resulted in 1,076 DEGs with a log₂ fold change > 0.3 and FDR < 0.05. 478 of these genes were found to be up-regulated (increased expression level) and 598 genes were found to be down-regulated (decreased expression level) in cKO mice. A subsequent study (Cosgrove et al., 2021) further refined this list to 1,055 genes by converting to human orthologues using Ensembl Biomart (<https://www.ensembl.org/biomart>). This gene-set was used for analysis as is denoted as the homozygous knockout (KO_hom) (n=1055). It was also separated into up and down-regulated gene-sets. The up-regulated gene-set (n=465) will be known as “KO_hom_up” and the down-regulated gene-set (n=590) is known as “KO_hom_down”.

A second set of DEGs was identified by (Harrington et al., 2020). This gene-set was obtained from *MEF2C* knockout mice that were generated to mimic MEF2C haploinsufficiency syndrome and contained 476 genes at FDR < 0.05. Once converted to human orthologues using Ensembl Biomart (<https://www.ensembl.org/biomart>) this gene-set contained 460 genes that were carried forward for analysis and are denoted as the heterozygous knockout (KO_het) in this chapter (Fahey et al., 2023).

3.3.1.2 Single-cell Data

In order to investigate which cell types were enriched for *MEF2C* associated genes, we integrated the *MEF2C* gene-sets with single-nucleus gene expression data generated using single-nucleus RNAseq (snRNA-seq). As noted throughout this thesis, any reference to single-cell data refers to single-nucleus data unless otherwise stated. All data was isolated from postmortem cortical samples and was obtained from the following studies. The first two datasets were used in the cell-type enrichment analysis while the third dataset was used in the the co-expression enrichment analysis.

(Zhu et al., 2022) generated snRNAseq data from 12 human neocortex samples from six developmental periods (early mid gestation fetal, late mid gestation fetal, infancy, childhood, adolescence and adulthood) resulting in 45,549 cells. Twenty-eight distinct clusters were identified comprising 15 cell types including inhibitory neurons, excitatory neurons, vascular smooth muscle cell (VSMC), pericytes, endothelial cells, OPCs, oligodendrocytes and microglia. This will be referred to as the developmental dataset and is the first dataset

used for the cell-type enrichment analysis.

(Ruzicka et al., 2022) generated snRNAseq data from schizophrenia cases and controls from two cohorts. In this analysis we use data from the second cohort which is referred to as the Mount Sinai School of Medicine (MSSM) cohort. This cohort consisted of 92 samples (n=41 schizophrenia cases, n=51 controls) from the prefrontal cortex (PFC) which were subsequently profiled using snRNAseq. These individuals were exclusively adult samples and ranged in age from 24 to 101. This resulted in with a total of 106,761 cells which was found to contain 24 individual cell types following clustering. These cell types included inhibitory neurons, excitatory neurons, interneurons, oligodendrocytes, oligodendrocyte precursor cells (OPCs), microglia and endothelial cells. This will be referred to as the SCZ dataset and is the second dataset used for the cell-type enrichment analysis.

(Zhong et al., 2018) generated snRNAseq data from the developing human PFC. Samples were taken from at different gestational timepoints ranging from gestational weeks 6-26. Overall, 2,300 cells were profiled across the six main developing brain cells types: neural progenitor cells (NPCs), excitatory neurons, interneurons, astrocytes and microglia. This dataset will be referred to as the fetal dataset and is used for the co-expression analysis only.

3.3.2 Enrichment of *MEF2C* gene-sets in single-cell data

Expression weighted cell enrichment (EWCE) was used to identify cell types in which the *MEF2C* gene-sets were enriched. EWCE is implemented as an R package (<https://github.com/NathanSkene/EWCE>) and is a statistical method that determines whether a gene-set has higher expression in a given cell-type than would be expected by chance (Skene and Grant, 2016). This is estimated by testing to identify whether the expression of a given gene-set in a cell-type is higher than random gene-sets created using background genes.

Previous studies (Cosgrove et al., 2021), have investigated which cell types are enriched for the genes present in the *MEF2C* gene-sets described above using mouse scRNAseq data and the R package EWCE (Skene and Grant, 2016). Using more recent human scRNAseq (Ruzicka et al., 2022; Zhu et al., 2022), we constructed two CellTypeDatasets (CTDs) using EWCE. This was carried out by first converting both snRNAseq datasets to the SingleCellExperiment format which is a format which stores expression data and associated metadata for single-cell data sets and is available as an R package (Amezquita et al., 2020). Genes that are not sufficiently expressed or are uninformative were then removed from the datasets using the `drop_uninformative_genes()` in order to generate robust

results and reduce noise. Both single-cell datasets were then converted to separate CTDs using the `generate_celltype_data()` function within the EWCE package. This resulted in two CTDs which contained a normalised sparse expression matrix for each dataset.

Both CTDs were then used to identify the cell-types in which the *MEF2C* associated genes were enriched. The following process was carried out for each CTD. Taking the developmental CTD from (Zhu et al., 2023) as an example, the CTD was loaded using the `load_rdata()` function within EWCE. The KO_hom gene-set from (Cosgrove et al., 2021) (*MEF2C21*) was split into up-regulated (KO_hom_up) and down-regulated (KO_hom_down) as previously mentioned. These three gene-sets and the gene-set from (Fahey et al., 2023) (KO_het) were then tested individually to identify enriched cell-types from the CTD. This was carried out using the `bootstrap_enrichment_test()` function controls for transcript length and GC content by selecting bootstrap lists of genes with similar features to the genes in the target list. It was run for 10,000 permutations in order to generate robust FDR corrected p-values. This resulted in four sets of results table, one for each gene-set (Tables 1-4). This was then carried out using the other CTD generated using the data from (Ruzicka et al., 2022) and results are available in Tables 5-8.

3.3.3 Enrichment of *MEF2C* Gene-sets in DEGs from Schizophrenia Single-nuclei Analysis

We obtained the list of DEGs from each cell-type in the SCZ dataset. Each *MEF2C* gene-set was tested for enrichment in each of the DEG lists using a hypergeometric test. The p-value significance threshold was obtained by dividing 0.05 by 100 (25 cell-types multiplied by 4 gene-sets).

3.3.4 Co-expression Network Construction

Similar to a previous study (Pang et al., 2020), we constructed co-expression networks for each of the cell types contained in the fetal dataset from (Zhong et al., 2018). The data was downloaded from the Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104276>). Counts were normalized to transcripts per million (TPM) and then log-transformed ($\log_2(\text{TPM}+1)$). Cells were then classified into the six major cell types (neural progenitor cells (NPCs), excitatory neurons, interneurons, astrocytes and microglia) using the sample annotation file from (Zhong et al., 2018). Genes which did not have an expression level greater than 0 in at least 10% of cells of a particular cell-type were removed.

Genes expressed in each cell type were used as background genes and the Spearman corre-

lation coefficient between background genes in each cell type was calculated to generate a correlation matrix. Following the procedure used by (Pang et al., 2020) we determined the correlation threshold for the top 0.5% of the Spearman’s correlation coefficients that were calculated and 0.5% was defined as the co-expression network density for the background genes. Networks were then constructed for a given gene-set with the same correlation threshold.

3.3.5 Co-expression Enrichment Analysis

The correlation threshold value of 0.5% used above to construct a network for the background genes in a given cell type is defined as the co-expression network density for the background genes. The co-expression network density for each gene-set is the number of significant co-expressed pairs divided by the number of all pairs between genes in the gene-set.

Co-expression fold enrichment was then calculated per gene-set and is defined as the ratio of the network density for the background genes to the density of the gene-set. This was also calculated at different timepoints for excitatory neurons, inhibitory neurons and NPCs.

3.4 Results

3.4.1 Cell-type Enrichment in the Developmental dataset

We performed cell-type enrichment analysis on single-cell gene expression data from the human brain. We aimed to identify cell-types in which DEGs from *MEF2C* cKOs were enriched. The developmental dataset from (Zhu et al., 2023) profiled 15 cell-types across fetal and post-natal stages. Each gene-set, described earlier, was tested to identify enriched cell-types. The differing enrichment per gene-set are visualised in a heatmap using the log fold-change statistic from EWCE (Figure 3.1). The KO_hom gene-set had seven enriched cell-types following BH correction (Table 3.1). Enriched cell-types were primarily neuronal and included multiple time-points including the early and late fetal stages. Inhibitory neurons derived from both the medial and caudal ganglionic eminence were also significantly enriched (FDR <0.05), while OPCs were also enriched following multiple testing correction and were the only non-neuronal cell type enriched in this gene-set.

The KO_hom_down gene-set were enriched for four neuronal cell-types present in the developmental snRNA-seq dataset (Table 3.2). Adult and late-fetal excitatory neuronal cell types were implicated in this analysis (Figure 3.1). While only adult inhibitory neurons were found to be enriched in this gene-set. In contrast, KO_hom_up gene-set were enriched for five neuronal cell-types from all three time-points (Table 3.3). Early-fetal excitatory neurons were enriched in the up-regulated genes which was not seen when restricted to down-regulated genes only. Furthermore, only adult inhibitory neurons were enriched which echoes the result from the KO_hom gene-set.

Finally, the KO_het gene-set was only significantly enriched for excitatory neurons at the late-fetal stage following BH correction. This was in contrast to the other gene-set and its subsets which contained multiple enriched cell-types (Table 3.4). Enrichment of adult excitatory neurons was nominally significant but it did not survive multiple testing correction.

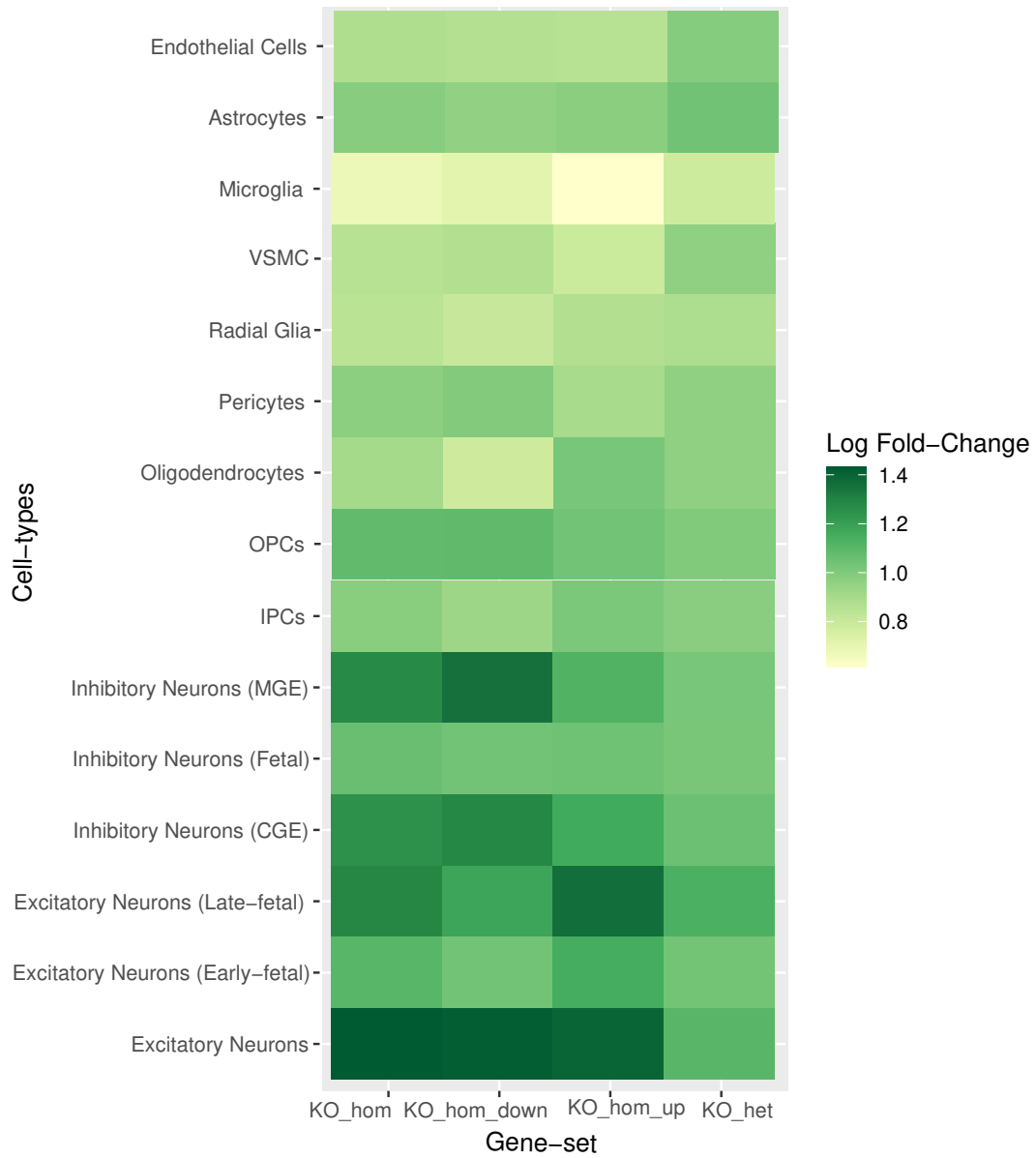


Figure 3.1: Heatmap displaying the differing cell-type enrichments in the developmental dataset for each tested gene-set using the log-fold change statistic calculated by the EWCE package.

Table 3.1: EWCE results using the KO_hom gene-set and the 15 cell-types from the developmental dataset. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Excitatory Neurons	<0.00001	<0.00001	1.4309
Excitatory Neurons (Late-fetal)	<0.00001	<0.00001	1.2832
Inhibitory Neurons (MGE)	<0.00001	<0.00001	1.2745
Inhibitory Neurons (CGE)	<0.00001	<0.00001	1.2495
Excitatory Neurons (Early-fetal)	0.0002	0.0006	1.1053
Oligodendrocyte Precursor Cells (OPCs)	0.0078	0.0195	1.0807
Inhibitory Neurons (Fetal)	0.0102	0.0219	1.0604
Astrocytes	0.6967	1	0.9819
Intermediate progenitor cells (IPCs)	0.7392	1	0.9803
Pericytes	0.7608	1	0.9697
Oligodendrocytes	0.9876	1	0.9066
Endothelial Cells	0.9975	1	0.8757
Vascular Smooth Muscle Cells (VSMC)	1	1	0.8508
Radial Glia (RG)	1	1	0.8468
Microglia	1	1	0.6804

Table 3.2: EWCE results using the KO_hom_down gene-set and the 15 cell-types from the developmental dataset. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Excitatory Neurons	<0.00001	<0.00001	1.4174
Inhibitory Neurons (MGE)	<0.00001	<0.00001	1.3574
Inhibitory Neurons (CGE)	<0.00001	<0.00001	1.2816
Excitatory Neurons (Late-fetal)	<0.00001	<0.00001	1.1789
Oligodendrocyte Precursor Cells (OPCs)	0.0317	0.0951	1.0826
Inhibitory Neurons (Fetal)	0.1278	0.3036	1.0383
Excitatory Neurons (Early-fetal)	0.1417	0.3036	1.0379
Pericytes	0.5160	0.9675	0.9965
Astrocytes	0.8339	1	0.9571
Intermediate progenitor cells (IPCs)	0.9713	1	0.9269
Endothelial Cells	0.9865	1	0.8655
Vascular Smooth Muscle Cells (VSMC)	0.9948	1	0.8684
Oligodendrocytes	1	1	0.7849
Radial Glia (RG)	1	1	0.8038
Microglia	1	1	0.7137

Table 3.3: EWCE results using the KO_hom_up gene-set and the 15 cell-types from the developmental dataset. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Excitatory Neurons (Late-fetal)	<0.00001	<0.00001	1.3662
Excitatory Neurons	<0.00001	<0.00001	1.3958
Excitatory Neurons (Early-fetal)	0.0003	0.0015	1.1485
Inhibitory Neurons (CGE)	0.0005	0.0019	1.1599
Inhibitory Neurons (MGE)	0.0018	0.0054	1.122
Inhibitory Neurons (Fetal)	0.1082	0.2705	1.0469
Oligodendrocyte Precursor Cells (OPCs)	0.2195	0.4704	1.0367
Oligodendrocytes	0.3495	0.6553	1.0234
Intermediate progenitor cells (IPCs)	0.3968	0.6613	1.0105
Astrocytes	0.6845	1	0.9736
Pericytes	0.9496	1	0.899
Endothelial Cells	0.9829	1	0.8537
Radial Glia (RG)	0.9982	1	0.8696
Vascular Smooth Muscle Cells (VSMC)	0.9998	1	0.7957
Microglia	1	1	0.6131

Table 3.4: EWCE results using the KO_het gene-set and the 15 cell-types from the developmental dataset. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Excitatory Neurons (Late-fetal)	0.0009	0.0135	1.13328
Excitatory Neurons	0.0191	0.1432	1.1022
Inhibitory Neurons (CGE)	0.1198	0.5365	1.0527
Excitatory Neurons (Early-fetal)	0.1895	0.5365	1.0342
Astrocytes	0.2049	0.5365	1.0417
Microglia	0.2146	0.5365	1.0575
Inhibitory Neurons (MGE)	0.2851	0.5373	1.0226
Inhibitory Neurons (Fetal)	0.2866	0.5374	1.0202
Oligodendrocyte Precursor Cells (OPCs)	0.4923	0.8173	0.9989
Intermediate progenitor cells (IPCs)	0.6788	0.8173	0.9782
Oligodendrocytes	0.6906	0.8173	0.9654
Vascular Smooth Muscle Cells (VSMC)	0.6999	0.8173	0.9652
Pericytes	0.7083	0.8173	0.96467
Radial Glia (RG)	0.9927	0.9994	0.8851
Endothelial Cells	0.9994	0.9994	0.7868

3.4.2 Cell-type Enrichment in the SCZ dataset

The SCZ dataset from (Ruzicka et al., 2022) was also used to test for cell-type enrichment and differed from the previous single-cell dataset in that it contained cells from adult schizophrenia cases and controls. Again, each gene-set was tested for enriched cell-types (Figure 3.2). The KO_hom gene-set was found to be enriched in 16 cell-types (Table 3.5). The enriched cell-types were consistent with the results from the other single-cell dataset in that they primarily consisted of excitatory and inhibitory neurons. Other cell types such as *SNCG* expressing cells. *PAX6* expressing cells and near-projecting neurons were also significant following multiple testing correction. This data also contained cell-types from multiple brain layers with layer two/three, four, five and six inhibitory neurons found to be significantly enriched in this gene-set. A more diverse and subtype specific enrichment was observed due to the increased cell-type diversity in the SCZ dataset.



Figure 3.2: Heatmap displaying the differing cell-type enrichments in the SCZ dataset for each tested gene-set using the log-fold change statistic calculated by the EWCE package.

The KO_hom_down gene-set were enriched for 14 cell-types (Table 3.6). They were consistent with the enriched cell-types identified using the KO_hom gene-set with the majority of the observed enrichment coming from inhibitory and excitatory neurons (Figure 3.2). Other cell types such as *SNCG* expressing cells, *PAX6* expressing cells and near-projecting neurons were again also significantly enriched. Cells from multiple brain layers

were also implicated with layer two/three, four, five and six inhibitory neurons found to be significantly enriched in this gene-set. In contrast, the KO_hom_up gene-set was only enriched for four cell-types. All cell-types were from Layer 6 and were predominantly inhibitory neurons but also included a mix of excitatory and inhibitory neurons in Layer 6b and excitatory projection neurons. (Table 3.7).

The KO_het gene-set was not found to contain any significantly enriched cell-types (Table 3.8) in the SCZ dataset. Four cell-types were found to be nominally significant but did not survive multiple testing correction. These included two inhibitory Lamp5 neuronal subtypes, Layer 6b neurons and excitatory projection neurons from Layer 6.

The SCZ dataset was also split into cases and controls to identify whether the enrichment results would differ. However, they were relatively consistent with the results of the combined dataset described above (data not shown).

Table 3.5: EWCE results using the KO_hom gene-set and the 24 cell-types from the SCZ dataset. EN denotes excitatory neurons and IN denotes inhibitory neurons. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Layer 6 Inhibitory Neurons	<0.00001	<0.00001	1.2414
Chandelier Neurons (IN)	<0.00001	<0.00001	1.2418
Layer 6 Cortico-cortical Projection Neurons (EN)	<0.00001	<0.00001	1.2035
Lamp5 Neurons (IN)	<0.00001	<0.00001	1.1944
Lamp5 Neurons (<i>Lhx6</i> expressing) (IN)	<0.00001	<0.00001	1.2025
Layer 6 Inhibitory Neurons (<i>Car3</i> expressing)	<0.00001	<0.00001	1.2198
Layer 5 Inhibitory Neurons	<0.00001	<0.00001	1.1813
Layer 4 Inhibitory Neurons	<0.00001	<0.00001	1.1681
Layer 5 Excitatory Neurons	<0.00001	<0.00001	1.1836
Layer 2/3 Inhibitory Neurons	0.0001	0.0002	1.1821
Neuropeptide Somatostatin (Sst) Neurons (IN)	0.0001	0.0002	1.1402
<i>SNCG</i> Expressing Cells	0.0002	0.0004	1.1107
Layer 6b Neurons	0.0003	0.0005	1.1664
Parvalbumin Neurons (IN)	0.0004	0.0007	1.1489
Layer 5/6 Near-projecting Neurons	0.0009	0.0014	1.1375
<i>Pax6</i> Expressing Cells	0.0012	0.0018	1.1122
Sst-Chodl Neurons (IN)	0.0507	0.0716	1.0777
<i>VIP</i> Expressing Neurons (IN)	0.0826	0.1101	1.049
Oligodendrocyte Precursor cells (OPCs)	0.8452	1	0.9504
Astrocytes	0.9860	1	0.8907
Endothelial Cells	0.9949	1	0.8479
Vascular and Leptomeningeal Cell (VLMC)	1	1	0.8155
Oligodendrocyte	1	1	0.7741
Microglia	1	1	0.5114

Table 3.6: EWCE results using the KO_hom_down gene-set and the 24 cell-types from the SCZ dataset. EN denotes excitatory neurons and IN denotes inhibitory neurons. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Chandelier Neurons (IN)	<0.00001	<0.00001	1.3658
Lamp5 Neurons (IN)	<0.00001	<0.00001	1.2402
Lamp5 Neurons (<i>Lhx6</i> expressing) (IN)	<0.00001	<0.00001	1.251
Parvalbumin Neurons (IN)	<0.00001	<0.00001	1.2598
Layer 4 Inhibitory Neurons	<0.00001	<0.00001	1.205
Neuropeptide Somatostatin (Sst) Neurons (IN)	0.0005	0.002	1.1876
Layer 5 Excitatory Neurons	0.001	0.0034	1.1869
Layer 5 Inhibitory Neurons	0.0012	0.0035	1.1721
Layer 6 Inhibitory Neurons	0.0013	0.0035	1.1556
Layer 2/3 Inhibitory Neurons	0.0016	0.0038	1.2098
<i>SNCG</i> Expressing Cells	0.0123	0.0266	1.0984
Layer 6 Cortico-cortical Projection Neurons (EN)	0.014	0.0266	1.1004
Layer 5/6 Near-projecting Neurons	0.0144	0.0266	1.1249
Layer 6 Inhibitory Neurons (<i>Car3</i> expressing)	0.0262	0.0449	1.1002
<i>VIP</i> Expressing Neurons (IN)	0.0427	0.0683	1.0838
Layer 6b Neurons	0.0679	0.1018	1.0829
<i>Pax6</i> Expressing Cells	0.0972	0.1372	1.0609
Sst-Chodl Neurons (IN)	0.1784	0.2379	1.0543
Oligodendrocyte Precursor Cells (OPCs)	0.8597	1	0.9322
Endothelial Cells	0.9906	1	0.8232
Vascular and Leptomeningeal Cells (VLMC)	0.9942	1	0.8252
Astrocytes	0.9992	1	0.8052
Oligodendrocytes	1	1	0.5931
Microglia	1	1	0.5131

Table 3.7: EWCE results using the KO_hom_up gene-set and the 24 cell-types from the SCZ dataset. EN denotes excitatory neurons and IN denotes inhibitory neurons. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Layer 6 Inhibitory Neurons	<0.00001	<0.00001	1.3013
Layer 6 Cortico-cortical Projection Neurons (EN)	<0.00001	<0.00001	1.2881
Layer 6 Inhibitory Neurons (<i>Car3</i> expressing)	<0.00001	<0.00001	1.3221
Layer 6b Neurons	0.0012	0.0072	1.2278
Layer 5 Inhibitory Neurons	0.0087	0.0418	1.1463
<i>Pax6</i> Expressing Cells	0.0123	0.0492	1.1354
Layer 5 Excitatory Neurons	0.0197	0.0675	1.1357
Lamp5 Neurons (<i>Lhx6</i> expressing) (IN)	0.0359	0.096	1.0954
Lamp5 Neurons	0.0360	0.096	1.0932
<i>SNCG</i> Expressing Cells	0.0461	0.1076	1.0863
Layer 5/6 Near-projecting Neurons	0.0493	0.1076	1.111
Layer 2/3 Inhibitory Neurons	0.0803	0.1606	1.1036
Layer 4 Inhibitory Neurons	0.0915	0.1689	1.0775
Sst-Chodl Neurons (IN)	0.1616	0.277	1.0705
Chandelier Neurons (IN)	0.2321	0.3673	1.0369
Neuropeptide Somatostatin (Sst) Neurons (IN)	0.2449	0.3673	1.0372
Oligodendrocytes	0.6176	0.8705	0.9677
Astrocytes	0.6529	0.8705	0.9643
Parvalbumin Neurons (IN)	0.7229	0.8756	0.9631
<i>VIP</i> Expressing Neurons (IN)	0.7297	0.8756	0.9656
Oligodendrocyte Precursor Cells (OPCs)	0.7960	0.9097	0.9355
Endothelial Cells	0.9554	1	0.8492
Vascular and Leptomeningeal Cells (VLMC)	0.9989	1	0.7764
Microglia	1	1	0.4914

Table 3.8: EWCE results using the KO_het gene-set and the 24 cell-types from the SCZ dataset. EN denotes excitatory neurons and IN denotes inhibitory neurons. P-values were corrected for multiple testing using the Benjamini-Hochberg method to generate Q-values.

Cell-type	P-Value	Q-Value	Log Fold Change
Lamp5 Neurons (<i>Lhx6</i> expressing) (IN)	0.0144	0.2004	1.1212
Lamp5 Neurons (IN)	0.0167	0.2	1.1149
Layer 6b Neurons	0.0355	0.258	1.1206
Layer 6 Cortico-cortical Projection Neurons (EN)	0.0430	0.258	1.0964
Chandelier Neurons (IN)	0.0605	0.2904	1.0933
Layer 6 Inhibitory Neurons (<i>Car3</i> expressing)	0.0849	0.3396	1.08
<i>Pax6</i> Expressing Cells	0.1237	0.4241	1.0615
Layer 6 Inhibitory Neurons	0.1585	0.4265	1.0534
Astrocytes	0.1744	0.4265	1.0751
<i>SNCG</i> Expressing Cells	0.1840	0.4265	1.0425
Microglia	0.1955	0.4265	1.0813
Layer 4 Inhibitory Neurons	0.2784	0.5568	1.0309
Layer 5 Inhibitory Neurons	0.3345	0.6175	1.0197
Layer 2/3 Inhibitory Neurons	0.3898	0.663	1.0136
Sst-Chodl Neurons (IN)	0.4144	0.663	1.0132
Layer 5 Excitatory Neurons	0.504	0.7382	0.9957
Parvalbumin Neurons (IN)	0.5229	0.7382	0.9906
Layer 5/6 Near-projecting Neurons	0.5637	0.7516	0.9848
Neuropeptide Somatostatin (Sst) Neurons (IN)	0.6142	0.7594	0.9779
<i>VIP</i> Expressing Neurons (IN)	0.6328	0.7594	0.9768
Oligodendrocyte Precursor Cells (OPCs)	0.7008	0.8009	0.9556
Oligodendrocytes	0.9187	0.9598	0.8746
Vascular and Leptomeningeal Cells (VLMC)	0.9198	0.9598	0.8847
Endothelial Cells	0.9745	0.9745	0.8249

3.4.3 Enrichment of *MEF2C* Gene-sets in DEGs from Schizophrenia Single-nuclei Analysis

We tested the *MEF2C* gene-sets for enrichment in cell-type specific DEGs from the SCZ dataset using a hypergeometric test. Similar to the cell-type enrichment analysis, a variety of neuronal cell-types were implicated (Figure 3.3). However this analysis also implicated endothelial cells, OPCs and microglia which were not detected by the other analysis. In particular, the KO_het gene-set was enriched for DEGs in microglia, astrocytes, endothelial cells in addition to neuronal cell-types despite not appearing to be enriched for any particular cell-type in previous analysis.

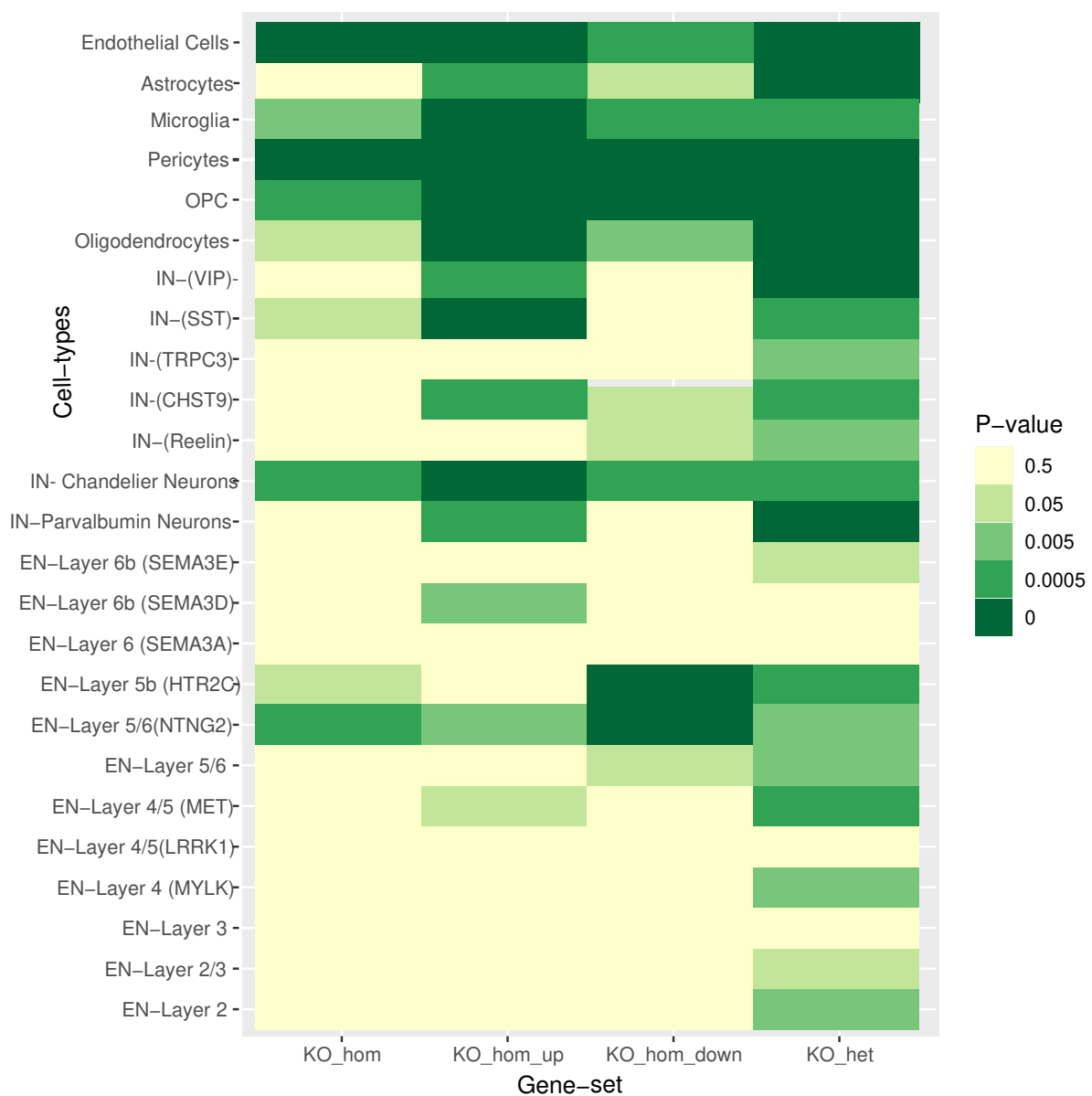


Figure 3.3: Heatmap displaying the differing cell-type enrichments in the DEGs for each tested gene-set using the p-value calculated by a hypergeometric test.

3.4.4 Co-expression fold enrichment

Co-expression fold enrichment was calculated for each of the four gene-sets in each cell-type contained in the fetal dataset (Figure 3.4). Higher fold enrichment scores imply that the genes in the set are more significantly co-expressed than genes in the background set. The highest fold enrichments were observed in interneurons followed by NPCs. Fold enrichment was generally lower in non-neuronal cell-types across gene-sets. The KO_hom_down gene-set showed differing fold enrichment to the KO_hom_up gene-set particularly for interneurons and microglia. The KO_het showed much lower fold enrichment across cell-types.

Co-expression fold enrichment was also calculated for excitatory neurons, inhibitory neurons and NPCs at different developmental time points (Figure 3.5). In NPCs, fold enrichment was highest at gestational week (GW) 10 and was higher for genes in the KO_hom gene-set. Fold enrichment did not differ extensively between GWs in excitatory neurons. Finally, co-expression fold enrichment was highest at GW 23 in interneurons.

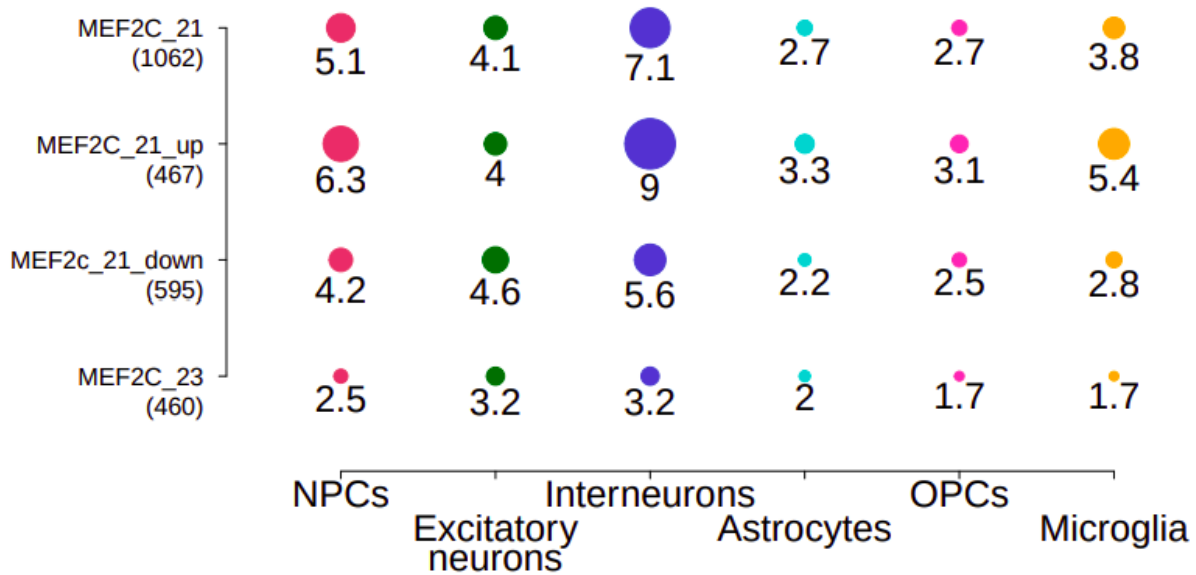
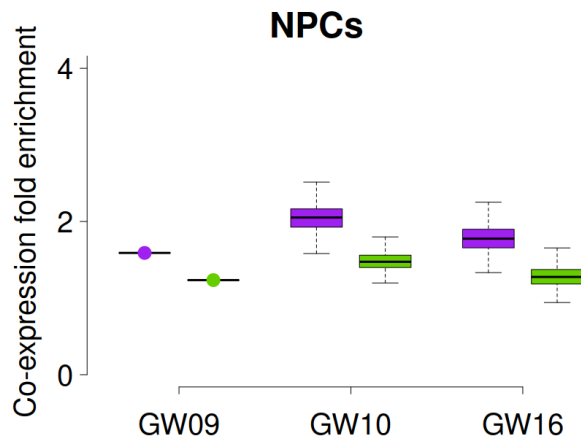
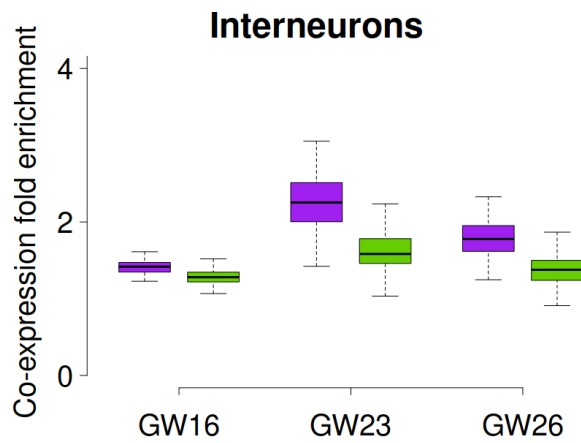


Figure 3.4: Co-expression fold enrichment of *MEF2C* gene-sets in six cell types from the prefrontal cortex. The name and sample size of each gene-set is included on the left side of the graph. The co-expression fold enrichment score for each gene-set per cell-type is denoted by a circle which is proportional to enrichment score

[a]



[b]



[c]

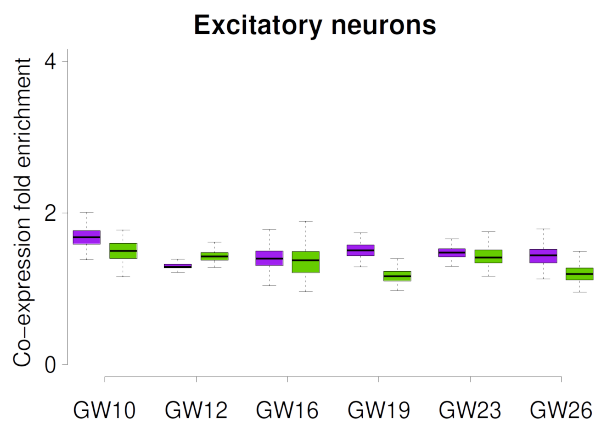


Figure 3.5: Coexpression fold enrichment of *MEF2C* gene-sets in three cell types from the prefrontal cortex at different gestational weeks (GW). Purple boxplots denote results for genes in the KO_hom gene-set while green denote results for genes in the KO_het gene-set.

3.5 Discussion

The analysis of gene-sets derived from gene-expression studies provides an insight into how these genes can affect a given phenotype. In the case of *MEF2C*, previous work has implicated this gene in the dysfunction of synaptic processes and neuronal differentiation (Cosgrove et al., 2021). This suggests that neuronal cell-types will be greatly impacted by *MEF2C* mutation. However, while gene ontology and over-representation analysis can provide clues as to which cell-types are at risk of being impacted, they do not utilise gene expression from those cell-types. In addition, the vast majority of gene-sets are derived from bulk-tissue samples which again limits our ability to understand which cell-types they affect. The generation of single-cell dataset which profile gene expression has been a major development in the area of functional genomics. These data have enabled us to identify the cell-types that are likely to be impacted by cKO of *MEF2C* or a *MEF2C* knockout designed to recapitulate the effects of *MEF2C* haploinsufficiency syndrome. The genes present in the KO_hom gene-set were found to be enriched in excitatory neuronal cell-types in both of the snRNAseq dataset used. While this is consistent with enrichment analysis from (Cosgrove et al., 2021), it provides confirmation that *MEF2C* dysregulation impacts human cell types as previous studies relied solely is on expression data from the mouse brain. The developmental dataset also enabled us to discern that while excitatory neurons across timepoints were enriched for this gene-set, adult inhibitory neurons were primarily implicated. This remained consistent for the full gene-set and up and down regulated portions of the gene-set. This suggests that *MEF2C* associated genes primarily affect excitatory neurons in development but inhibitory types are more affected in adulthood. Additionally, cortical inhibitory neurons derived from both the caudal and medial ganglionic eminence were found to be enriched for all variants of the KO_hom gene-set. Cortical interneurons have previously been shown to have reduced density and activity in ASD (Contractor et al., 2021). While this gene-set was not enriched for common variants associated with ASD (Cosgrove et al., 2021), *MEF2C* is associated with neurodevelopmental disorders and the cKO mice display autism-like symptoms (Harrington et al., 2016; Zhang and Zhao, 2022b). The use of this human developmental transcriptomic dataset also identified OPCs as an enriched cell-type which is in contrast to previous studies which exclusively identified neuronal cell-types (Cosgrove et al., 2021). This suggests that some enriched cell-types may be missed when exclusively using mouse data for this analysis. Furthermore, the oligodendrocyte transcriptome is thought to be perturbed in ASD according to postmortem studies and mouse models of ASD and the OPC enrichment may suggest changes that affect mature oligodendrocytes (Phan et al., 2020).

The use of the SCZ single-cell dataset also provided useful insights into the cell-type enrichment for the KO_hom gene-set. While lacking a developmental insight it did contain

9 additional cell-types and provided information regarding the regional locations of particular cell-types. Consistent with both the above results and that of (Cosgrove et al., 2021), enriched cell-types were predominantly neuronal. However, there was a stronger enrichment for inhibitory neuronal cell-types and impaired inhibitory signalling is associated with neurodevelopmental disorders (Tang et al., 2021). Parvalbumin neurons which are associated with both schizophrenia and ASD (Enwright et al., 2018; Filice et al., 2020) were found to be highly enriched in both the full gene-set and the KO_hom_down gene-set. Parvalbumin neuron perturbation in schizophrenia is thought to be due to mitochondrial dysfunction while their role in ASD is yet to be elucidated despite a decrease in their number observed in postmortem tissue (Enwright et al., 2018; Filice et al., 2016). Regionally, cells from layers five and six were the most highly enriched cell-types. Layer five excitatory neurons has been previously shown to have distinct differential gene expression between schizophrenia cases and controls which is consistent with this result (Arion et al., 2017). In addition, snRNAseq of schizophrenia cases and controls shows that genes preferentially expressed in layer five are located in genomic regions enriched for common variants associated with schizophrenia (Ruzicka et al., 2022). Furthermore, this study found that many DEGs across cell-types in this study were *MEF2C* target genes which again provides support for the role of *MEF2C* in schizophrenia (Ruzicka et al., 2022). All variants of the KO_hom gene-set were also observed to be enriched in the DEGs from this dataset across multiple cell-types including excitatory and inhibitory neurons, which supports the cell-type enrichment results. These gene-sets were also enriched in glial cell types such as astrocytes and microglia, which was not observed with the cell-type enrichment analysis. This suggests that these cell-types may also be affected by the homozygous knockout.

There was a distinct lack of enrichment in either single-cell dataset using the KO_het gene-set. There was no significant enrichment using the SCZ single-cell dataset while only excitatory neurons from the late-fetal stage were enriched using the developmental dataset. While the excitatory enrichment is expected, more than one enriched cell-type would have also been expected. This gene-set is smaller than the KO_hom gene-set so it is possible that this is impacting the enrichment or that the genes in this set are not distinctive to any one cell-type. However, this gene-set was enriched in DEGs obtained from a variety of neuronal and non-neuronal cells which suggests that it may recapitulate some of the biology observed in neurodevelopmental disorders.

Coexpression fold enrichment was highest in interneurons across gene-sets which suggests that interneurons are affected in both *MEF2C* knockouts despite the experimental differences between them. Fold enrichment was also high in excitatory neurons which, is consistent with previous studies in addition to the cell-type enrichment analysis from this chapter (Fahey et al., 2023; Cosgrove et al., 2021). This shows that using gene-sets indi-

vidually and by taking co-expression patterns into account will still implicate similar cell types. The KO_hom_up gene-set showed much higher fold enrichment in excitatory neurons that down-regulated genes, which suggests that the co-expression dynamics are different between the gene-sets despite convergence on the same cell-type. The KO_hom_up gene-set also showed high enrichment in microglia, which is not as strong in the results from other gene-sets. Microglia were also not found to be enriched in the EWCE analysis despite high expression of *MEF2C* in microglia (Ruzicka et al., 2022). This suggests that when considering co-expression of multiple genes, cell-types that were missed by other analysis may be identified. When comparing co-expression fold enrichment at different gestational weeks, fold enrichment in NPCs was highest at GW 10 while it was highest at GW 23 in interneurons. This illustrates the difference in co-expression at different timepoints.

There are however a number of limitations to this analysis. The single-cell dataset used contain much fewer cell-types and decreased subtype diversity when compared to the mouse cell-types used in previous analysis (Cosgrove et al., 2021; Fahey et al., 2023). Furthermore the cKO from which the (Cosgrove et al., 2021) gene-set was derived was only carried out in excitatory neurons of the forebrain and the RNA-seq analysis was performed only on tissue from the somatosensory cortex. This implies that we are only getting a region specific view of the dysregulation induced by the cKO. Additionally, while the mutated neuronal population may affect other cell types, this does not reflect the changes that would be observed if the knockout was carried out in these cell-types. The KO_het gene-set is thought to be more reflective of the human condition yet both gene-sets are derived from mice and converted to human orthologs. Finally, the tissue obtained from the knockout models was exclusively adult tissue. As such, genes that are affected in pre-natal cell-types may not be detected as differentially expressed. Ideally, in order to characterise the dysregulation as a result of *MEF2C* mutation, a snRNAseq of multiple brain regions from embryonic or postmortem tissue would provide better insight into cell-type specific changes. This would however be much more difficult as postmortem and embryonic tissue is difficult to obtain and snRNAseq is expensive to carry out. The co-expression analysis was also limited to the six major cell-types in the fetal dataset. Extending this analysis to the other snRNAseq dataset proved difficult using the software from (Pang et al., 2020) yet may provide additional information were it to be extended in future. While this analysis is flawed in this respect, it does make use of existing data to provide new insights into the cell-types impacted by *MEF2C* knockout.

Chapter 4

Investigation of the Chromatin Accessibility and Transcriptional Landscapes of Schizophrenia at the Cell-type Level

Rebecca Mahoney¹, Jaroslav Bendl², Alexey Kozlenkov², Courtney Micallef², Zhiping Shao², Jonathan Edelstien², Gabriel E. Hoffman², Cathal Seoighe³, Derek W. Morris¹, Vahram Haroutunian⁴, Stella Dracheva⁴, John F. Fullard², Panos Roussos²

1. Neuroimaging, Cognition and Genomics (NICOG), School of Biological and Chemical Sciences, University of Galway, Ireland
2. Center for Disease Neurogenomics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
3. School of Mathematics, Statistics and Applied Mathematics, University of Galway, Ireland
4. Friedman Brain Institute, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA

4.1 Preamble to Chapter 4

This research presented in this chapter was the result of a collaboration between the University of Galway and the Centre for Disease Neurogenomics at the Icahn School of Medicine at Mount Sinai. My supervisors at the University of Galway were Prof. Derek Morris and Prof. Cathal Seoighe. All of my analysis for this project was supervised by Prof. Panos Roussos and Dr. Jaroslav Bendl from the Centre for Disease Neurogenomics at the Icahn School of Medicine at Mount Sinai.

Prof. Panos Roussos and Dr. Jaroslav Bendl supervised all analysis relating to this project and were involved in study design. Data generation was supervised by Dr. John Fullard and conducted by Dr. John Fullard, Dr. Alexey Kozlenkov, Courtney Micallef, Zhiping Shao and Jonathan Edelstien. Dr. Vahram Haroutunian and Prof. Stella Dracheva provided human brain tissue dissections and biopsies. Dr. Jaroslav Bendl and Dr. Gabriel Hoffman were involved in the initial quality control and processing of the generated data. I carried out all downstream analysis. This study is currently being prepared for submission.

4.2 Abstract

Schizophrenia (SCZ) is a severe neuropsychiatric disorder with a complex pathophysiology which is clinically characterised by psychosis, social withdrawal and cognitive deficits. Its etiology is still poorly understood but it thought to be impacted by a variety of genetic and environmental factors. The majority of previous studies that have utilised functional genomic data to explore the dysregulation in SCZ have focused on gene expression in bulk tissue. Here we performed transcriptome ($n=400$ libraries) and chromatin accessibility ($n=400$ libraries) profiling in two neuronal (GABAergic and glutamatergic neurons) and two non-neuronal (Oligodendrocytes and Microglia/Astrocytes) cell-types from 50 SCZ cases and 50 controls. We have expanded the current knowledge of open chromatin regions (OCRs) in the human brain by detecting many novel OCRs. We present disease-associated changes in gene expression and chromatin accessibility at the cell-type level. We observe an enrichment of common variants associated with SCZ in GABAergic and oligodendrocyte OCRs and identify extensive cell-specificity of our differentially expressed genes. To fully leverage the potential of our deeply sequenced transcriptome data, we utilised our 392 RNA-seq samples to conduct a differential transcript analysis, thus offering a complementary view that deepens our understanding of the transcriptomic nuances in SCZ. A large majority of the significant genes identified in the transcript analysis were not observed in the gene level analysis, indicating the need to study differences in transcript expression at cell-type resolution. Isoforms of *CACNA1C*, a well-known SCZ risk gene were differentially expressed at transcript-level only in oligodendrocytes an effect not observed in the other cell-types. We have identified cell-type-specific enhancer–promoter interactions by integrating our OCRs with Hi-C data which captures the three-dimensional structure of the genome. We then link disease-associated epigenetic change with SCZ GWAS signal and predict target genes using our enhancer–promoter interactions. Finally we conduct a cell-type specific eQTL analysis by imputing gene expression profiles for 848 individuals using our cell-type specific gene expression profiles as a reference. This greatly increased our power to carry out eQTL analysis to link expression changes to genetic variants. Overall, this analysis provides a comprehensive resource for characterising the transcriptomic and chromatin accessibility landscapes of SCZ.

4.3 Introduction

Schizophrenia (SCZ) is a severe, psychiatric disorder that affects between 0.7% and 1% of the world's population (Kahn et al., 2015). Symptomatically it usually manifests in late adolescence or early adulthood despite changes beginning in neurodevelopment and remains a chronic debilitating illness throughout the life of an individual (Ochoa et al., 2012). Phenotypic changes that are observed include positive symptoms such as delusions or hallucinations which are the most commonly known but also extends to negative symptoms which result in a reduction of normal behaviours related to motivation and interest and cognitive deficits (Kahn et al., 2015; Correll and Schooler, 2020; McCutcheon et al., 2023). Whilst SCZ is one of the most widely studied neuropsychiatric disorders, treatment still remains limited and affected individuals have a life expectancy of 10-20 years less than the general population (Correll et al., 2022b). Most available treatments rely on atypical antipsychotics which have remained the first line of treatment for the last 70 years despite many clinical trials (Kantrowitz et al., 2023). These treatments are also only partially effective with 10-30% of patients found to be refractory to treatment and only 10-20% of all patients achieving full recovery criteria (Kane et al., 2019; Jääskeläinen et al., 2015; Huxley et al., 2021).

As a result of the limited treatment options for this disorder, extensive research has been undertaken in an attempt to further elucidate the biology of this disorder. The etiology is not fully understood yet it is believed to be a complex mixture of genetic and environmental causes. Environment factors are thought to include obstetric complications, maternal immune activation, season of birth and cannabis use (Ursini et al., 2018; Choudhury and Lennox, 2021; French et al., 2015). Genetic factors which are of more interest in the context of this thesis are also thought to play a role both alone and by interaction with environmental known as genotype x environment effects (GxE) (Fan et al., 2018). In terms of genetics, SCZ is highly heritable with the heritability of SCZ thought to be approximately 79% (Hilker et al., 2018). It is a complex polygenic disorder with many loci implicated and 120 high confidence genes associated with SCZ in the latest GWAS (Trubetskoy et al., 2022). These genes are thought to impact a wide variety of pathways and functions including synaptic functions, calcium channels and excitatory neuronal functioning (Pardiñas et al., 2018; Trubetskoy et al., 2022). Yet GWAS is thought to only explain 24% of SNP heritability which implies that common variants only explain a subset of the heritability of SCZ (Trubetskoy et al., 2022). Rare variants are also thought to contribute to SCZ but until recently were hard to identify due to the lack of power in small samples sizes. The recent SCHEMA exome sequencing study identified 10 genes with ultra-rare coding variants at exome-wide significance (Singh et al., 2022). Yet, both common and rare genes associated with SCZ appear to converge on the same pathways

and functions. As a result it is becoming increasingly important to understand how exactly these changes contribute to SCZ development.

Functional genomics studies have aimed to characterise the underlying neurobiology of SCZ. Studies of both the transcriptome and the epigenome have uncovered a number changes in gene expression and molecular phenotypes related to gene regulation such as chromatin accessibility, methylation and histone modifications. As this chapter will focus on gene expression and chromatin accessibility changes, I will focus on these two molecular changes. Postmortem gene expression studies have revealed changes in brain regions including the dorsolateral prefrontal cortex (DLPFC), which is the most studied region, and the hippocampus (Horváth et al., 2011). Subsequent larger scale studies by the CommonMind Consortium (CMC) and PsychENCODE have identified DEGs between cases and controls with the greater number of DEGs in the latter study attributable to the larger sample size (Fromer et al., 2016; Gandal et al., 2018). Both studies identified genes involved in synaptic components and neurodevelopment while pathway analysis from the PsychENCODE study showed that downregulated genes significantly implicated transmembrane and signalling receptors. Transcript expression analysis has also been invaluable in highlighting changes that may not be observed at gene level. While there is often a large overlap between DEGs and differentially expressed transcripts (DETs), it has been observed that DETs have larger effect size than gene-level results particularly for protein coding genes (Gandal et al., 2018). Additionally, genes that were only significant at transcript level are of particular interest as their effects may be masked by the expression of highly expressed genes when only gene-level analysis is conducted. Cell-type enrichment analysis has also enabled us to pinpoint cells in which the observed expression alterations are enriched. In relation to SCZ, this has highlighted excitatory neuronal cells and oligodendrocytes for down-regulated genes while astrocytes were enriched for upregulated DEGs (Gandal et al., 2018). Isoform-only DEGs also exhibited overlap with excitatory neurons confirming them as a major cell type of interest in SCZ due to the concordance with prior studies (Gandal et al., 2018; Skene et al., 2018). These results also highlight a major limitation of many studies of the transcriptome in SCZ. Despite the ability to detect enriched cell-types, these studies are not truly able to inform us of cell-type specific changes in SCZ as they utilise bulk tissue which is an overview of all expression in a tissue rather than particular cell-types. As such, there has been an increasing focus on cell-type specific transcriptomic changes in SCZ.

Cell-type specific transcriptomic changes in SCZ have been explored by a small number of studies, yet our knowledge on this subject remains limited. These studies have used targeted methods such as FACS or laser microdissection to isolate individual cell-types. Findings from these studies include disruption of the ubiquitin protease and mitochon-

drial pathways in pyramidal neurons, which was only observed in SCZ cases and not cases of schizoaffective disorder thus identifying cell-type and disease specific changes (Arion et al., 2017). Transcriptomic profiling of the human dentate gyrus granule cell layers which is primarily composed of granule cell neurons also uncovered changes that have previously been missed by bulk studies in addition to eQTL associations for *CACNA1C* a known SCZ risk gene (Jaffe et al., 2020). The number of cell-enriched eQTL associations was also greatly increased from previous studies of smaller sample size showing the need for cell-specific profiling (Jaffe et al., 2020; Dong et al., 2018). Profiling of parvalbumin neurons similarly implicated mitochondrial dysfunction in addition to tight junction signalling (Enwright et al., 2018). As the individuals in this study were also profiled in a prior study of pyramidal neurons (Arion et al., 2015), the changes across both cell types were compared with more than 80% of alterations found to be specific to a particular cell-type.

Other molecular alterations in SCZ can also be informative, particularly in relation to gene regulation. As with transcriptomic studies the majority of these studies are in bulk tissue but some have profiled particular cell-types. Analysis of bulk chromatin accessibility in SCZ identified three regions which were differentially accessible between cases and controls but were not found to be enriched for genes associated with SCZ (Bryois et al., 2018). A number of DEGs identified by a previous study were found to have nearby ATAC-seq peaks yet none of these were differentially accessible, which was thought to be attributed to either cell-type specific effects that were masked in bulk tissue or lower statistical power (Bryois et al., 2018). Indeed, cell-type specific effects in chromatin accessibility have been observed in controls with glutamatergic neuron peaks enriched for SCZ risk variants (Hauberg et al., 2020).

Cell-type specific datasets in SCZ are extremely limited and often have smaller sample sizes and limited numbers of cell-types. In this study we have isolated nuclei from two neuronal (GABA:GABAergic and GLU:glutamatergic neurons) and two non-neuronal (OLIG:Oligodendrocytes and MGAS:Microglia/Astrocytes) cell-types from the dorsolateral prefrontal cortex (DLPFC) in SCZ cases and controls in order to characterise cell-type-specific changes at both the transcriptomic and the epigenomic levels using RNA-seq and ATAC-seq profiling. We identified differentially accessible open chromatin regions (OCRs) in addition to the identification of genes showing significantly different expression between cases and controls. Given the limited research on SCZ at the transcript level, our study aimed to address this gap by prioritising genes with cell-type-specific DETs. This often revealed dysregulation of transcripts for genes that did not individually reach statistical significance. Finally, we leveraged these data to enable a QTL analysis by using these data to identify cell-type proportions in bulk RNA-seq data and using those

proportions to impute expression for 848 bulk RNA-seq profiles. This then allowed for eQTL detection and colocalization analysis resulting in cell-type specific eQTLs which colocalized with SCZ risk loci. Overall, we aim to characterise cell-specific changes in SCZ at both the transcriptomic and epigenetic level.

4.4 Methods

4.4.1 Description of Samples

Dorsolateral prefrontal cortex (DLPFC) brain tissue was obtained from 100 individuals in the Mount Sinai Brain Bank comprising 50 SCZ cases and 50 controls. These individuals ranged in age from 24-101 and consisted of 34 females and 66 males. Full demographic information for these samples is available in Supplementary Table 4.1 (https://drive.google.com/drive/folders/1lR3Ccix_ccf4pWhXY6AjqtoLdI0Qym86?usp=sharing).

4.4.2 Fluorescence Activated Nuclear Sorting (FANS) of Four Different Cell-types

For each sample, 50 mg of frozen postmortem brain tissue was homogenised in cold lysis buffer (0.32M Sucrose, 5mM CaCl₂, 3mM Magnesium acetate, 0.1mM, EDTA, 10mM Tris-HCl, pH8, 1mM DTT, 0.1% Triton X-100). Following homogenisation, each sample was filtered through a 40µm cell strainer, and the flow through underlaid with sucrose solution (1.8M Sucrose, 3mM Magnesium acetate, 1mM DTT, 10mM Tris-HCl, pH8) prior to ultracentrifugation at 107,000xg for 1 hour at 4°C. Pellets were then resuspended in 500µl DPBS and incubated in BSA at a final concentration of 0.1% together with anti-NeuN antibody (1:1000, PE conjugated, Millipore Cat FCMAB317PE), anti-SOX6 and anti-SOX10. DAPI (Thermoscientific) was added to a final concentration of 1µg/ml directly prior to FANS sorting. GABAergic neurons (DAPI+NeuN+ SOX6+), Glutamatergic neurons (DAPI+NeuN+ SOX6-), oligodendrocytes (DAPI+NeuN- SOX10+) and MGAS (DAPI+NeuN- SOX10-) nuclei were sorted into individual tubes (pre-coated with 5% BSA) using a FACS Aria flow cytometer (BD Biosciences) equipped with FACS-Diva Version 8.0.1 software. FANS sorting yielded 100 samples per cell type resulting in a total of 400 samples.

4.4.3 RNA-seq Data Generation

RNA-seq libraries were generated from nuclei for each of the four cell types (GABAergic neurons, glutamatergic neurons, oligodendrocytes, a mixture of microglia and astrocytes). Where available, 100,000 FANS sorted nuclei were incubated in extraction buffer (Arcturus PicoPure RNA Isolation Kit, applied biosystems) at 42°C for 30 mins while shaking at 850 rpm. Samples were stored at -80°C prior to RNA extraction, according to manufacturer's instructions. RNA-sequencing libraries were generated using the SMARTer Stranded Total RNA-seq Kit v2 (Takara Bio USA). Libraries were quantified by Qubit HS DNA kit (Life technologies) and by quantitative PCR (KAPA Biosystems) prior to se-

quencing. Finally, RNA-seq libraries were sequenced on Hi-Seq2500 (Illumina) obtaining 2x100 paired-end reads.

4.4.4 ATAC-seq Data Generation

Nuclei from each of the four cell types (GABAergic neurons, glutamatergic neurons, oligodendrocytes, a mixture of microglia and astrocytes) were used to generate ATAC-seq libraries using a previously established protocol (Buenrostro et al., 2015). For each sample, 75,000 sorted nuclei were centrifuged at 500g for 10min at 4°C followed by thorough resuspension of pellets in transposase reaction mix (25µL 2× TD Buffer (Illumina Cat #FC-121-1030) 2.5µL Tn5 Transposase (Illumina Cat #FC-121-1030) and 22.5µL Nuclease Free H₂O) on ice. Reactions were incubated at 37°C for 30 minutes, after which they were purified using the MinElute Reaction Cleanup kit (Qiagen Cat #28204), eluting in 10µL of buffer EB. Following amplification, barcoded libraries were resolved on 2% agarose gels and fragments ranging in size from 100-1000bp were excised and purified (Qiagen Minelute Gel Extraction Kit—Qiagen Cat#28604). Prior to sequencing, libraries were quantified with the Qubit dsDNA HS assay kit (Invitrogen Cat#Q32851) and using quantitative PCR (KAPA Biosystems Cat#KK4873). In addition, fragment sizes were estimated using TapeStation D5000 ScreenTapes (Agilent technologies Cat# 5067-5588) and libraries were sequenced on Hi-Seq2500 (Illumina) obtaining 2×150 paired-end reads.

4.4.5 Alignment of Raw Sequencing Files

FASTQ files for the respective samples were matched based on their pooling IDs and barcodes. Raw reads were then trimmed for adapter sequences using the following parameters for Trimmomatic (v0.36) (Bolger et al., 2014):

```
ILLUMINACLIP::2:30:10:8:TRUE LEADING:3 TRAILING:3  
SLIDINGWINDOW:4:15 MINLEN:36
```

The STAR alignment tool was used to map the trimmed reads on a modified version of hg38 (Dobin et al., 2013). We corrected for allelic bias resulting from person-specific genome variation by running STAR with the WASP module enabled (van de Geijn et al., 2015). To do this we provided the ATAC-seq FASTQ file and the whole genome sequencing (WGS) file or SNP array genotype of each individual. The following settings of the STAR aligner (v.2.7.0e) were used:

```
-alignIntronMax 1 -outFilterMismatchNmax 100 -alignEndsType Local  
-outFilterScoreMinOverLread 0.66 -outFilterMatchNminOverLread 0.66
```

This resulted in a BAM file per sample which contained mapped paired-end reads sorted by genomic coordinates. Reads that mapped to multiple loci were excluded using samtools (Li et al., 2009), duplicated reads were excluded with PICARD (v2.2.4), (<http://broadinstitute.github.io/picard>), and reads mapping to the mitochondrial genome were also removed.

4.4.6 Peak Calling

Peaks were called for the ATAC-seq samples using MACS (MACS, v2.1) for each cell type (Zhang et al., 2008). The settings for MACS were as follows :

```
“-keep-dup all -shift -100 -extsize 200 -nomodel”
```

For each cell type, we used the same number of reads from each sample to create a per-cell-type BAM file that is set to have 500 million reads. MACS was used to call peaks with an FDR (-q parameter) of 0.01. Four sets of peaks (one from each cell-type) were then merged to create a consensus set comprising 464,152 peaks.

4.4.7 Quality Control

Quality control metrics used for both RNA-seq and ATAC-seq samples included the total number of initial reads, the fraction of reads that were uniquely mapped; the number of uniquely mapped reads, GC content, insert and duplication metrics from PICARD and the rate of reads mapping to the mitochondrial genome. In addition, for the ATAC-seq samples the fraction of reads in peaks (FRiP), which is the fraction of reads that fall in called peaks, the fraction of reads in only the blacklisted peaks and the ratio between these two metrics. On average the number of reads per sample was 83 million and 84 million for the ATAC-seq and RNA-seq samples respectively. ATAC-seq samples with low FRiP (less than 3% in a peak set created from one hundred randomly selected samples by requiring a peak to be called in one or more of the sample BAM-files) were removed. In addition, for both sets of samples, individuals with low mappability (less than 50%), low GC content or low final read count (less than 5,000,000) were excluded from further analysis. This resulted in a final set of 317 ATAC-seq samples and 392 RNA-seq samples which were used to construct count matrices for the respective data-types. These matrices were then used as an input to Rtsne package with a perplexity parameter of 7 in order to perform a t-SNE based clustering of the samples.

4.4.8 Overlap of ATAC-seq peaks with Existing Annotation

The R package ChIPSeeker (Yu et al., 2015) was used to assign a genomic context to our detected OCRs with these contexts defined as promoter ($\pm 3\text{kb}$ of any TSS), 5'-UTR, 3'-UTR, exon, intron, and distal intergenic. The closest gene of each OCR was also assigned. We compared the ATAC-seq OCRs to previously reported open chromatin regions (from Roadmap Epigenomics Consortium (REMC) ((Roadmap Epigenomics Consortium et al., 2015)), The Cancer Genome Atlas (TCGA) (Corces et al., 2018), and the Brain Open Chromatin Atlas (BOCA) (Fullard et al., 2018)). We calculated this overlap using the Jaccard index. The Jaccard index was taken as the intersection of base pairs divided by union of base pairs. Only previously known OCRs were considered and the novel peaks identified in this study were not included. We utilised the imputed datasets for REMC as they were of higher quality and the samples were grouped as previously described by (Fullard et al., 2018) in order to reduce dimensionality. For TCGA, we considered Glioblastoma Multiforme, and Low Grade Glioma samples as brain samples. For BOCA, only cortical samples were used and samples from the amygdala, insula, hippocampus, mediodorsal thalamus, nucleus accumbens, putamen, and the primary visual cortex were removed.

4.4.9 Differential Chromatin Accessibility Analysis

In order to identify OCRs that showed differential accessibility in SCZ, a differential chromatin accessibility analysis was performed. Chromatin accessibility was assessed by calculating the number of ATAC-seq reads that overlapped a given OCR with higher counts representing more accessible regions. A count matrix was constructed for each cell type and filtered to include only OCRs where counts per million (CPM) was ≥ 1 in at least 20% of the samples. The Bayesian information criterion was then used to select covariates which improved the overall model. The covariates tested included fraction of peaks in reads (FRiP), GC content metrics, mapping metrics, insert metrics, predicted cell type ratios, ethnicities, the rate of reads mapping to the mitochondrial genome, polymerase chain reaction (PCR) bottle neck coefficient (PBC), relative strand cross-correlation coefficient (RSC), and barcode. Sex and diagnosis were included in the base model. Following this we identified for each additional covariate the number of OCRs which showed an improved BIC score minus the number of OCRs which showed a worse BIC score when the covariate was included in the linear regression model compared to when it was not included. Covariates were then required to improve the mean BIC per OCR by at least 5 for it to be included in the final model. The GC content (represented by two complementary metrics) emerged as the sole technical covariate and was included in the model.

The normalised count matrices were then modelled jointly using the limma R package

for each cell type using the `voomWithDreamWeights` function (v.3.38.3) (Ritchie et al., 2015) which uses sample-level and observational-level weights. Firstly, Voom residualises the read counts and fits a mean-variance function across all OCRs to account for the fact that more accessible OCRs show lower variance. The observation level weights are then set as the inverse of the estimated variance. Sample weights are similarly estimated and used to calculate a final set of weights. The normalised read count matrices from `voomWithDreamWeights` were then modelled by fitting weighted least-squares linear regression models estimating the effect of the right hand side variables on the accessibility of each OCR. The `dream` function from the `variancePartition` R package (v1.0.27) (Hoffman and Schadt, 2016; Hoffman and Roussos, 2021) was then used to estimate the number of differentially accessible OCRs and a multiple testing correction was applied (BH-corrected P-value < 0.05). `Dream` was used as the same individuals reoccur for each cell type and it accounts for correlation structure in repeated measures which enabled us to avoid inflating the false discovery rate. Residualised count matrices for each cell-type were subsequently constructed by removing the effect of the variables mentioned above but retaining the effect of case/control status.

4.4.10 Differential Gene Expression Analysis

Genes which displayed differential expression in SCZ cases were identified by performing a differential gene expression analysis. A count matrix was constructed for each cell type and filtered to include only genes where the CPM was ≥ 1 in at least 20% of the samples. The Bayesian information criterion approach described above was then used to select covariates and the GC content (represented by two complementary metrics) emerged as the sole technical covariate. The normalised count matrices were then modelled jointly using the `limma` R package (Ritchie et al., 2015). In order to identify the differentially expressed genes (DEGs), the `dream` function from the `variancePartition` R package was used (Hoffman and Schadt, 2016; Hoffman and Roussos, 2021). A multiple testing correction was also applied (BH-corrected P-value < 0.05).

4.4.11 Differential Transcript Analysis

`Limma` (Ritchie et al., 2015) was used to model differential expression at transcript level using the same process for the gene-level analysis described above. No additional covariates were used as none were observed to improve the model following testing using the Bayesian information criterion. We then wished to prioritise genes which showed differential transcript expression. To do this we used the `remaCor` R package (<https://gabrielhoffman.github.io/remaCor>) which implements a random effects meta-analysis per gene to test for both deviation from the mean of zero and effect size heterogeneity (Han and Eskin, 2011; Han et al., 2016; Lee et al., 2017). The `LS()` function was used

to implement fixed effect meta-analysis for correlated test statistics using the method of (Lin and Sullivan, 2009). The correlation is set to the identity matrix for independent test statistics. The RE2C() function was used to implement a random effect meta-analysis for correlated test statistics to test for deviation of the mean from zero in addition to effect size heterogeneity. This function uses the RE2 method of (Han and Eskin, 2011) when test statistics are not correlated, or the RE2C method described by (Han et al., 2016) for correlated test statistics. Again, correlation is set to the identity matrix for independent test statistics. The analysis was carried out per gene with all applicable transcripts tested and subsequently corrected for multiple testing (BH-corrected RE2C P-value < 0.05). Our results focus on output from the RE2C as it is more powerful for experimental designs with correlated test statistics.

4.4.12 Gene Set Enrichment Analysis

Gene set enrichment analysis was carried out for both DEGs and differentially accessible OCRs. For OCRs, the GREAT approach was used to assign OCRs to genes (McLean et al., 2010). This involved restructuring the peak-sets into gene-sets which were comprised of the union of OCRs thought to regulate one or more genes in those gene-sets. These gene-sets were then supplied to the CameraPR() function within the limma R package (Ritchie et al., 2015). Gene-sets identified by the differential gene expression analysis were supplied directly to CameraPR(). This function tests to see if the gene set supplied is more significantly enriched than those not in the set. It considers all expressed genes or accessible OCRs and all genes/OCRs are pre-ranked using the output from Dream mentioned above. The gene-sets we tested against were from MSigDB 7.0 which allowed for a general analysis (Subramanian et al., 2005). We also used gene-sets from SynGO from a targeted approach to identify enriched synaptic processes (Koopmans et al., 2019).

4.4.13 Gene-set Analysis of Differentially Expressed Genes

MAGMA was also used to test enrichment of the DEGs per cell type in GWAS summary statistics for various brain-related traits (de Leeuw et al., 2015)

(<http://ctg.cncr.nl/software/magma>). We used summary statistics from published GWAS on SCZ (Trubetskoy et al., 2022) ($n=76,755$ SCZ cases, $n=243,649$ controls), intelligence (IQ) (Savage et al., 2018) ($n=269,867$) and educational attainment (EA) (Lee et al., 2018) ($n=766,345$). SNPs with available GWAS results are mapped to genes using GRCh37/hg19 coordinates. Gene P-values are then calculated for each set of summary statistics and the genes were padded by 35kb upstream and 10kb downstream in order to include regulatory region variants. Linkage disequilibrium (LD) was estimated using the European panel of the 1000 Genomes Project. This gene analysis is based on a mul-

multiple linear principal components regression model that accounts for LD between SNPs in each gene, number of SNPs in each gene, inverse of the mean minor allele count of variants in each gene and the GWAS sample size. We corrected for multiple testing using Benjamini-Hochberg correction to compute Q values.

4.4.14 Disease Heritability of Open Chromatin Regions using LD-score Partitioned Heritability

LD-score partitioned heritability (v.1.0.0) (Bulik-Sullivan et al., 2015) (<https://github.com/bulik/ldsc>) was used to ascertain whether the previously identified OCRs were enriched for common variants associated with various GWAS traits. Summary statistics from the following GWAS were used: SCZ (Trubetskoy et al., 2022), Bipolar Disorder (Mullins et al., 2021), Educational Attainment (Lee et al., 2018), Amyotrophic Lateral Sclerosis (ALS) (van Rheenen et al., 2021), Parkinson’s Disease (Nalls et al., 2019), Alzheimer’s Disease (AD) (Jansen et al., 2019), Autism (Grove et al., 2019), Depression (Als et al., 2023), Crohn’s Disease (Liu et al., 2015), Ulcerative Colitis (Liu et al., 2015), BMI (Locke et al., 2015) and drinks per week (Karlsson Linnér et al., 2019). LD-score partitioned heritability tests to see if common genetic variants that are located in regions of the genome that are of interest and explain more of the heritability than variants not located in these regions. It corrects for the number of variants in either group and enabled us to correct for possible biases present in the regions of interest. A baseline model of general genomic annotation was used and included coding regions and conserved regions. The regression model outputs both a P-value and a regression coefficient. We normalised the regression coefficient by the per-SNP heritability and called it the “heritability coefficient” in order to compare the coefficient across traits. This is different to the enrichment which is calculated by this tool as the heritability coefficient takes the baseline into account unlike the enrichment metric. We tested a number of phenotypes including SCZ, brain related traits and a variety of negative controls using European summary statistics only. LD-score analyses were only implemented with sets of OCRs which covered 0.05% or more of the human genome. This analysis was run separately for differentially accessible OCRs and general sets of OCRs.

4.4.15 Prediction of Enhancer-Gene Interactions

We used the Activity by Contact (ABC) model (ABC v.0.2) (Fulco et al., 2019) (<https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction>) to build a regulatory map of enhancer-promoter interactions in each cell-type (GABAergic neurons, glutamatergic neurons, oligodendrocytes and MGAS). This model required: i) information regarding the contact frequency between putative enhancers and promoters of regulated genes, and ii)

enhancer activity data. Contact frequency matrices were generated from Hi-C datasets for each of the cell-types mentioned. Enhancer activity data was ascertained using the ATAC-seq data from the current study. We filtered out predictions for genes on chromosome Y and genes which were lowly expressed. A minimum ABC score of 0.02 (default threshold) was used for validating interactions and the default screening window was 5 Mb around the TSS of each gene. We employed the ABC-MAX approach whereby we only retained E-P with this highest ABC score in a particular locus.

4.4.16 Identification of Genes Implicated by Dysregulated ATAC-seq Peaks

We sought to link the epigenetic changes observed in our ATAC-seq data with SCZ GWAS signal and predict genes that may be involved in this change. In order to accomplish this we first investigated whether our differentially accessible OCRs contained a SCZ GWAS SNP. We compiled a list of SCZ index SNPs and SNPs in LD with them ($R^2 \geq 0.8$) from the latest SCZ GWAS (Trubetsky et al., 2022). We used the Genomic Ranges R package (Lawrence et al., 2013) to find OCRs which overlapped these SNPs by using the `subsetByOverlaps()` function. We took these results and identified the overlap between these OCRs with a GWAS SNP and predicted E-P interactions in order to predict genes which may be contributing to dysregulation in these OCRs.

4.4.17 Cell-type Deconvolution and Gene Expression Imputation

4.4.17.1 Using FANS Data as a reference

The proportion of each cell-type (GABAergic neurons, glutamatergic neurons, oligodendrocytes and MGAS) contained within each sample present in bulk RNA-seq data ($n=494$ controls, $n=354$ SCZ cases) was estimated using dTangle (Hunt et al., 2019) using the FANS data as a reference. dTangle was chosen as it has been found to be one of the most reliable deconvolution tools for brain expression data and gives more accurate cell-type proportion estimates than other methods (Sutton et al., 2022). dTangle uses a linear mixed model in order to estimate cell-type proportions (Hunt et al., 2019). The bulk expression data is part of the CMC dataset which includes functional and genotypic data on controls and individuals with neuropsychiatric disorders (Hoffman et al., 2019). The 100 individuals present in the FANS data were also present in the bulk RNAseq count matrix. This bulk matrix was filtered to remove genes which had a CPM < 1 in 30% of samples in order to avoid lowly expressed genes affecting the deconvolution. Both the bulk and reference expression were normalised using the Trimmed Mean of M-values (TMM) method using the `limma` R package and converted to $\log(\text{CPM}+1)$. They were then sub-

setted to contain only common genes between the two matrices and combined into one matrix. The `normalizeBetweenArrays()` function within `limma` was used to normalise the expression intensities between the two datasets. `dTangle` was also used to identify marker genes in the FANS data using the function `find_markers()` which takes the bulk data and the reference data as input and outputs a list of marker genes per cell-type. Following this step, the proportions of each cell-type can be estimated using the `dtangle()` function. This function requires as input: i) our combined bulk and reference expression matrix, ii) the list of marker genes per cell-type, iii) a vector containing information on which samples of the combined matrix are to be considered reference profiles, and iv) the data-type which was “rna-seq”.

The resulting proportions were used as input for imputation of cell-specific gene expression profiles using `bMIND` (Wang et al., 2021). `bMIND` estimates per-sample cell-type specific expression profiles using a bulk expression matrix and an estimate of the proportion of each cell-type present in each bulk sample. It imputes expression data using Markov chain Monte Carlo (MCMC) (Wang et al., 2021). This resulted in the creation of four cell-type-specific count matrices ($n=494$ controls, $n=354$ SCZ cases). Spearman correlation coefficient was calculated between gene expression values of the imputed data and the FANS data in order to ascertain whether the imputed gene expression was representative of the gene expression in those cell-types.

4.4.17.2 Using Single-cell Data as a Reference

In order to ascertain the proportion of cell-subtypes present in the four cell-types profiled using FANS we used a single-nucleus RNA-seq (snRNA-seq) dataset as a reference for each of the four FANS expression matrices. snRNA-seq data was taken from a multi-omic study profiling chromatin accessibility and gene expression in both the adult and developing neocortex (Zhu et al., 2023). This dataset was filtered to include only adult individuals and contained nine cell-types (excitatory neurons, medial ganglionic eminence derived inhibitory neurons (IN-MGE), caudal ganglionic eminence derived inhibitory neurons (IN-CGE), oligodendrocyte precursor cells (OPCs), astrocytes, oligodendrocytes, microglia, endothelial cells, pericytes and vascular smooth muscle cells (VSMC)). Both expression matrices were filtered to remove genes which had a CPM < 1 in 30% of samples in order to avoid lowly expressed genes affecting the deconvolution. The snRNA-seq data was converted to “pseudobulk” by summing the expression values across all cells in a cell-type. Both the FANS and snRNA-seq expression were normalised using the TMM method using the `limma` R package and converted to $\log\text{CPM}+1$. They were then subsetted to contain only common genes between the two matrices and combined into one matrix. The `normalizeBetweenArrays()` function within `limma` was used to normalise the expression

intensities between the two datasets. Marker genes for the snRNA-seq dataset were chosen by using the cell-type marker genes identified in (Zhu et al., 2023). The proportions of each cell-type were then estimated using the `dtangle()` function. This function requires as input: i) our combined FANS cell-type and reference expression matrix, ii) the list of marker genes per cell-type, iii) a vector containing information on which samples of the combined matrix are to be considered reference profiles, and iv) the data-type which was “rna-seq”. This process was run separately for each of the FANS profiles except glutamatergic neurons as the snRNA-seq dataset did not contain excitatory neuronal subtypes.

Similar to the process followed for the proportions estimated for the bulk expression data, the proportions estimated for each FANS cell-type were used as input for imputation of cell-specific gene expression profiles using `bMIND` (Wang et al., 2021). This resulted in the creation of nine cell-type-specific count matrices (IN-MGE, IN-CGE, OPCs, astrocytes, oligodendrocytes, microglia, endothelial cells, pericytes, VSMC). Spearman correlation coefficient was calculated between gene expression values of the imputed data and the single-nucleus data in order to ascertain whether the imputed profiles were a good representation of the gene expression in those cell-types.

4.4.18 eQTL Detection in Expression Data Imputed from Bulk RNA-seq

In order to detect eQTLs using our imputed expression matrices generated from bulk RNA-seq data we used `mmQTL` which is a statistical package that better accounts for population structure (Zeng et al., 2022)(<https://github.com/jxzb1988/MMQTL>). Genotype data for these individuals is publicly available as part of the CMC (Hoffman et al., 2019). Genotypes were converted from variant call format (VCF) to PLINK format which is required to `mmQTL` using the PLINK software (V1.9) (Chang et al., 2015) (<https://www.cog-genomics.org/plink/>). Genotypes were subjected to standard quality control measures using PLINK. SNPs that were missing in at least 10% of individuals were removed (`-geno 0.1`) while no individuals were removed for excessive missingness (`-mind 0.1`). Variants with a minor allele frequency of less than 0.05 were also excluded (`-maf 0.05`) and Hardy-Weinberg Equilibrium (HWE) outliers were also removed (`-hwe 1e-6`). A genetic relatedness matrix was required for `mmQTL` and was generated using `GEMMA` (`-gk 2`) (Zhou and Stephens, 2012) (<https://github.com/genetics-statistics/GEMMA>) .

Bayesian Information Criterion (BIC) analysis was used to identify which technical variables needed to be corrected for in the imputed expression matrices in a manner explained in the differential analyses section of this chapter. We normalised the count matrices using the TMM method and transformed values to \log_2 CPM +1 using `voom` within the

limma R package (Ritchie et al., 2015). To improve eQTL detection, PEER factors were estimated by generating a range of latent factors (from 1 to 30) by using the peer R package (Stegle et al., 2010) in addition to the consideration of the variables of interest pre-selected by the BIC analysis. These variables and the latent variables generated by the peer R package were used to create a residualised matrix per cell-type. These residualised matrices were then used as input for mmQTL.

mmQTL was then used for QTL detection using the processed expression and genotype data. We carried out two rounds of multiple testing correction following (Kosoy et al., 2022). Multiple test correction was firstly applied per gene using the Benjamini-Hochberg (BH) and a BH-corrected P-value cutoff of 5% retaining the variant with the minimum q-value per gene. BH correction was then applied at genome-level on these retained variants with a cutoff of 5%.

4.4.19 Colocalization Analysis

Genetic colocalization analysis was carried out using coloc (<https://github.com/chr1swallace/coloc>) (Giambartolomei et al., 2014; Wallace, 2020). SCZ risk loci coordinates ($r^2 > 0.1$ with index SNP) were taken from the coordinates reported in the latest GWAS (Trubetskoy et al., 2022). For each locus we used the coloc.abf() function within coloc to test colocalization between the GWAS and eQTL signals. We used the default prior in coloc in addition to the beta coefficients from the GWAS and eQTL analysis and the sdY parameter was set to 1 for eQTL dataset. Significant colocalizations were ascertained by using the PP H4 cutoff of >0.75 .

4.4.20 Integration of ATAC-seq and RNA-seq Data

4.4.20.1 Canonical Correlation Analysis

In order to assess the shared correlation structure of the ATAC-seq and RNA-seq data we used Canonical Correlation Analysis (CCA). We subsetting both count matrices to ensure that the samples are matched between data-types, which is required for CCA. We used the Seurat R packages to identify variable features in each of the count matrices and then focused on the protein-coding genes that are highly variable. We then used a variant of CCA, known as diagonal CCA (Butler et al., 2018) to build our canonical correlation vectors.

4.4.20.2 Multi-omic Intergation with SNFtool

We sought to identify molecular subtypes of SCZ by using the multi-omic integration software, SNFtool (Wang et al., 2014). Similarity network fusion (SNF) constructs net-

works of samples, in this case the individuals in our count matrices for each data type and then fuses them into one network that represents all of the underlying data (Wang et al., 2014). A prior study, (Ramaswami et al., 2020) utilised this approach to identify molecular subtypes of autism spectrum disorder (ASD) which could then be subsetted from the original count matrices and analysed separately in order to identify changes that were previously masked.

Following the methods from (Ramaswami et al., 2020), we first regressed out technical and biological covariates including RNA integrity number (RIN), age and sex. Matrices were then subsetted to include only matched samples and matched genes. Data was normalised using TMM normalization which has been previously described in this section. Similarity matrices were calculated for each data-type using the `dist2()` and `affinityMatrix` functions within `SNFtool` using the default parameters of `SNFtool`. Both similarity matrices are then supplied to the `SNF()` function which constructs similarity networks and fuses the matrices. The resulting clusters were then viewed using the `spectralClustering()` function.

4.5 Results

4.5.1 Dataset Description and Quality Control

We performed ATAC-seq and RNA-seq profiling in nuclei isolated from SCZ cases ($n=50$) and controls ($n=50$) from four cell types (GABAergic neurons, glutamatergic neurons, oligodendrocytes, a mixture of microglia and astrocytes) derived from prefrontal cortex (PFC) tissue using FANS. FANS was used to differentiate neuronal and non-neuronal nuclei using an anti-NeuN (RBFOX3) antibody. The inclusion of an anti-Sox10 antibody allowed us to distinguish GABAergic (NeuN+ Sox10+) and glutamatergic nuclei (NeuN+ Sox10-). Similarly, within the non-neuronal cell population, an anti-Sox6 antibody was used to separate oligodendrocytes (NeuN- Sox6+) from remaining non-neuronal cells, predominantly microglia and astrocytes. Following quality control measures (Methods), 317 ATAC-seq and 392 RNA-seq samples remained comprising 464,152 OCRs and 22,600 expressed genes. An overall schematic of the experimental design is shown in Figure 4.1.

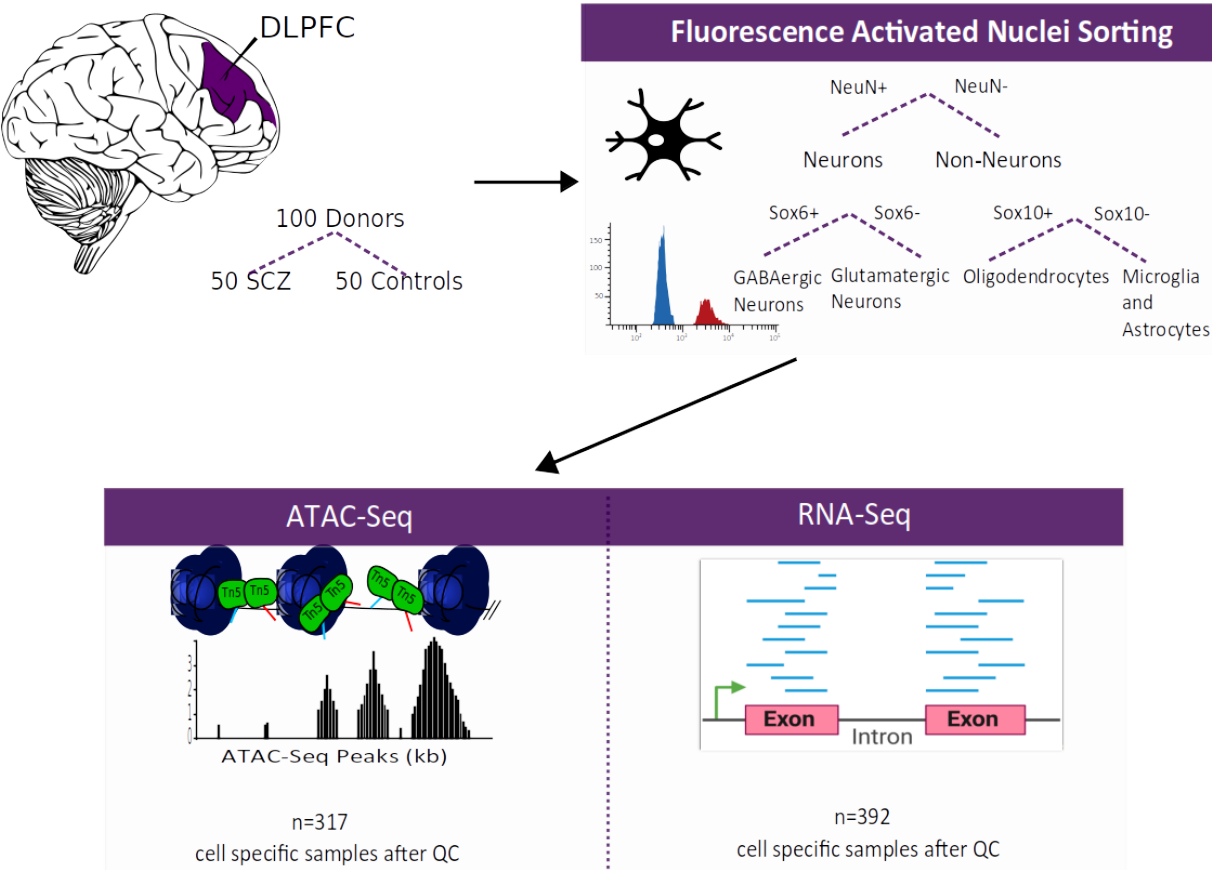


Figure 4.1: Schematic outlining the overall study design

t-SNE clustering was also performed using the respective count matrices generated per data-type and showed differences between the overall cell types (neuronal/non-neuronal). Subtype level difference were also observed between neuronal cell-types (GABAergic and glutamatergic neurons) and non-neuronal cell-types (oligodendrocyte and MGAS) (Figure 4.2).

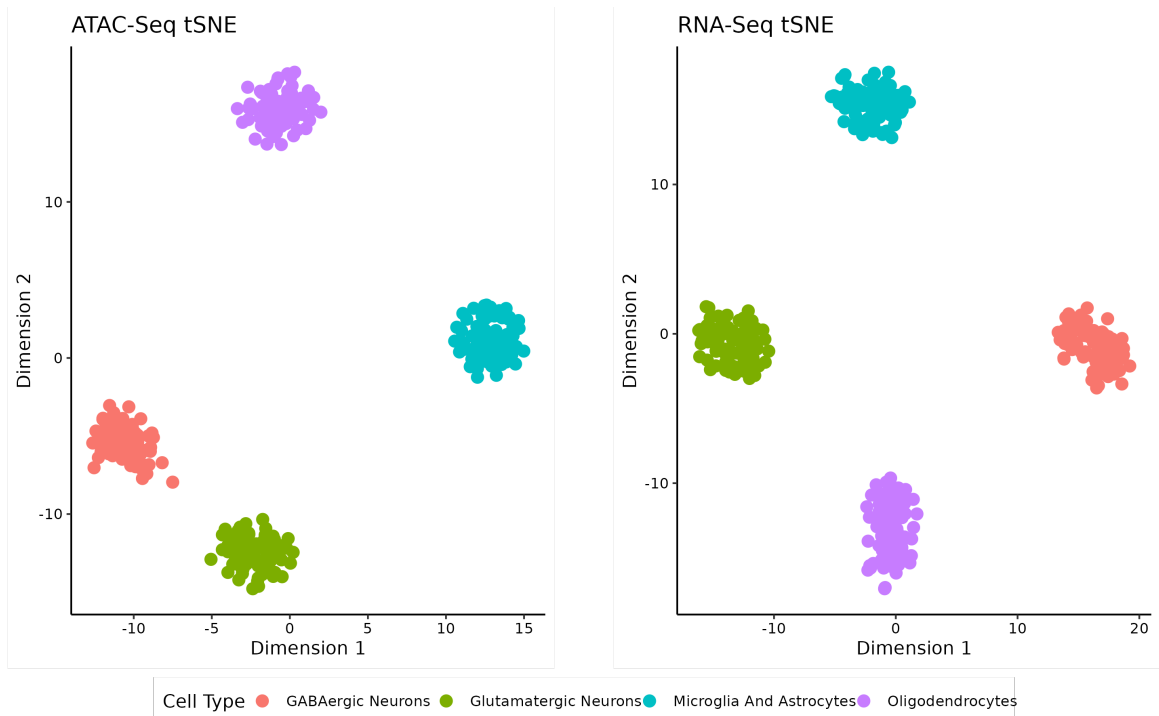


Figure 4.2: t-SNE clustering of samples using chromatin accessibility (left) and gene expression data (right)

4.5.2 Annotation of Open Chromatin Regions

We assigned genomic contexts to each OCR using the R package ChIPSeeker (Yu et al., 2015) with each OCR assigned one of the following annotations: promoter (± 3 kb of any TSS), 5'-UTR, 3'-UTR, exon, intron and distal intergenic. Of the OCRs identified, the highest proportion were located in intergenic (39.2 %) and intronic (39 %) genomic regions (Figure 4.3). We also compared our identified OCRs with previously published epigenome annotations from (Bryois et al., 2018; Corces et al., 2018; Bendl et al., 2022; Fullard et al., 2018; Hauberg et al., 2020) and (Nott et al., 2019), revealing that around 20% of them appear to be novel OCRs (Figure 4.4). We also visualised this overlap using the Jaccard index in the form of various heatmaps (Figure 4.5). The oligodendrocyte OCRs from this study showed strong overlap with brain OCRs from TCGA, bulk brain ATAC-seq data and oligodendrocyte OCRs (Figure 4.5(a,b,e,f)). Microglia/Astrocyte OCRs showed a weaker overlap with published studies but a higher overlap was observed

with microglia OCRs from cell-specific ATAC-seq data (Figure 4.5 (e,f)). GABAergic neurons and glutamatergic neurons displayed a strong overlap with neuronal cell populations as well as OCRs from purified GABAergic and glutamatergic neuron populations respectively (Figure 4.5 (b-f)). Stronger overlap was also observed within subtypes with GABAergic and glutamatergic neurons showing stronger overlap with each other or neuronal cells than non-neuronal cell-types (Figure 4.5). This trend was also observed for microglia/astrocytes and oligodendrocytes.

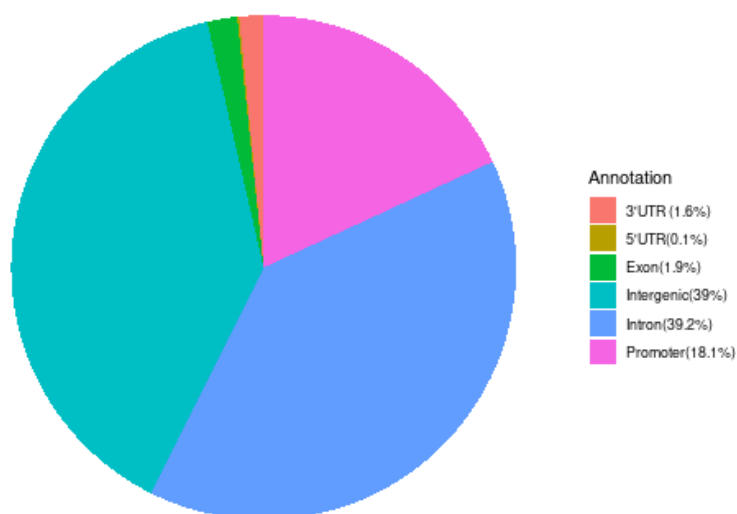


Figure 4.3: Proportion of OCRs corresponding to different genomic annotations. OCR regions located within a 3kb proximity to a transcription start site (TSS) were categorised as promoter OCRs.

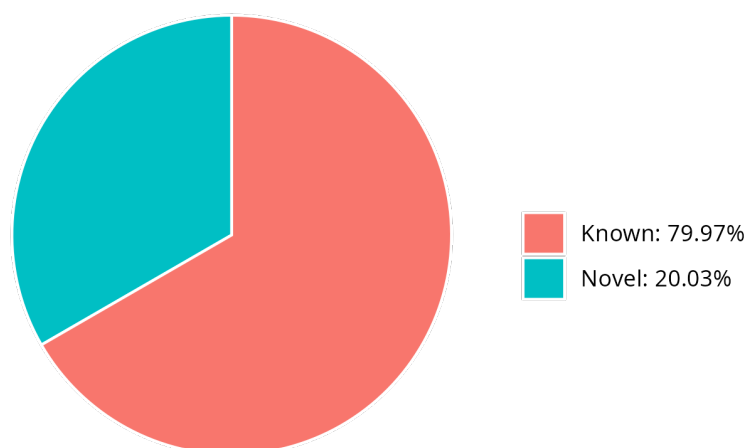
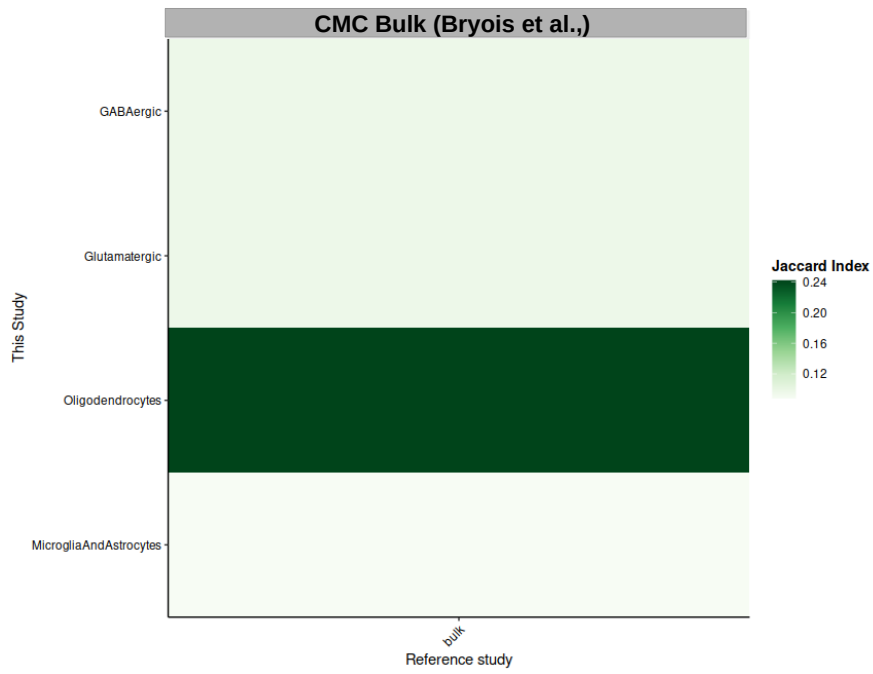
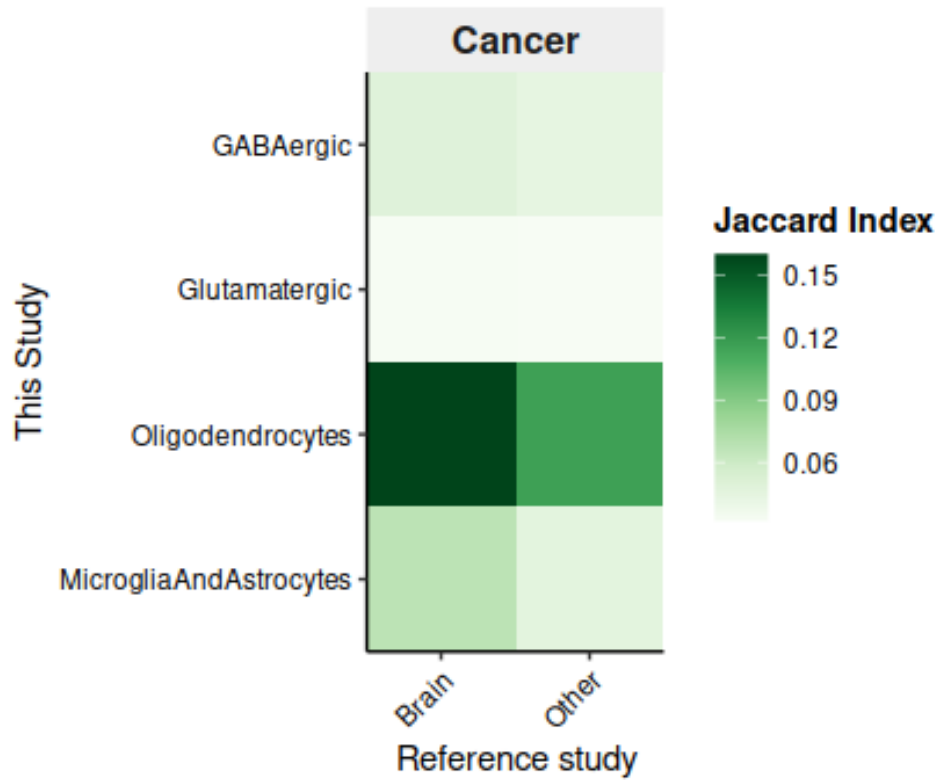


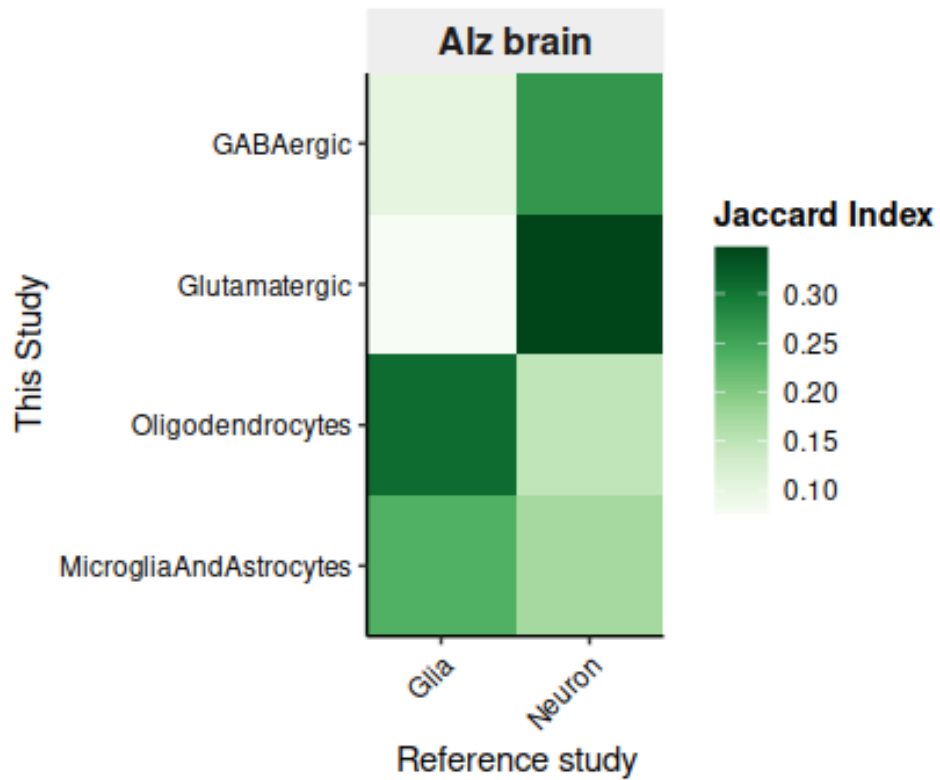
Figure 4.4: Proportion of known and novel OCRs. Novel OCRs are calculated based on comparison with established atlases of chromatin accessibility.



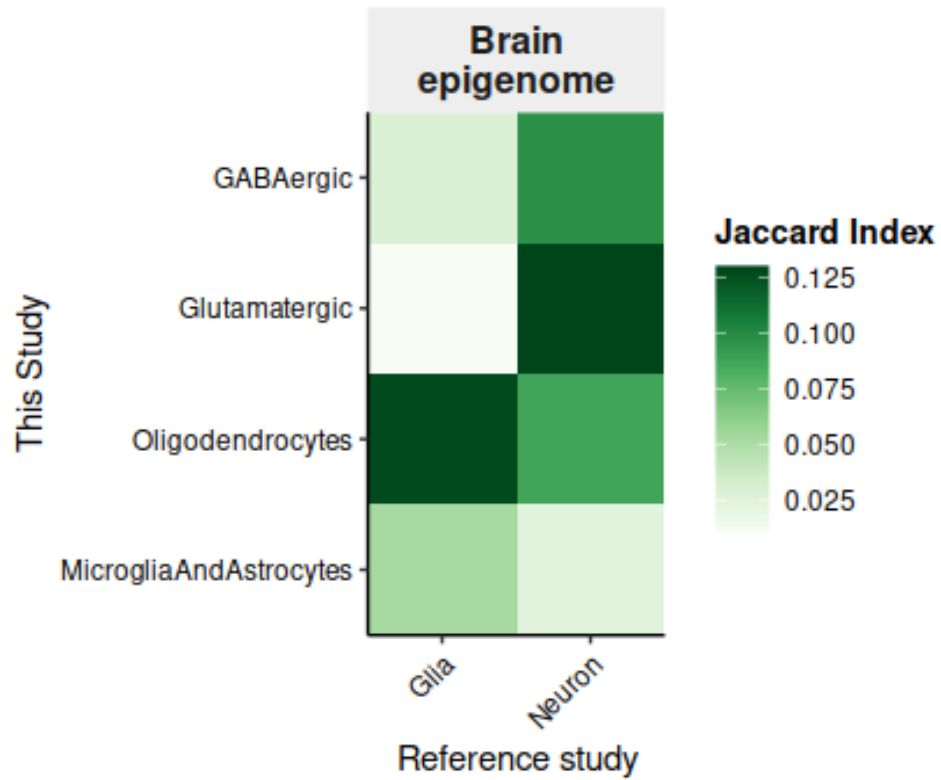
(a) Overlap with OCR annotation from CMC bulk ATAC-seq (Bryo et al., 2018)



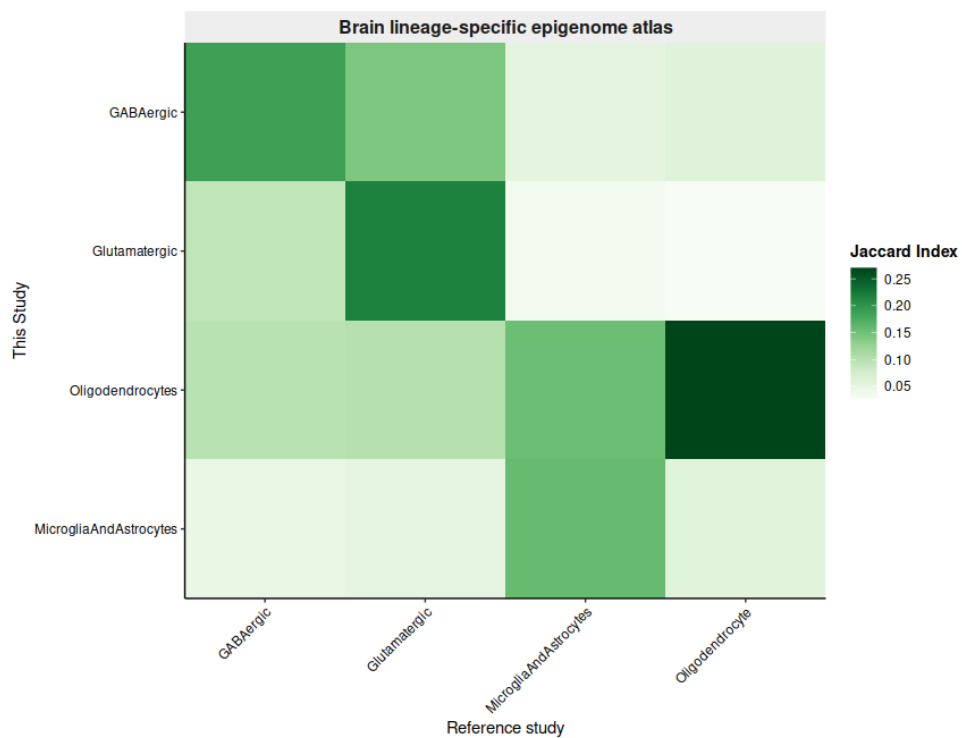
(b) Overlap with OCR annotation from TCGA ATAC-seq



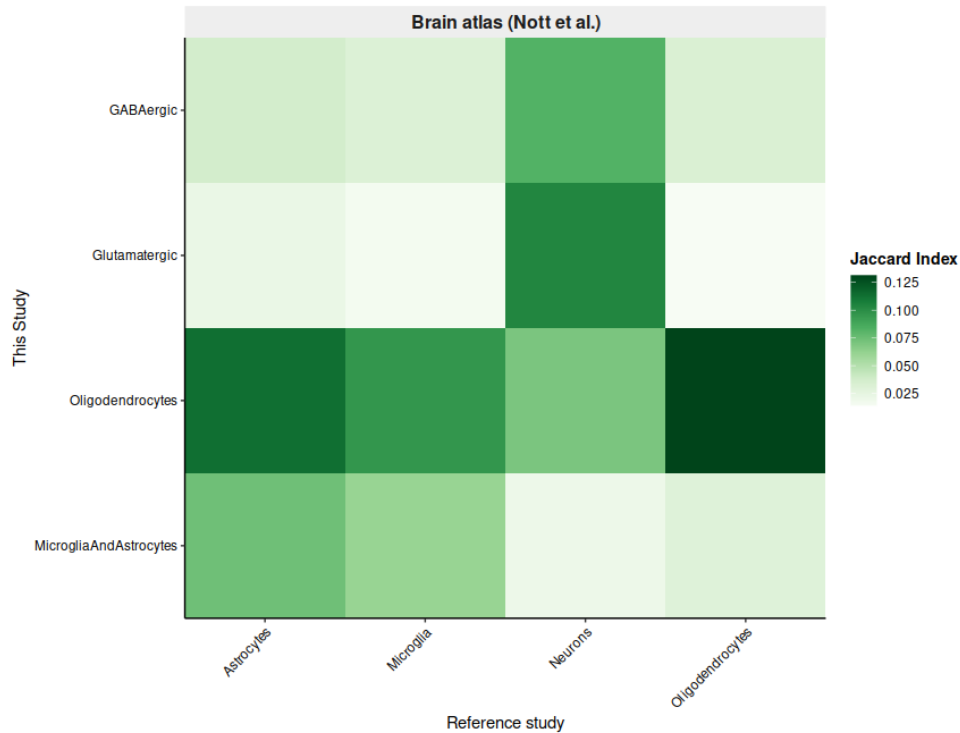
(c) Overlap with OCR annotation from Alzheimer's disease (AD) brain ATAC-seq from (Bendl et al., 2022)



(d) Overlap with OCR annotation from brain ATAC-seq data from (Fullard et al., 2018)



(e) Overlap with OCR annotation from brain ATAC-seq data from (Hauberg et al., 2020)



(f) Overlap with OCR annotation from the epigenomic brain atlas generated by (Nott et al., 2019)

Figure 4.5: Heatmaps visualising the overlap between the OCRs from this study and various reference studies using Jaccard index.

4.5.3 Differential Accessibility Analysis

To explore SCZ-associated changes in chromatin accessibility, we performed cell-specific differential analyses on 317 ATAC-seq samples from SCZ cases and controls. The combined influence of cell type and diagnosis status explained 39.7% of the variability in the ATAC-seq data. In a stepwise Bayesian information criterion approach (Methods), the GC content (represented by two complementary metrics) emerged as the sole technical covariate for the ATAC-seq data, contributing to 13.8% of the variance. The largest numbers of SCZ-associated OCRs were observed in non-neuronal cell types (OLIG: $n=24,633$, MGAS: $n=1,810$ at BH-corrected P-value < 0.05), followed by GABAergic neurons ($n=1,366$) (Figure 4.6). Additionally, the vast majority of differentially accessible OCRs were up-regulated across cell-types with the exception of MGAS. This was also observed in the differential analysis where cell-types were merged and considered as a bulk sample. A full table of differentially accessible OCRs is available in Supplementary Table 4.2 (https://drive.google.com/drive/folders/11R3Ccix_ccf4pWhXY6AjqtoLdI0Qym86?usp=sharing).

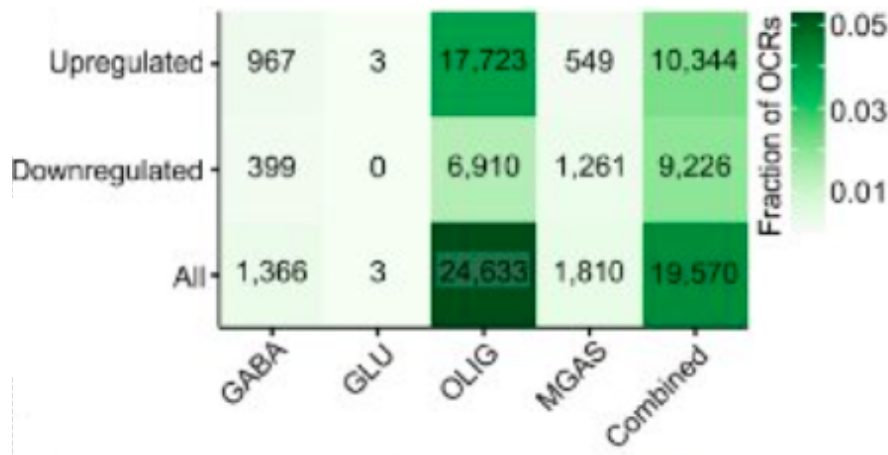


Figure 4.6: Heatmap displaying the total number of differentially accessible OCRs per cell-type (ATAC Merged) as well as the number of up-regulated OCRs (ATAC Up) and the number of down-regulated OCRs (ATAC Down) per cell-type. Fraction of OCRs refers to how many of the total OCRs were differentially accessible, darker shading indicates more of the total OCRs were differentially accessible between cases and controls.

4.5.4 Gene-set Enrichment Analysis of Open Chromatin Regions

Gene set enrichment analysis (GSEA) was performed in order to identify the biological processes potentially implicated in SCZ using the differentially accessible OCRs per cell-type (Figure 4.7). The altered pathways associated with SCZ in GABAergic neurons included general molecular pathways such as “C Terminal Protein Amino Acid Modification” and “DNA Topological Change”. Both oligodendrocytes and the combined set of differentially accessible OCRs implicated N-glycan synthesis, which is consistent with studies that have reported glycosylation abnormalities in SCZ postmortem brain tissue (Bauer et al., 2010; Tucholski et al., 2013; Mueller et al., 2014), in addition to mouse models which report altered glycosylation when modelling SCZ risk variants (Mealer et al., 2022). The “DNA fragment pathway” was also implicated in GABAergic neurons and oligodendrocytes. Interestingly, DNA fragmentation is a sign of apoptosis, which has been hypothesised to be affected in schizophrenia resulting in excessive dendritic spine loss (Parellada and Gassó, 2021). Indeed, SCZ organoids have been observed to have increased DNA fragmentation (Notaras et al., 2022).

We also employed a targeted strategy using the SynGO ontology resource. This analysis primarily implicated postsynaptic locations in GABAergic neurons such as the postsynaptic membrane and cytoskeleton (Figure 4.8).

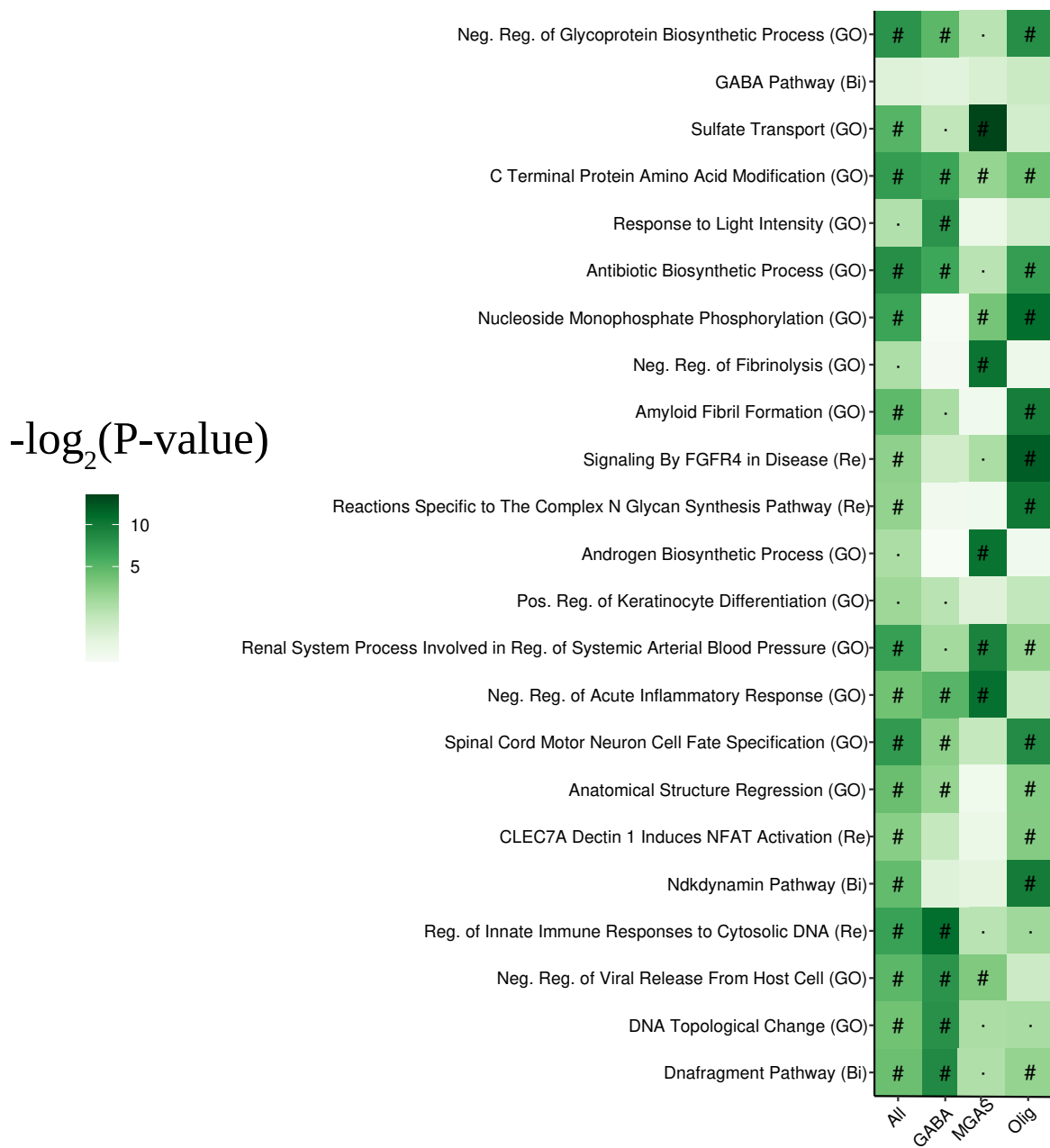


Figure 4.7: Heatmap visualising GSEA results for differentially accessible OCRs using MSigDb. #: significance at BH-corrected P-value < 0.05, a dot indicates nominal significance, Bi: Biocarta, GO: Gene Ontology, Re: Reactome.

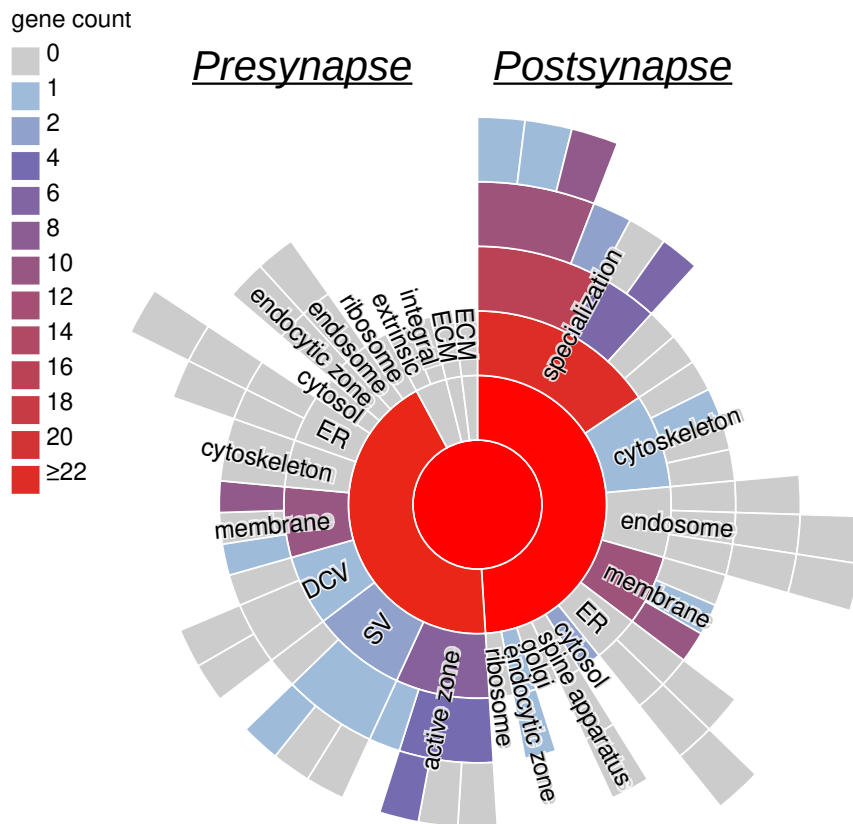


Figure 4.8: Sunburst plots depicting synaptic locations implicated by the GABAergic OCRs. This plot features the synapse at the center, pre- and post-synaptic locations in the first ring, and child terms in subsequent rings. The color scheme in the legend indicates the number of genes within each term.

4.5.5 Disease Heritability of Open Chromatin Regions

In order to examine the role that the OCRs detected in this study may have on disease, we tested whether they were enriched for common variants associated with various disorders using LDSC (Methods). This analysis was first carried out for all OCRs detected per cell-type as well as the consensus set. We found nominal enrichment of MGAS and oligodendrocyte OCRs for common variants associated with ALS and AD but no enrichment for SCZ (Figure 4.9). In addition, none of the results survived multiple testing correction. When these sets of OCRs were restricted to those that were differentially accessible, we

do observe enrichment for common variants with schizophrenia. SCZ-associated OCRs were significantly enriched in oligodendrocytes and GABAergic neurons (Figure 4.10). The full set of SCZ associated OCRs were also significant in the LD-score regression analysis at BH-corrected P-value < 0.05 . Other brain-related disorders and traits such as bipolar disorder and educational attainment were also nominally significant while other neuropsychiatric traits were not significant, perhaps due to lack of power in the GWAS.

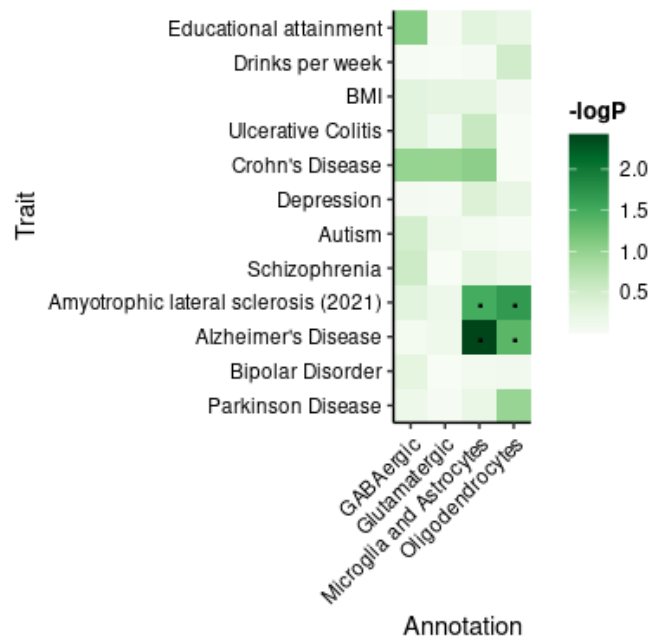


Figure 4.9: Heatmap for p-value enrichment of OCRs in common variants associated with a variety of traits using LDSC

We also used the regression coefficient normalised by the per-SNP heritability of the trait for each set of OCRs to further explore the overlap with risk of that trait. If the coefficient is positive it implies enrichment in heritability. Heritability enrichment was observed for ALS and AD in MGAS and oligodendrocytes, which was consistent with the previous analysis (Figure 4.11). When OCRs were restricted to those which were differentially accessible, significant enrichment in SCZ was observed for the combined set of OCRs (Figure 4.12). Nominal enrichment was observed for oligodendrocytes and GABAergic neurons, while no enrichment was observed for MGAS. This coefficient was also compared to those calculated by previous studies (Fullard et al., 2018; Hauberg et al., 2020) and shows stronger enrichment compared to previous studies.

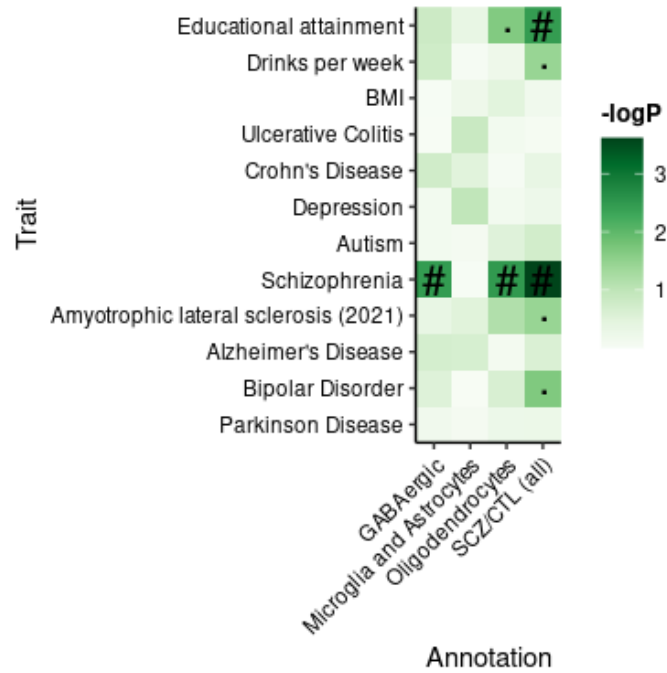


Figure 4.10: Heatmap for p-value enrichment of differentially accessible OCRs in common variants associated with a variety of traits using LDSC

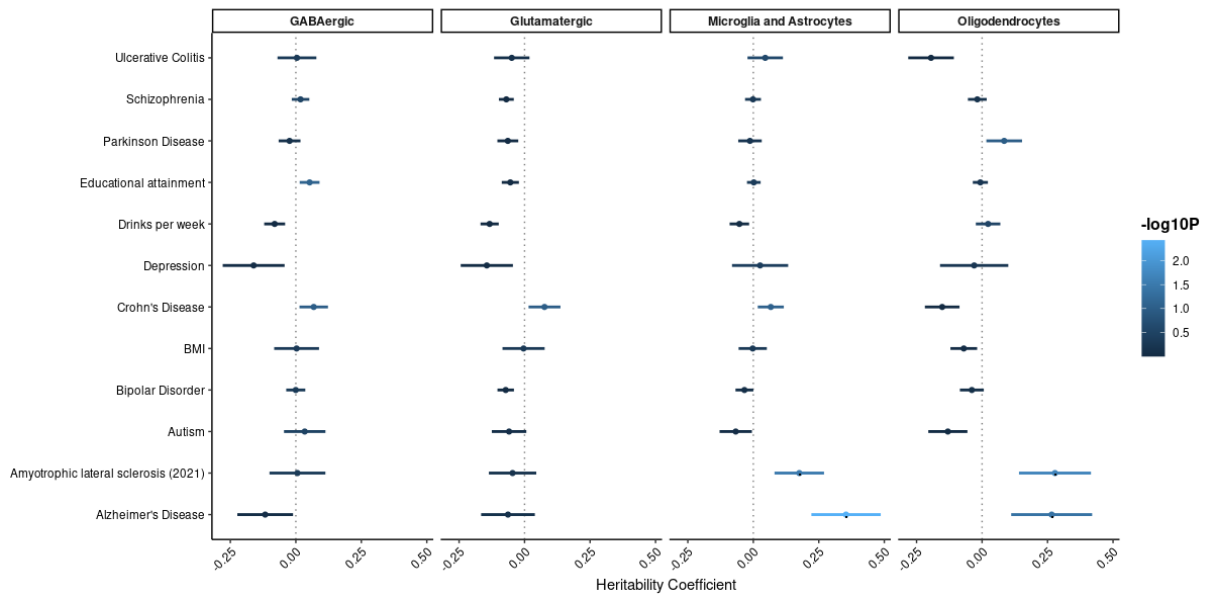


Figure 4.11: Heritability coefficients for various traits for all OCRs in a given cell type. This was calculated by using the regression coefficient normalised by the per-SNP heritability of a given trait.

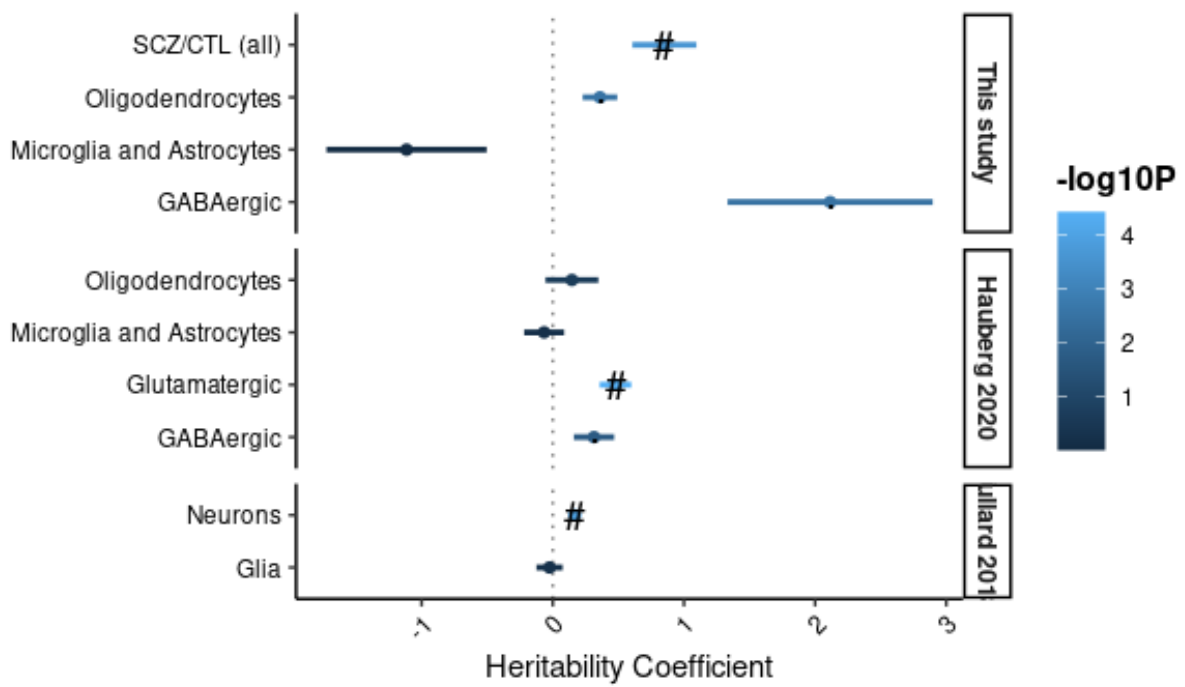


Figure 4.12: Heritability coefficients for SCZ for differentially accessible OCRs in a given cell type. Heritability coefficients observed for two previous studies, (Hauberg et al., 2020) and (Fullard et al., 2018) are also included for comparison.

4.5.6 Identification of Enhancer-Promoter Interactions and their Relationship with Differentially Accessible Open Chromatin Regions

We aimed to identify enhancer-promoter (E-P) interactions by integrating our ATAC-seq data with cell-specific Hi-C and ChIP-seq data using the ABC approach developed by (Fulco et al., 2019) (Methods). Across the four cell-types (GABAergic neurons, glutamatergic neurons, microglia/astrocytes and oligodendrocytes) we identified 75,135, 107,495, 74,881, 84,829 E-P interactions respectively. Most genes are linked to a distal OCR (OCR_{ABC}) (85%) with the majority linked to five enhancers or less (Figure 4.13). As a general pattern we observe, that most of the identified enhancers regulate a single gene, however a substantial number regulate two or more genes (Figure 4.14). A full table of E-P links estimated by the ABC analysis is available in Supplementary Table 4.3 (https://drive.google.com/drive/folders/1lR3Ccix_ccf4pWhXY6Ajqt0LdI0Qym86?usp=sharing).

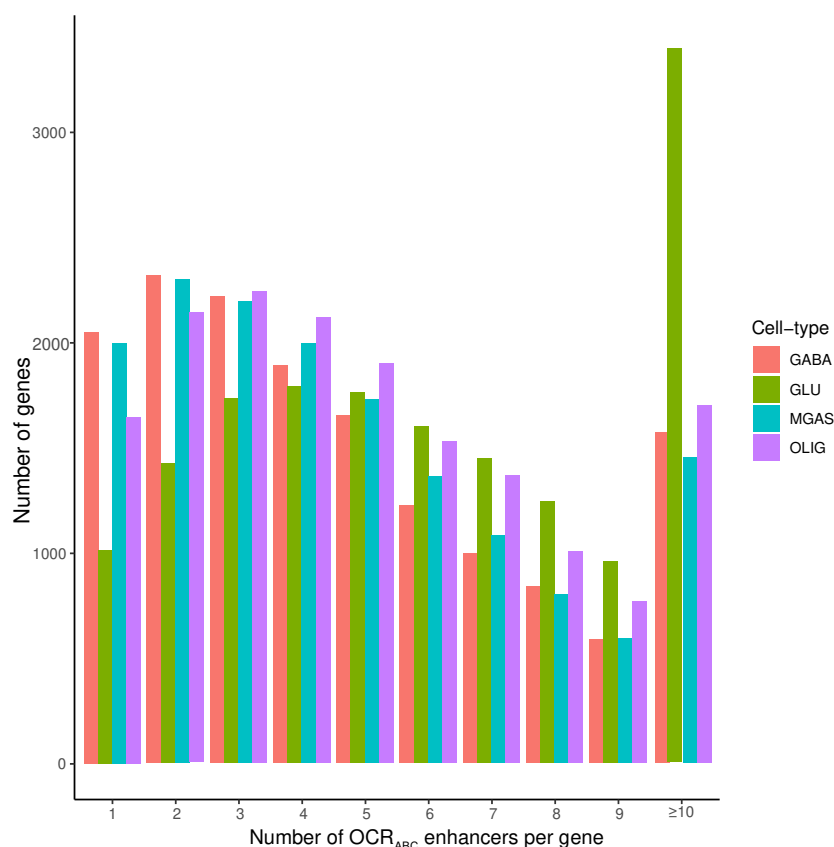


Figure 4.13: Histogram of the number of OCR_{ABC} linked per gene across the four cell-types.

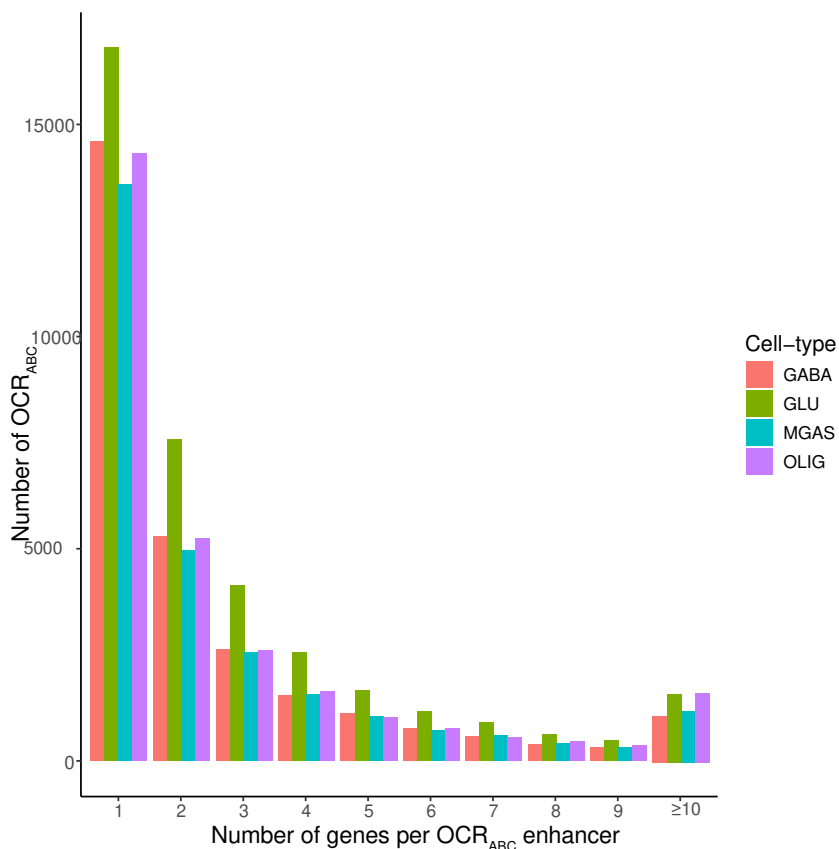


Figure 4.14: Histogram of the number of genes linked per OCR_{ABC} across the four cell-types.

We next aimed to leverage our E-P links to see if we could link genetic signal from SCZ GWAS with epigenetic changes. Our overall approach is illustrated in Figure 4.15. We first collected index SNPs from SCZ GWAS ($n=326$) and then included SNPs that were in high linkage disequilibrium ($R^2 \geq 0.8$) resulting in a total set of 11,445 SNPs. We then subsetted our differentially accessible OCRs per cell-type (excluding glutamatergic neurons due to the low number of differentially accessible OCRs) to include only those which overlapped a SCZ-associated SNP. We did not observe any overlap for MGAS and there were only two overlaps for GABAergic neurons (two distinct loci). However, there were 66 overlaps across 31 distinct loci for oligodendrocytes. We then took the OCRs which overlapped a GWAS SNP and assessed whether this region also overlapped an E-P link from our ABC analysis. In order to avoid multiple associations for one locus, we followed the ABC-Max approach and for each peak, we kept the E-P ABC link with the highest ABC score. While we did not observe overlap for GABAergic neurons we identified 12 dysregulated peaks which overlapped both an E-P link and a SCZ GWAS SNP thus linking epigenetic changes with genetic signal. We then visualised this cell-specific regulation using IGV genome viewer (<https://www.igv.org/>) to show the E-P link from the ATAC-seq region containing a GWAS SNP to the transcription start site of the predicted target gene. An example of this for *ZNF281* is shown in Figure 4.16.

ZNF281 is highly expressed in oligodendrocytes and is involved in the DNA damage response, a process that has been found to be perturbed in postmortem tissue isolated from schizophrenia cases (Pieraccioli et al., 2016; Shishido et al., 2023).

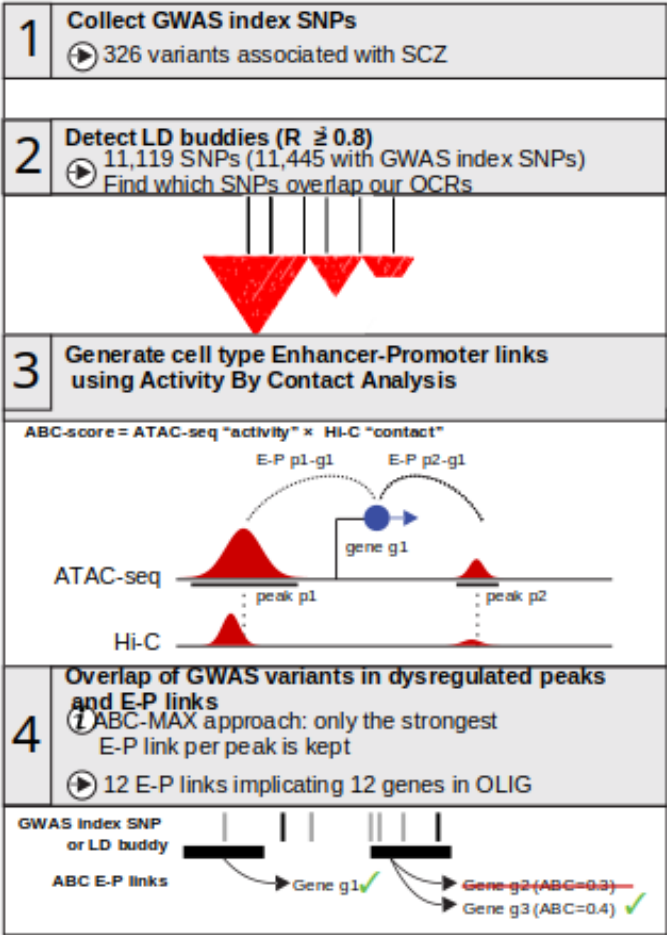


Figure 4.15: Schematic outlining the strategy for linking risk variants associated with SCZ to differentially accessible OCRs and their respective causal genes in a cell-specific manner.

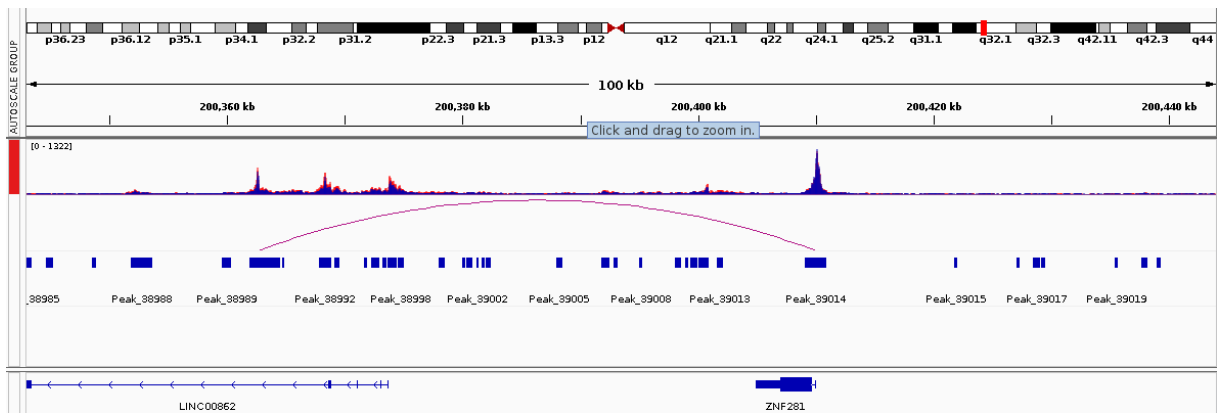


Figure 4.16: Genome browser visualisation of E-P link and ATAC-seq data for *ZNF281*. The top track represents genomic location followed by ATAC-seq data for cases (red) and controls (blue), the enhancer link between the ATAC-seq peak and *ZNF281* is then illustrated below.

4.5.7 Differential Gene Expression Analysis

Three hundred and ninety-two samples remained following processing of the RNA-seq counts and were used to identify DEGs between SCZ cases and controls. The combined influence of cell type and diagnosis status explained 66.4% in the RNA-seq data. In a stepwise Bayesian information criterion approach (Methods), the GC content (represented by two complementary metrics) emerged as the sole technical covariate for the RNA-seq contributing to 6.3% of the variance. The number of unique genes across cell types that were significant at BH-corrected P-value < 0.05 was 2,083. Oligodendrocytes were shown to have the most SCZ-associated changes in gene expression ($n=1,564$) followed by glutamatergic neurons ($n=266$) (Figure 4.17). Overall, non-neuronal cell types were observed to contain higher numbers of DEGs than neuronal cell types. (Figure 4.17). We also explored the overlap of our DEGs between cell-types. Overall, there appeared to little overlap of DEGs between cell-types with only one shared across all four cell-types and the majority of DEGs appearing to have cell-type specific dysregulation (Figure 4.18). A full table of DEGs is available in Supplementary Table 4.4 (https://drive.google.com/drive/folders/1lR3Ccix_ccf4pWhXY6Ajqt0LdI0Qym86?usp=sharing).

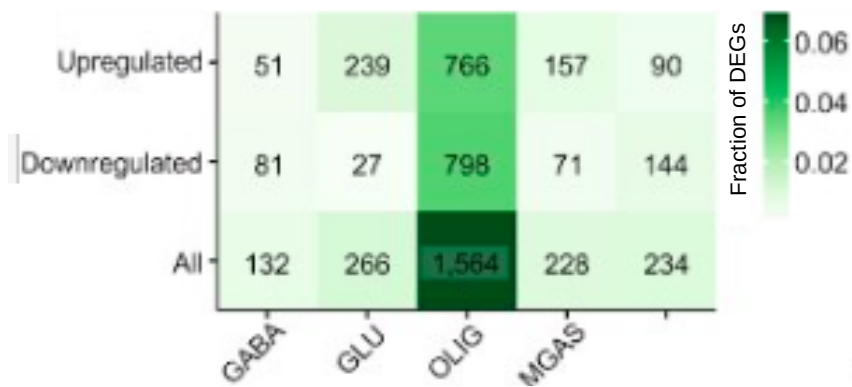


Figure 4.17: Heatmap summarising the number of differentially expressed genes (DEGs) per cell type. The total number of DEGs per cell-type (RNA merged), the number of up-regulated genes per cell-type (RNA Up) and the number of down-regulated genes per cell-type (RNA Down) are shown. Fraction of DEGs refers to how many of the total gene were differentially expressed, darker shading indicates more of the total number of genes were differentially expressed between cases and controls.

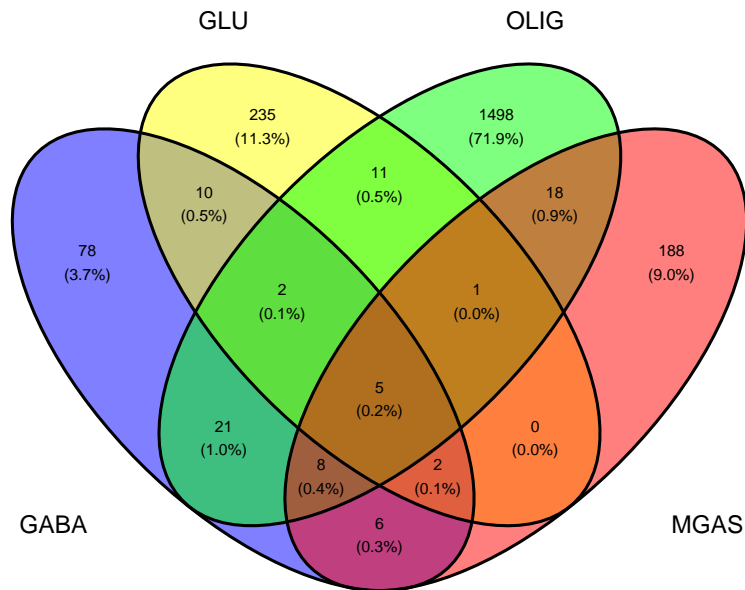


Figure 4.18: Venn diagram showing overlap of DEGs between cell types.

4.5.8 Gene-set Enrichment Analysis of Differentially Expressed Genes

We employed GSEA to identify the biological functions potentially impacted by the DEGs identified by our RNA-seq (Figure 4.19). Significant terms in the neuronal cell-types included synaptic terms such as “Regulation of Synaptic Vesicle Cycle”, “Regulation of Postsynaptic Membrane Potential” and “Glutamate Binding Activation of AMPA Receptors and Synaptic Plasticity”. MGAS DEGs implicated “Axon Ensheathment in Central Nervous System” which is consistent with recent research which shows that microglia are necessary for myelin sheath maintenance (McNamara et al., 2023). Terms related to the endoplasmic reticulum (ER) were also significant for the combined DEG set and the glutamatergic DEGs. Oligodendrocyte DEGs implicated processes including “Cholesterol Biosynthesis” and “Sterol Biosynthesis”. Oligodendrocytes require cholesterol for myelin production and lipids have been investigated as biomarkers for SCZ as they are thought to play a role in the cognitive symptoms associated with the disorder (Maas et al., 2020; Sun et al., 2021).

We also utilised a targeted strategy using the SynGO ontology resource to identify synaptic locations associated with our DEGs (Figure 4.20). The locations implicated by the GABAergic DEGs included the postsynaptic cytoskeleton and presynaptic synaptic vesi-

cle (SV) while glutamatergic DEGs implicated the postsynaptic specialisation membrane and the presynaptic ER (Figure 4.20 (a-b)).

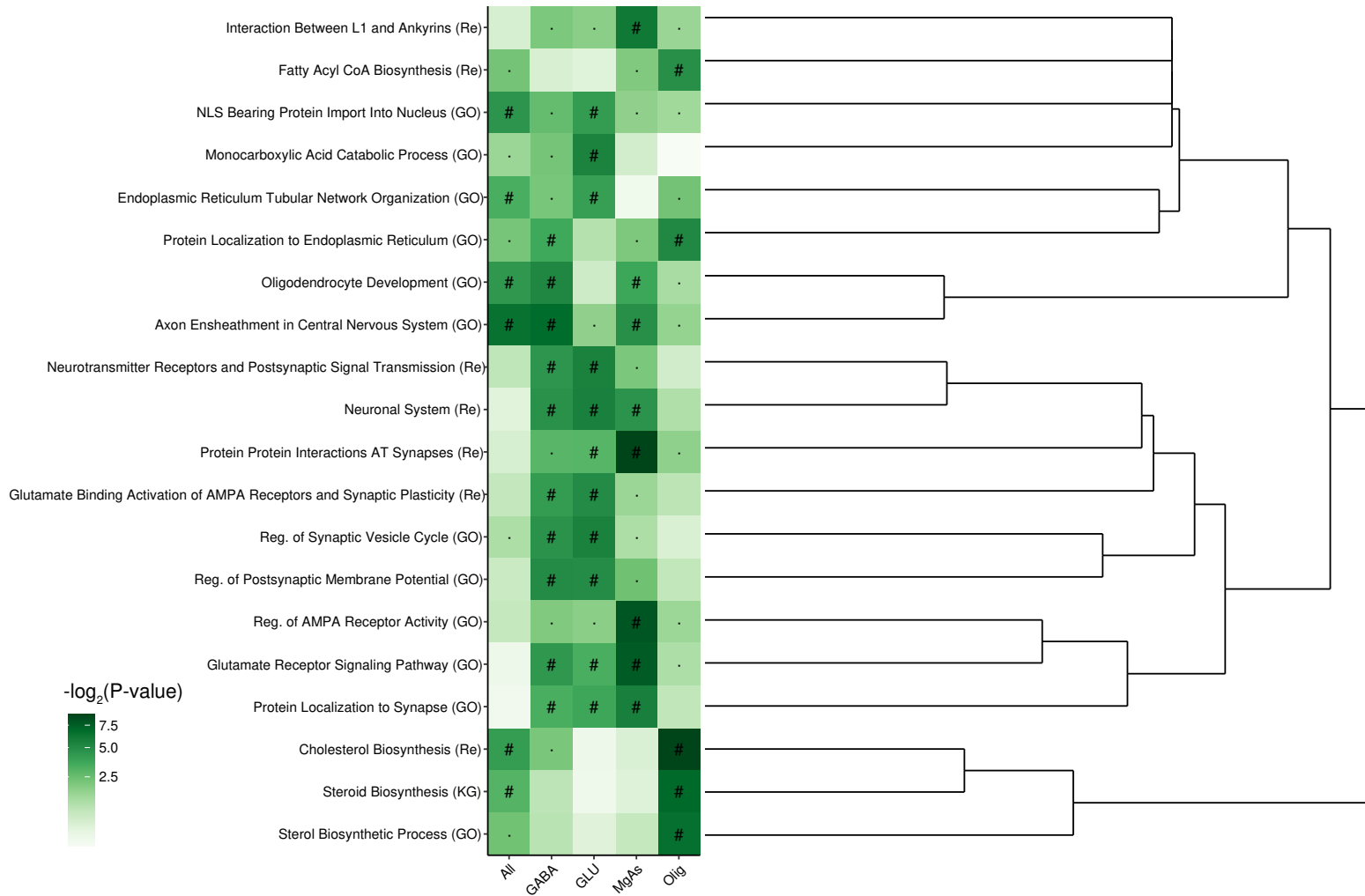


Figure 4.19: Heatmap visualisation of GSEA results for DEGs using MSigDB. #: Significance with BH-corrected P-value < 0.05, a dot indicates nominal significance, Bi: Biocarta, GO: Gene Ontology, Re: Reactome.

Figure 4.20: Sunburst plots depicting synaptic locations implicated by the (a) GABAergic DEGs and (b) glutamatergic DEGs. This plot features the synapse at the center, pre- and post-synaptic locations in the first ring, and child terms in subsequent rings. The color scheme in the legend indicates the number of genes within each term.

MAGMA gene-set analysis was used to identify whether our DEGs were enriched for genes containing common variants associated with SCZ, IQ and educational attainment (EA). The latter two were used as they are associated with SCZ and contain larger sample sizes for a more powerful analysis. However, none of these gene-sets were significantly enriched for genes associated with any of the aforementioned phenotypes (Table 4.1). The 12 independent tests were corrected for multiple testing by calculating BH-corrected P values (Q values).

Table 4.1: Results from MAGMA gene-set analysis of the DEGs per cell-type using GWAS data for SCZ, IQ and EA.

Cell Type	Phenotype	Beta	P-value	Q-value
GABAergic Neurons	SCZ	-0.046	0.634	0.841
Glutamatergic Neurons	SCZ	0.037	0.321	0.642
Oligodendrocytes	SCZ	0.036	0.141	0.423
Microglia/Astrocytes	SCZ	0.173	0.038	0.228
GABAergic Neurons	IQ	-0.097	0.771	0.841
Glutamatergic Neurons	IQ	-0.002	0.509	0.764
Oligodendrocytes	IQ	0.05	0.058	0.233
Microglia/Astrocytes	IQ	0.09	0.177	0.425
GABAergic Neurons	EA	-0.184	0.896	0.896
Glutamatergic Neurons	EA	0.003	0.487	0.764
Oligodendrocytes	EA	-0.023	0.742	0.841
Microglia/Astrocytes	EA	0.26	0.008	0.094

4.5.9 Differential Transcript Expression Analysis

To fully leverage the potential of our deeply sequenced transcriptome data, we used our 392 RNA-seq samples to carry out a differential transcript analysis. We first carried out a differential expression analysis using limma (Methods). We then wished to prioritise genes with differential transcript expression using the remaCor R package. This implements a random effect meta-analysis per gene using both the differential transcript analysis objects created with limma and a list of transcripts for a given gene to prioritise genes with differential transcript expression. Following correction, there were 3,666 unique genes that contained a transcript showing a disease association (BH corrected RE2C statistic <0.05). Again, oligodendrocytes show the highest number of associations

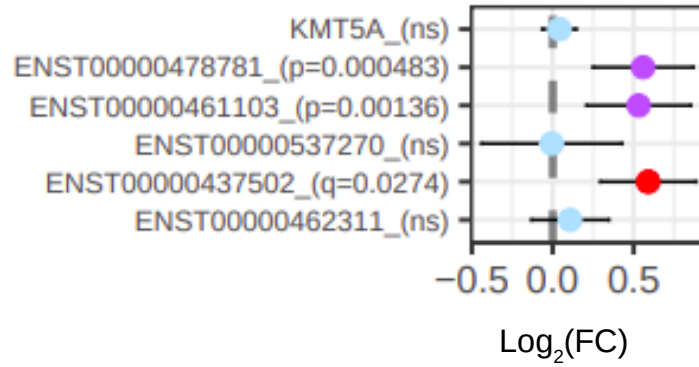
Table 4.2: Number of genes containing a significant transcript per cell-type (BH corrected RE2C P-value < 0.05)

Cell-type	Number of Genes
GABAergic Neurons	432
Glutamatergic Neurons	114
Oligodendrocytes	2270
Microglia/Astrocytes	1465

with SCZ ($n=2,270$) followed by the microglia/astrocyte mixture ($n=1,465$), which is consistent with the gene level analysis (Table 4.2). A full table of prioritised genes containing a DET is available in Supplementary Table 4.5 (https://drive.google.com/drive/folders/1lR3Ccix_ccf4pWhXY6Ajqt0LdI0Qym86?usp=sharing).

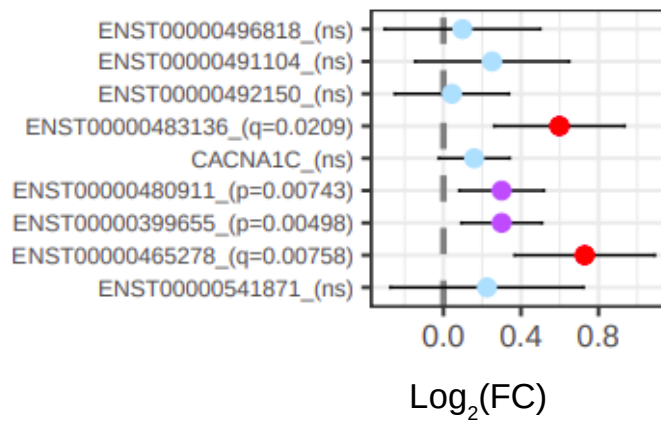
Analysis of transcript expression also identified a number of SCZ risk genes that did not reach significance in the gene-level analysis. *KMT5A*, a gene of interest in relation to SCZ, was not significant in the gene-level analysis but was implicated in the remaCor DET analysis in both oligodendrocytes (BH corrected RE2C P-value =0.024) and GABAergic (BH corrected RE2C P-value =0.012) neurons. Of particular interest is the alternative protein-coding transcript implicated in GABAergic neurons (ENST00000437502) which is not significant in oligodendrocytes and contains an exon not shared by the other alternative transcripts (Figure 4.21 (a)). Other non-protein coding isoforms of *KMT5A* have been shown to be of interest in SCZ though not in any particular cell type. In addition, *CACNA1C* in oligodendrocytes (BH corrected RE2C P-value =0.007) contains two significant transcripts (Figure 4.21(b)). This difference is only observed in the oligodendrocyte samples, highlighting the need for cell-type specific analysis of transcript expression.

ENSG00000183955 (*KMT5A*)



(a) Graphical illustration of transcript expression *KMT5A*

ENSG00000151067 (*CACNA1C*)



(b) Graphical illustration of transcript expression *CACNA1C*

Figure 4.21: Graphical illustration of transcript expression for (a) *KMT5A* and (b) *CACNA1C*. Gene-level expression is denoted by gene symbol, transcript level expression is denoted by Ensembl transcript ID. Red denotes BH-corrected significant P-value results while purple indicates nominal significance and blue indicates that the gene/transcript is not significant (ns).

4.5.10 Estimation of Cell-type Proportions using Cell-type Deconvolution

In order to increase our power to detect cell-type specific eQTLs, we decided to impute cell-type expression profiles for each of the FANS cell-types using bulk gene expression data from the CMC ($n=494$ controls, $n=354$ SCZ cases). Expression imputation requires both bulk expression data and the estimated proportion of each cell-type in the bulk expression data. We estimated these proportions using the cell-type deconvolution tool dTangle (Hunt et al., 2019) by utilising the FANS data from this study as a reference profile of the gene expression within each cell type. This resulted in the estimation of the proportion of each of the four FANS cell-types (GABAergic neurons, glutamatergic neurons, MGAS and oligodendrocytes) for each individual in the bulk expression data (Figure 4.22).

The bulk expression data and the estimated cell-type proportions were then supplied to the cell-type expression imputation tool bMIND (Wang et al., 2021). This resulted in the creation of four count matrices, one per cell-type. In order to confirm that the imputed expression profiles accurately represented the expression for a given cell-type, we compared the Spearman correlation coefficient of the mean expression value per gene between the FANS data and the imputed expression profiles (Figure 4.23). When comparing identical cell-types, this ranged from 0.63 (MGAS) to 0.86 (oligodendrocytes and glutamatergic neurons). This indicates that our imputed profiles characterise the gene expression from purified populations to a large extent.

We also aimed to estimate the proportion of cell-subtypes present in the four cell-types profiled using FANS using a single-nucleus RNA-seq (snRNA-seq) dataset as a reference for each of the four FANS expression matrices and impute expression matrices for ten cell-types based on these proportions. However, unlike the above data we observed very low Spearman correlation coefficients when comparing our imputed gene expression to reference cell-type specific expression from (Ruzicka et al., 2022). This suggested that these count matrices were not accurate possibly due to the number of subtypes we were trying to impute, many of which had low sample size. As such, we did not move forward with this angle of the analysis and focused instead on the count matrices imputed from the bulk expression data.

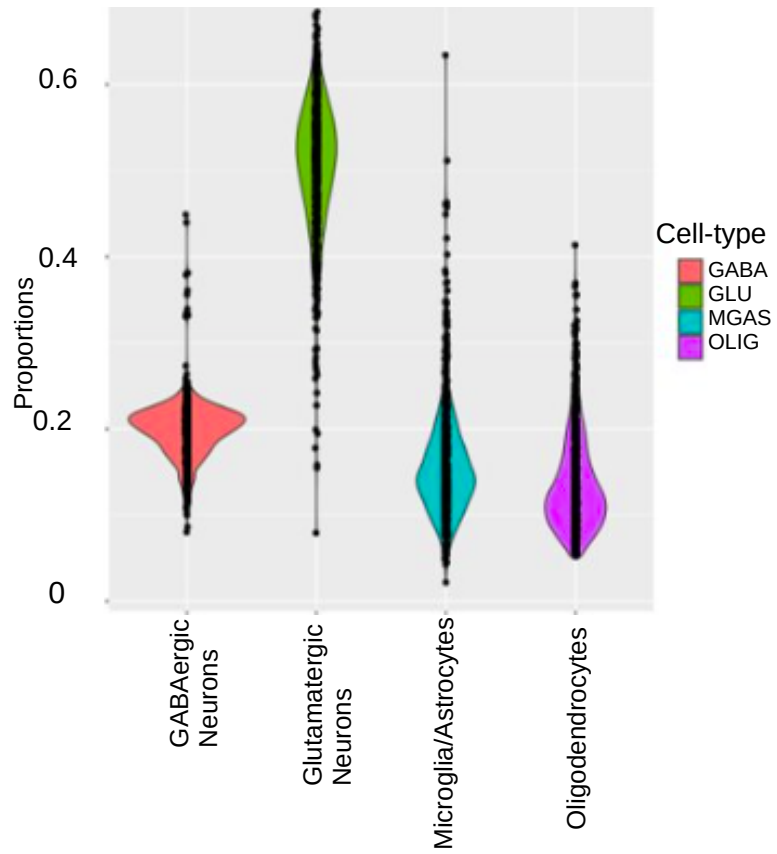


Figure 4.22: Violin plot showing the distribution of estimated proportions per cell-type.

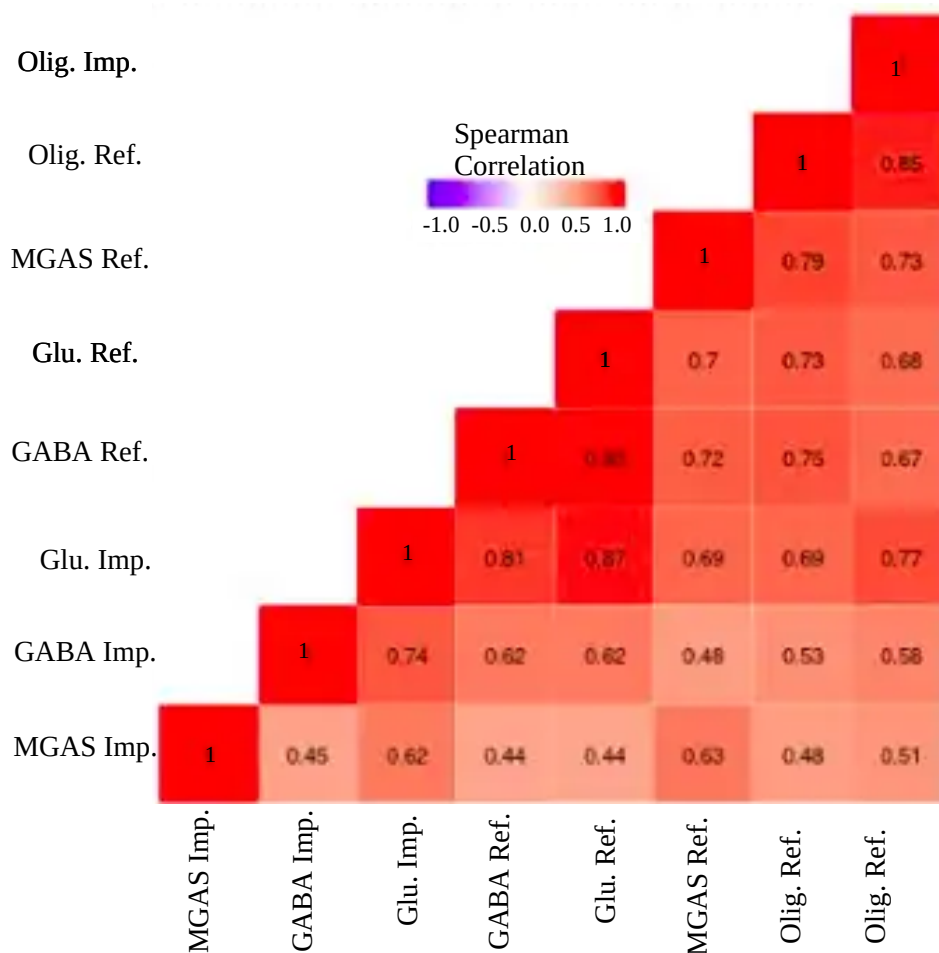


Figure 4.23: Spearman correlation coefficient between mean gene expression values of FANS (Ref.) data and imputed expression profiles (Imp.)

4.5.11 Detection of Cell-type Specific eQTLs

As our imputed expression profiles appeared to have sufficient accuracy based on the Spearman correlation coefficient, we decided to take advantage of the increased sample size to undertake an eQTL analysis. We used the mmQTL method (Zeng et al., 2022) to detect eQTLs per cell type as our data contained individuals from multiple ancestries. We used residualised matrices adjusted for technical factors and PEER factors to increase the power for the detection of eQTLs. Following two rounds of mul-

multiple testing correction at both the locus and genome-wide levels, we detected approximately 5,500-9,000 lead eQTLs per cell type (Figure 4.24), a number of which were found to be enriched for common variants associated with SCZ using LD score regression (Figure 4.25) (Bulik-Sullivan et al., 2015). A full table of genome-wide significant eQTLs is available in Supplementary Table 4.6 (https://drive.google.com/drive/folders/11R3Ccix_ccf4pWhXY6AjqtoLdI0Qym86?usp=sharing).

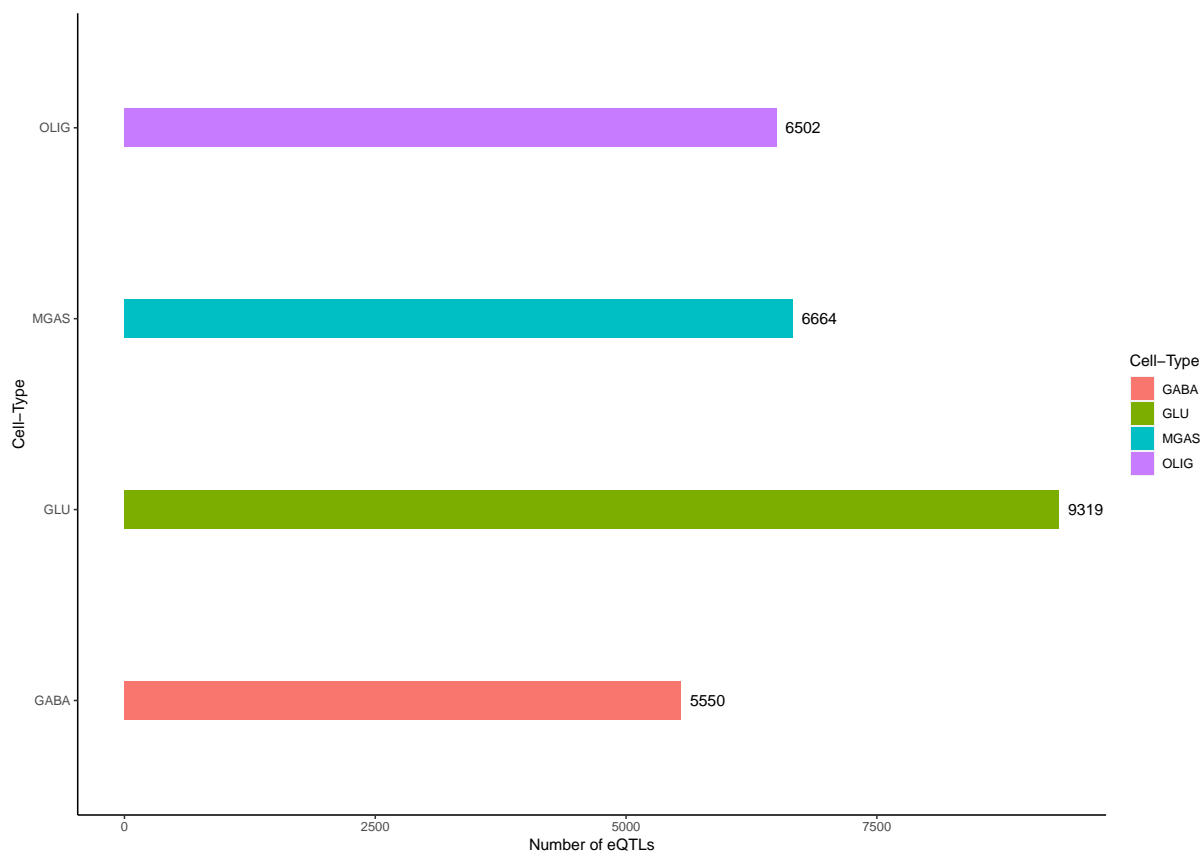


Figure 4.24: Bar chart showing the number of lead eQTLs per cell-type.

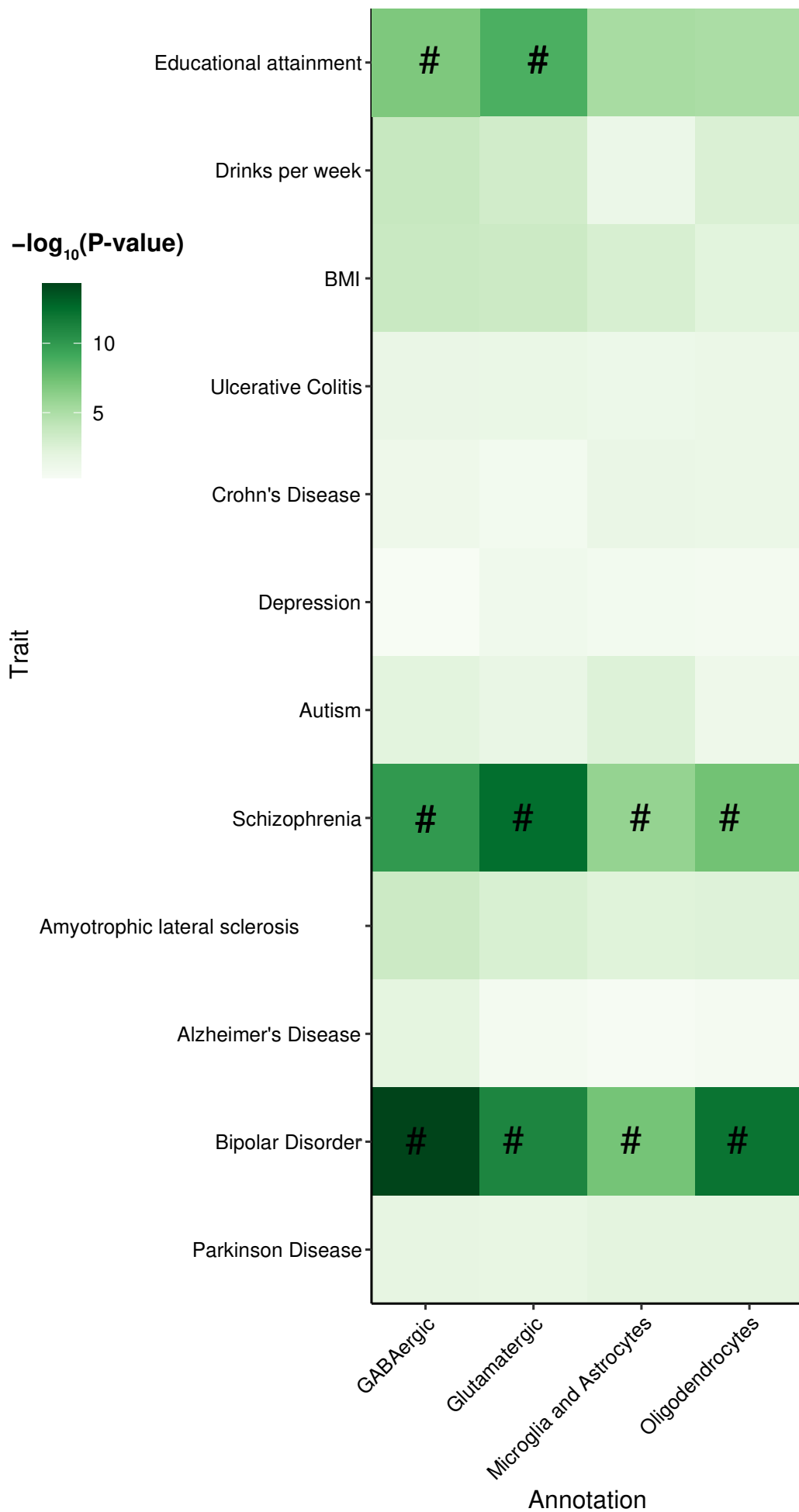


Figure 4.25: Heatmap for p-value enrichment of eQTLs in common variants associated with a variety of traits using LDSC. # denotes significant enrichment (BH-corrected P-value < 0.05).

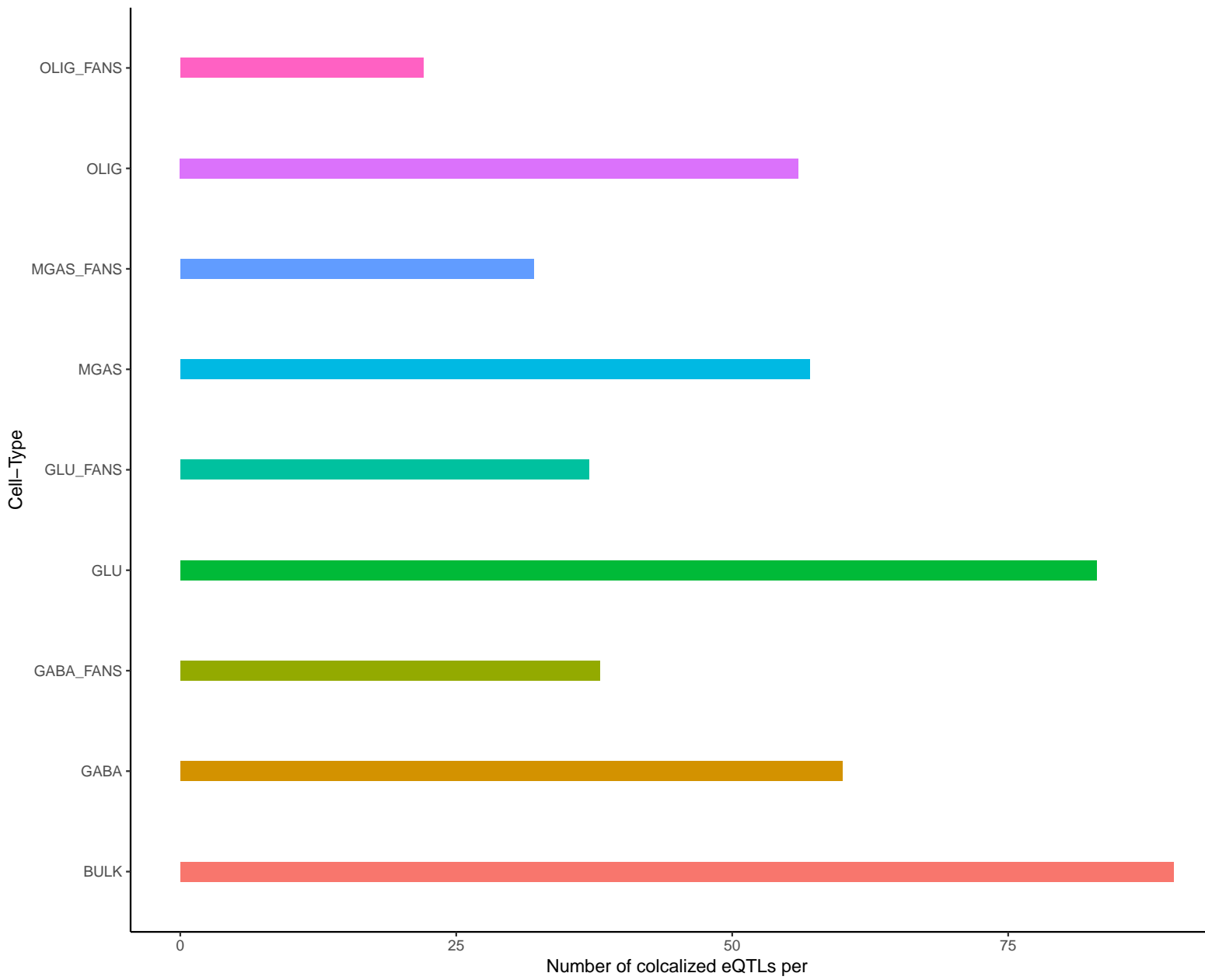


Figure 4.27: Bar chart showing the number of eQTLs which colocalized with SCZ risk loci in the imputed, FANS and bulk data.

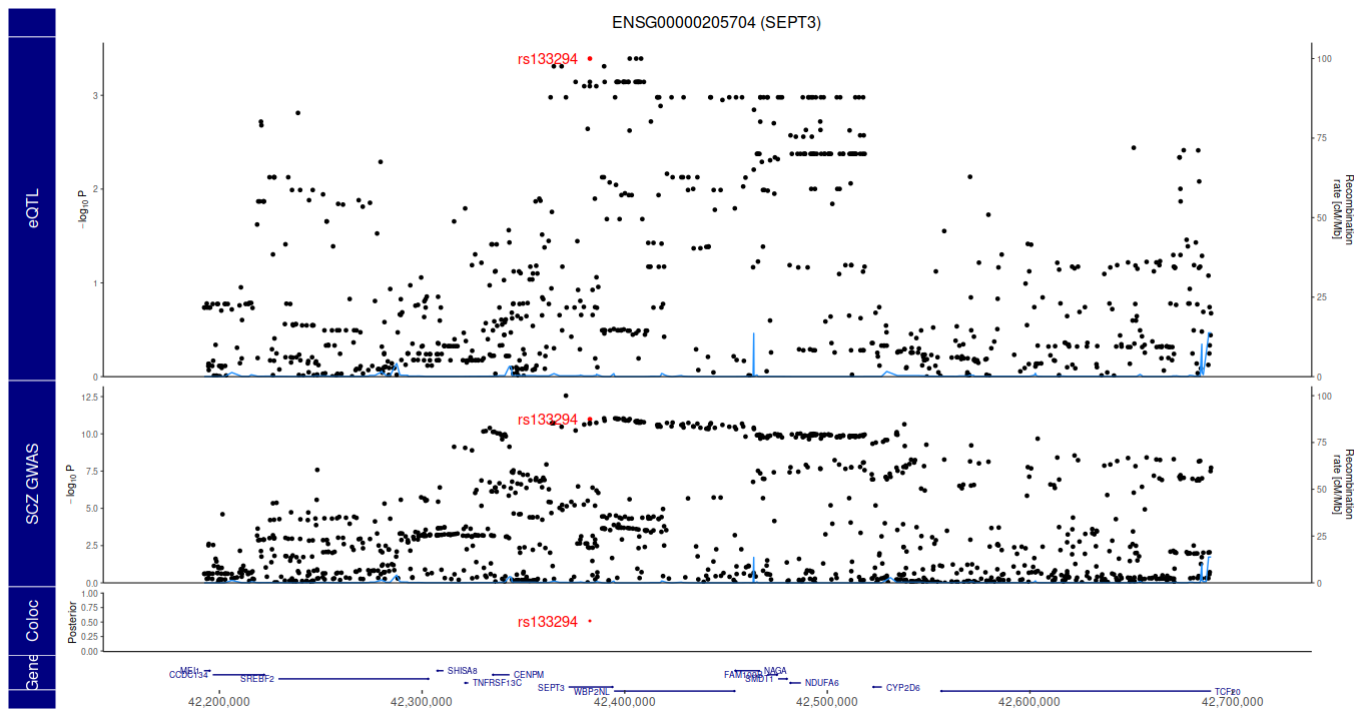


Figure 4.28: Colocalization of signals from eQTL and SCZ GWAS for SEPT3 in Microglia/Astrocytes. The top panel shows the eQTL Manhattan plot for this region, followed by the GWAS Manhattan plot and genes in this region.

4.5.12 Integration of ATAC-seq and RNA-seq Data

4.5.12.1 Canonical Correlation Analysis

We used canonical correlation analysis to integrate our RNA-seq and ATAC-seq data and assess their shared correlation structure. Both neuronal cell-types clustered closely together regardless of data type (Figure 4.29). However, while non-neuronal cell-types also clustered together, they did also separate by data type.

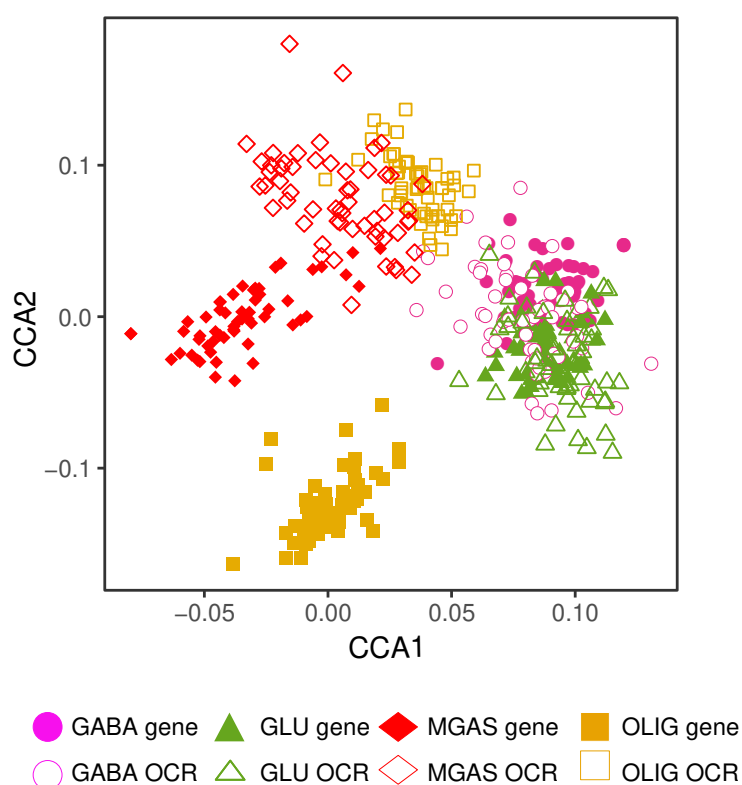


Figure 4.29: Plot showing the shared correlation structures using canonical correlation vectors (CCA1 vs CCA2) between ATAC-seq and RNA-seq profiles per cell-type. Plot points annotated as gene refers to RNA-seq data while points denoted as OCR refer to ATAC-seq data.

4.5.12.2 Integration of Multi-omic Data with SNFtool

We aimed to integrate our RNA-seq and ATAC-seq data using SNF, which first created a sample-sample similarity network per data-type before integrating these relationships

across the datasets. This process was carried out separately for each of the four cell-types. However, the clusters that we obtained did not appear to classify the samples into any particular subtype with the majority of clusters containing a mixture of cases and controls. A previous study (Ramaswami et al., 2020) that used this method observed their case samples divided into two distinct groups, one of which clustered strongly with control samples while the other consisted solely of case samples. As we did not observe this effect we concluded that this integration was not successful with our data.

4.6 Discussion

This study has involved the characterisation of cell-type specific dysregulation of both the transcriptome and chromatin accessibility landscape in human postmortem brain tissue derived from SCZ cases and controls. This was achieved through the exploration of 392 RNA-seq and 317 ATAC-seq profiles which encompassed four distinct cell-types (GABAergic neurons, glutamatergic neurons, microglia/astrocytes and oligodendrocytes). Distinct separation of cell-types was observed across both data types using t-SNE clustering which again suggests cell-specificity and underlines the need to undertake functional genomics studies at the cell-type level. ATAC-seq analysis resulted in the identification of 464,152 OCRs. This has expanded our knowledge of the chromatin accessibility landscape of the human brain as $\sim 20\%$ of the detected OCRs had not been identified in previous efforts by other large-scale studies (Bendl et al., 2022; Hauberg et al., 2020; Nott et al., 2019; Fullard et al., 2018; Bryois et al., 2018; Fullard et al., 2018; Corces et al., 2018; Roadmap Epigenomics Consortium et al., 2015). Further exploration of the overlap between the OCRs identified in this study and previous studies also showed a stronger overlap with OCRs from either brain tissue or a matching cell-type which suggests regional and cell-specific patterns of chromatin accessibility. General inferences were made regarding the brain epigenome by exploring the enrichment of these OCRs in regions that contain common variants associated with various GWAS traits. While we did not observe any enrichment for SCZ, we detected nominal significance for ALS and AD which is in contrast to a previous study which found that glutamatergic OCRs were enriched for common variants associated with SCZ (Hauberg et al., 2020). Our study however has a much larger sample size ($n=79$ ATAC-seq samples) compared to the previous study ($n=4$), which suggests that it may be more a representative sample. Additionally, the full set of OCRs used to conduct this analysis ranged in the hundreds of thousands per cell-type which indicates that they may just represent brain-related OCRs, which did not result in any significant enrichment due to lack of specificity.

Disease associated changes in chromatin accessibility were sizeable in the case of non-neuronal cell-types involving thousands of OCRs. The dysregulation identified was particularly extensive in oligodendrocytes, which have been associated with SCZ but to a lesser extent than neuronal cell-types. Some theories posit that they are responsible for the long gap between biological changes that occur in neurodevelopment and clinical presentation, which is usually in early adulthood (Fessel, 2022). This is hypothesised due to the fact they only double in population every two years and a long period of accumulation of aberrant cells is required in order to produce symptoms. The differentially accessible OCRs identified in oligodendrocytes were also observed to be enriched for common variants associated with SCZ as were differentially accessible OCRs in GABAergic neurons.

Inhibitory neuronal dysfunction is thought to be linked to the observed cognitive deficits which are perhaps caused by impaired neurotransmission (Birnbaum and Weinberger, 2017).

In contrast, glutamatergic neuron dysfunction in SCZ is well documented (Ruzicka et al., 2022; Gandal et al., 2018; Arion et al., 2017), however we did not observe substantial numbers of differentially accessible OCRs in excitatory neurons or any significant enrichment of common variants associated with SCZ. Much of the prior evidence however does come from transcriptomic studies rather than chromatin accessibility analysis. Chromatin accessibility profiling of glutamatergic neurons has previously been carried out although not in SCZ cases (Zhu et al., 2022). That study did identify an enrichment of common variants associated with SCZ in OCRs but using control data only so it does not provide a direct comparison to our study. Additionally, chromatin accessibility profiling of bulk data from SCZ cases and controls did not observe substantial differential accessibility and did not run cell enrichment analysis (Bryois et al., 2018). Thus, further studies will be required to fully elucidate the chromatin accessibility landscape of glutamatergic neurons in SCZ.

Disease associated changes in gene expression were largely consistent with chromatin accessibility in that oligodendrocytes were again the most perturbed cell-type. However glutamatergic neurons also showed substantial changes in contrast with the results from the ATAC-seq data. Changes were also observed in MGAS and GABAergic neurons, however disease associated changes were largely cell-specific with only 5 genes (0.2%) of all DEGs were shared across all cell-types, a trend which is echoed by a recent single-cell study of SCZ (Ruzicka et al., 2022). Additionally these genes did not overlap common variants associated with SCZ in any cell-type unlike the observed epigenomic changes.

By applying gene-set enrichment analysis we identified potential functional effects of the dysregulation observed in the aforementioned genes and OCRs. Processes such as the DNA fragment pathway and the regulation of the innate immune response were implicated by the GABAergic OCRs. DNA fragmentation is a hallmark of apoptosis, which is thought to be affected in SCZ resulting in dendritic spine loss (Pardiñas et al., 2018) and is supported by organoid research (Notaras et al., 2022). Additionally the GABAergic system is involved in the negative regulation of immune responses (Huang et al., 2022; Shan et al., 2023). Immune system dysfunction is well studied in SCZ with several lines of research showing elevated inflammation due to increased levels of C-reactive protein, cytokine levels and altered immune cell ratios (Wang et al., 2017; Gallego et al., 2021; Ermakov et al., 2022). Analysis of GABAergic DEGs identified more specific neuronal processes such as regulation of the synaptic vesicle cycle and postsynaptic membrane

potential. This was complemented by a targeted analysis using the SynGO ontology which implicated the presynaptic synaptic vesicle and the postsynaptic membrane as cellular locations impacted by the DEGs, which further supports the findings from GSEA. Affected processes in oligodendrocytes included terms involving glycosylation and cholesterol biosynthesis, which were identified using the OCRs and DEGs respectively. As oligodendrocytes are vital for both producing and maintaining the myelin sheath, of which cholesterol comprises $\sim 40\%$ of the lipid component (Kister and Kister, 2022), this suggests that some aspect of myelination may be aberrant in SCZ. Indeed, the possibility of myelin dysfunction was also observed using MGAS DEGs, which implicated axon ensheathment, supporting a recent study that showed that microglia are necessary for myelin sheath maintenance (McNamara et al., 2023). Across the cell-types a variety of pre and postsynaptic locations were revealed using the DEGs and OCRs. While many of the genes implicated were mapped to postsynaptic locations, which is consistent with previous literature, we also observe pre-synaptic locations (Trubetskoy et al., 2022). A presynaptic pathology was also suggested by the latest SCZ GWAS (Trubetskoy et al., 2022) using common variants while our findings from functional data supports this.

As our data is cell-specific we explored differential transcript expression which cannot be studied using current single-cell technologies. Current single-cell technologies capture only the beginning of a transcript meaning that isoform diversity cannot be profiled (Kharchenko, 2021). There are long read technologies that have been developed to address this issue but many are low throughput (Shi et al., 2023). Furthermore, there are no current SCZ datasets that can be used to study differential transcript expression in the context of SCZ. We observed extensive dysregulation across cell-types with oligodendrocytes again showing the most disease-associated changes confirming them as the most perturbed cell-type at both the epigenomic and transcriptomic levels. We then leveraged the novelty of cell-type specific transcript expression to identify DETs which were only dysregulated at transcript level and not dysregulated in all cell-types. Of note, *CACNA1C*, a gene involved in voltage-dependent gated calcium channels, has long been implicated in the etiology of SCZ (Nyegaard et al., 2010; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Trubetskoy et al., 2022) and showed cell-type specific transcript expression in oligodendrocytes. The effects of *CACNA1C* variants are thought to disrupt these calcium channels resulting in abnormal brain development and altered morphology in cortical regions (Zheng et al., 2016). Deletion of *CACNA1C* in neuronal cell-types does result in abnormal brain development in mice in addition to an anxiety-like phenotype and volumetric differences in a number of brain regions as it disturbs spontaneous calcium activity (Smedler et al., 2022). In oligodendrocytes however it is thought to promote the maturation of OPCs to mature oligodendrocytes but this effect is thought to have regional differences (Cheli et al., 2015; Pitman et al., 2020). *KMT5A*

was also observed to have differential transcript expression in GABAergic neurons and oligodendrocytes. It encodes a lysine methyltransferase and is involved in the DNA damage response and cell cycle regulation (Milite et al., 2016). These general themes were also identified through GSEA analysis of our GABAergic DEGs, however *KMT5A* was not a gene that was differentially expressed. This suggests that while both analyses converge on similar functions, we have uncovered through the DET analysis a contribution that would not have been detected if we have only focused on gene-level analysis. Moreover, one DET of *KMT5A* in GABAergic neurons was observed to contain an exon not present in any of the other transcripts, which may indicate a dysregulation of specific function that it carries out. Differential transcript expression of *KMT5A* is also supported by TWAS (Bhattacharya et al., 2022) but this analysis extends this by suggesting a cell-type specific context.

We also leveraged our multi-omic dataset to gain new insights by integrating it with other functional and genomic data. The integration of multiple data-types provides an opportunity to infer disease relevant changes that would not be detected through the analysis of a single dataset. To this end, we first assessed the shared correlation structure of our ATAC-seq and RNA-seq data using CCA. We found that broadly they clustered by cell-type rather than by data-type, which suggests high correlation between the gene expression and chromatin accessibility. This is in line with a previous study which profiled non-neuronal and neuronal cell-types (Dong et al., 2022a). However, we did observe a separation by data-type for our two non-neuronal cell-types. As the previous study did not have the same level of cell-type diversity a direct comparison is not possible. It is unclear whether this is a dataset specific effect or if lower correlation is observed at the cell subtype level.

Integration of our two data-types in order to classify our SCZ cases into subtypes based on molecular features was unsuccessful. While a previous study (Ramaswami et al., 2020) had success in defining subtypes of ASD using this approach, we were unable to separate our SCZ cases from controls in any cell-type. There are a number of reasons as to why this occurred. Firstly, the previous study identified subtypes in ASD, which has historically categorised cases based on symptomology (Grzadzinski et al., 2013). SCZ has no such clinical subtypes and it may not be possible to cluster cases as distinctly as in ASD. However, recent studies have employed machine-learning techniques to cluster SCZ cases using symptomology and brain-imaging (Castro-de Araujo et al., 2020; Chand et al., 2020) so it is possible that subtypes may exist but that using ATAC-seq and RNA-seq data alone is not informative. Furthermore, (Ramaswami et al., 2020) defined subtypes using bulk data in contrast to the cell-specific approach employed here, which may suggest that an overall picture of the molecular phenotype may be more successful for this type of classification.

The use of the activity-by contact model (Fulco et al., 2019) also enabled us to integrate information about the SCZ chromatin accessibility landscape with chromosomal conformation to detect E-P interactions. We inferred tens of thousands of these interactions per cell-type with over three quarters of the genes analysed having at least one E-P interaction. Determining these interactions is vital for understanding genetic regulation in the human brain and complements resources supplied by previous studies that have also detected E-P interactions in various brain cell types (Nott et al., 2019; Bendl et al., 2022; Kosoy et al., 2022). We extended this analysis to link these E-P interactions and SCZ-associated chromatin regions to genetic signal identified by the latest SCZ GWAS (Trubetskoy et al., 2022) for 12 oligodendrocyte E-P links. This allowed us to understand how risk variants can affect the epigenome through the dysregulation of an OCR and link this dysregulation to a target gene. One example, *ZNF281*, is highly expressed in oligodendrocytes and is involved in the DNA damage response (Pieraccioli et al., 2016). This response was also implicated by our transcript and GSEA analysis, which shows that our various analyses are converging on similar functional consequences.

By applying cell-type deconvolution and imputation techniques, we were able to compile a cell-type specific eQTL dataset. We leveraged our cell-type specific gene-expression profiles and previously published bulk expression data (Hoffman et al., 2019) to impute gene expression profiles for a much larger individual-level data, thus increasing our power to detect cell-specific eQTLs. This demonstrates the utility of leveraging older datasets to provide new insights into the genetic basis of the variation in gene expression. It also provides support for the use of imputation techniques as our estimated expression values correlate strongly with those observed in the cell-type specific profiles. While our dataset is limited to four cell-types, it provides a much greater sample size for eQTL detection compared to previous studies (Bryois et al., 2022), increasing the number of eQTLs discovered. As the eQTLs detected in our analysis showed significant overlap with common variants associated with SCZ, we undertook a colocalization analysis to determine whether they were important in the context of SCZ. We observed an increased number of colocalizations when compared to our original data or the unimputed bulk data with the added benefit of cell-type specificity. Cell-type specific eQTL analysis is indeed an important future direction in this area as the majority of colocalizations were cell-type specific. This cell-type specific effect is more pronounced when comparing combined neuronal versus non-neuronal cell-types. One cell-specific colocalization that we observed was *SEPTIN3* in MGAS. *SEPTIN3* is involved in both the cytoskeleton and the synapse (Marttinen et al., 2015) and while it was thought to be neuron specific, scRNAseq studies have shown it to be expressed in both astrocytes and microglia (Sjöstedt et al., 2020). More generally, the septin family has been linked to SCZ (Wang et al., 2018; Engmann et al.,

2011). Our eQTL analysis provides cell-type specific eQTLs detected in a case-control dataset unlike previous studies (Bryois et al., 2022).

Overall, this study provides an insight into cell-type specific effects observed at the epigenomic and transcriptomic level in SCZ. This study illustrates how the combination of new omics data with previously published data can increase sample size and power to identify disease associated changes. Furthermore, it links disease-associated epigenomic changes with SCZ GWAS findings providing a valuable way of interpreting the currently available data as a whole. We provide transcriptomic and epigenomic results that can be utilised for future downstream analysis in order to further understand and provide improvements for the the treatment of this disorder.

Chapter 5

Discussion

5.1 Summary of Main Findings

5.1.1 Chapter 2

This study utilised cell-type deconvolution to gain additional insights from published bulk gene expression data isolated from schizophrenia (SCZ) cases from the CommonMind Consortium (CMC). Cell-type deconvolution involved the use of reference data which contains information on cell-type specific expression to estimate the proportions of each cell-type in samples from a dataset where only bulk data was available. We estimated the proportions of ten different cell-types (medial ganglionic eminence derived inhibitory neurons (IN-MGE), caudal ganglionic eminence derived inhibitory neurons (IN-CGE), oligodendrocyte precursor cells (OPCs), astrocytes, oligodendrocytes, microglia, endothelial cells, pericytes and vascular smooth muscle cells (VSMC)) for 848 samples ($n=494$ controls, $n=354$ SCZ cases). We observed a marked reduction in neuronal proportion estimates compared to previous studies, which we can possibly attribute to the reference data that was used. Differences in the cell-type proportions between cases and controls for excitatory neurons did mirror a previous deconvolution study of SCZ (Wang et al., 2018) suggesting that overall trends are still observed. We also explored whether these proportions were associated with a polygenic risk score (PRS) for SCZ. Here, we observed an association only for the microglia cell-type. Finally, we used these proportions to impute gene expression for each cell-type using the bulk expression data. Gene ontology analysis of these imputed subtypes was then used to explore which pathways were implicated by the sets of differentially expressed genes (DEGs) between SCZ and controls. Biological processes enriched for neuronal DEGs included neuronal development, neuron projection and NCAM1 interactions. Non-neuronal cell-types implicated more general terms including signal transduction and various cellular locations including the cytoplasm and the extracellular matrix. Targeted analysis of the neuronal DEGs using the SynGO ontology resource was used to implicate potential synaptic locations such as the postsynaptic membrane and endosome and the presynaptic vesicle with different locations implicated between cell-types.

5.1.2 Chapter 3

This study investigated the genes under the control of *MEF2C*, which is a gene that has been implicated in a number of neurodevelopmental disorders including SCZ and autism spectrum disorder (ASD). *MEF2C* is a member of a family of transcription factors that are involved in transcriptional regulation throughout embryogenesis and continuing through adulthood (Flavell et al., 2008). Furthermore, it is highly expressed in the cortex of mouse models, which further suggests a role in nervous system development (Assali et al., 2019). We took gene-sets generated from a differential gene expression analysis of two *MEF2C* knockout mouse models (Harrington et al., 2016, 2020) and investigated them to determine which cell-types they were likely to affect. This was achieved by integrating them with snRNA-seq data from human postmortem brain tissue. This analysis primarily implicated neuronal cell-types for both gene-sets, which is consistent with previous studies that used mouse snRNAseq data for this purpose (Cosgrove et al., 2021; Fahey et al., 2023). However, it also implicated oligodendrocyte precursor cells (OPCs) suggesting the importance of using human brain data for these types of analyses. We also observed an enrichment of our *MEF2C* gene-sets in cell-type specific differentially expressed genes (DEGs) from SCZ cases and controls. We observed this enrichment in a number of cell-types including microglia, astrocytes and endothelial cells in addition to neuronal cell-types. Finally, we assessed which cell-types were implicated when considering the coexpression of the genes in each gene-set. We observed a higher coexpression fold enrichment of these gene-sets in excitatory neurons while we also observed a change in fold enrichment when considering different gestational timepoints.

5.1.3 Chapter 4

My final study involved the exploration of the transcriptomic and chromatin accessibility landscapes in SCZ. We used fluorescence activated nuclei sorting to isolate four cell types (GABAergic neurons, glutamatergic neurons, microglia/astrocytes and oligodendrocytes) from SCZ cases and controls ($n=50$ SCZ cases, $n=50$ controls). We then profiled the chromatin accessibility ($n=400$ samples, (4x100 per cell-type)) and gene expression ($n=400$ samples, (4x100 per cell-type)) in these cells to characterise their dysregulation in SCZ at the cell-type level. We expanded the known repertoire of open chromatin regions (OCRs) in the human brain with $\sim 20\%$ of detected OCRs appearing to be novel. The majority of detected OCRs were also found to be present in non-coding regions of the genome. We explored disease-associated changes at the transcriptomic and chromatin accessibility levels. We observed the most extensive chromatin accessibility changes in oligodendrocytes followed by microglia/astrocytes. Gene-set enrichment (GSEA) analysis of these OCRs implicated processes related to glycosylation and apoptosis. We also observed that common variants associated with SCZ were enriched in disease-associated GABAergic

and oligodendrocyte OCRs. Disease-associated changes in gene-expression were observed across cell-types with the most extensive changes in oligodendrocytes. However the majority of DEGs were cell-type specific. Biological processes enriched for our neuronal DEG sets included regulation of the synaptic vesicle cycle and synaptic plasticity. The non-neuronal cell-types implicated axon ensheathment and sterol biosynthesis, which indicate dysregulation related to the myelin sheath. We aimed to exploit the transcript-level information contained in our dataset as current single-cell studies do not have this advantage. We detected cell-type specific differentially expressed transcripts (DETs) for genes where the gene itself did not reach statistical significance. These results included *CACNA1C*, a well known SCZ risk gene, and *KMT5A* whose isoforms has been implicated in an isoT-WAS of SCZ (Bhattacharya et al., 2022).

We also integrated our multi-omic dataset with genetic and three-dimensional data to further explore deregulation in SCZ at the functional level. We predicted thousands of enhancer-promoter (E-P) interactions per cell-type by integrating our ATAC-seq data with Hi-C and ChIP-seq data. These interactions were then leveraged to link disease-associated epigenetic change with SCZ GWAS signal by identifying 12 OCRs that contain a SCZ GWAS SNP and overlap with our E-P interactions using the ABC-MAX approach. We also utilised our cell-type specific gene expression data to impute cell-type specific gene-expression in a bulk RNA-seq cohort. We did so by using our cell-type specific profiles to estimate cell-type proportions in the bulk data using our which were then used to impute gene expression. We showed that our imputed gene expression recapitulated cell-type expression to a large extent using Spearman correlation coefficient. As our imputed data was shown to be representative of the expression in these cell-types, we utilised the advantage of our larger sample size ($n=848$) to conduct an eQTL analysis. This increased the number of eQTLs that colocalize with SCZ GWAS signal compared to the bulk data or our cell-specific data. We identified a number of cell-type specific colocalizations highlighting the benefit of considering cell-type in QTL detection.

Overall, this study explored the transcriptomic and chromatin accessibility landscapes in SCZ and provides an extensive resource for future work once this data is shared with the research community. It also provides derived data-types that can be used with the aforementioned data to provide new insights into the neurobiology of SCZ at the cell-type specific level.

5.2 The Use of Cell-type Deconvolution to Gain New Insights from Existing Bulk Expression Data

Studies one and three of this thesis use cell-type deconvolution to estimate cell-type proportions in bulk expression data. The main advantage of cell-type deconvolution is that it can be used to leverage cell-type specific information from bulk tissue data. This is especially advantageous today as many bulk tissue resources have been built up making their sample sizes substantially larger than burgeoning single-cell studies. This is especially true in the cases of neuropsychiatric disorders which rely on postmortem brain tissue data which is much harder to obtain than in other research areas such as cancer research.

Beyond leveraging cell-type proportions to gain new insights from bulk data, estimated proportions can be used as a covariate in large-scale functional studies to correct for differences in proportion that may be due to technical factors such as cohort or tissue dissection. By not accounting for this, some studies may identify gene expression changes that are due to a difference in proportion rather than a true disease-associated change. Indeed, a study of sex differences in the SCZ transcriptome in the human brain included cell-type proportion as a covariate in their DEG analysis after identifying part of the expression variance explained by sex is affected by cell-type composition and was confounding the analysis (Hoffman et al., 2022). Chapter Two of this thesis aimed to utilise cell-type deconvolution for a different purpose, specifically using the proportions to identify differences between cases and controls, associate these changes with PRS and use the estimated proportions to impute cell-type specific gene expression. While we achieved these goals, we were concerned that our average proportions were not consistent with previous studies of brain tissue (Wang et al., 2018) and that this not only affected analysis using the cell-type proportions but the subsequent imputation. We did observe results that suggested we found trends consistent with previous studies (Wang et al., 2018) such as a difference between cases and controls for proportions of excitatory neurons and an enrichment of our imputation-derived neuronal DEGs being implicated in relevant processes. However, due to the limitations and discrepancies in this analysis, these results need further independent support. Furthermore, snRNA-seq analysis of SCZ (Ruzicka et al., 2022) identified differences in cell-type proportions between samples but did not attribute this to disease status and removed samples showing extensive differences in proportion from their DGE analysis. This suggests that proportion differences between cases and controls observed in the PsychENCODE analysis (Wang et al., 2018) may not be due to disease status but possible technical issues with the sample. Deconvolution may be suitable for estimating proportions but not for assessing disease-associated changes in cell-type proportion.

Chapter Four, which also used deconvolution and imputation methods, did however im-

prove on my initial study in a number of ways. We did observe average proportions consistent with previous published studies (Wang et al., 2018; Ruzicka et al., 2022; Hoffman et al., 2022). We also sought to ensure that our imputed expression matrices were accurate by identifying if they were consistent with the expression patterns you would expect in those cell-types. Using the Spearman correlation coefficient, we found that our imputed expression profiles correlated strongly with cell-type specific expression profiles. As such we were confident that our imputed data was recapitulating the expression of those cell-types to a large extent, enabling us to proceed with an eQTL analysis using a much larger sample size.

There are however a number of caveats to the use of cell-type deconvolution and there have been many concerns about accuracy of these tools. Studies have found that proportion estimates can vary widely based on the deconvolution tool that is used and other factors such as normalisation, choice of reference data and preprocessing can also have an impact (Avila Cobos et al., 2020). We aimed to overcome some of these issues by using dTangle (Hunt et al., 2019), the deconvolution tool that has been found to be most accurate for data derived from brain tissue samples (Sutton et al., 2022). Yet our proportions for Chapter Two were still inconsistent with published work as stated. As such, we thought that the reference data may have been affecting the analysis. This was confirmed in Chapter Four where we used the same reference data for an analysis that had to be abandoned as the imputed data did not reflect cell-type specific expression. Furthermore, a recent study has shown that even samples derived from neighbouring dissections in the same individual can result in vastly different proportion estimates and using different omic data such as methylation data can result in very different estimates to proportions generated using gene expression data (Toker et al., 2023). Additionally, while not an aim our our study, the use of cell-type proportions to account for confounding in gene expression studies may not be sufficient (Toker et al., 2023).

Overall, it is clear that studies using cell-type deconvolution and expression imputation must be undertaken with caution. While it is clear from Chapter Four of this thesis that they can be used to great effect, it is important that care is taken to fully ensure that the estimated proportions are accurate and the imputed expression correlates strongly with non-imputed cell-type specific expression data. With this in mind, if I were to design my first study again, I would determine marker genes for cell-type deconvolution using multiple studies similar to the strategy employed by (Sutton et al., 2022) in their benchmarking study to ensure that they are as representative as possible. I would also compare them to existing estimations to ensure consistency and check the correlation of any imputed expression with single-nucleus or cell-type specific profiles as shown in Chapter Four.

5.3 The Use of Multi-omic Datasets Derived from Postmortem Brain Tissue in the Study of Neuropsychiatric Disease

Each of my studies to some extent aimed to use multiple types of omic data to study neuropsychiatric disorders, primarily SCZ. Chapter Two aimed to use both snRNAseq and bulk expression data to study cell-type proportions as described above. While Chapter Three integrated a gene-set derived from a *MEF2C* mouse model with human snRNAseq data in order to identify cell-types enriched for those genes. This aimed to implicate human cell-types affected by genes under the control of *MEF2C*, a transcription factor implicated in neurodevelopmental disorders including SCZ (Zhang and Zhao, 2022b; Mitchell et al., 2018). We contrasted our results to two previous studies (Fahey et al., 2023; Cosgrove et al., 2021) that used murine expression data to implicate cell-types. We found similar results with neuronal cell-types primarily implicated, however OPCs were also implicated highlighting an advantage to using other datasets. Both studies one and two focused on gene expression datasets both bulk and single-nucleus, however my third study incorporated many omic datasets to explore SCZ etiology.

Chapter Four moved beyond gene expression alone and generated a primary multi-omic dataset of transcriptomic data and chromatin accessibility data. This enabled us to identify disease-associated transcriptomic and chromatin accessibility changes in SCZ. Among our findings were that common variants associated with schizophrenia were enriched in differentially accessible OCRs from GABAergic neurons and oligodendrocytes. However, we did not observe this enrichment using DEGs from those cell-types, this underscores the importance of multi-omic datasets because if we had only focused on gene expression, like the majority of previous studies, we would not have been able to find this enrichment. GSEA of both disease-associated features implicated processes such as glycosylation, apoptosis and axon ensheathment, the latter of which was only implicated by DEGs and not OCRs. Following initial analysis of the primary dataset, we integrated our dataset with other omics data to further explore functional changes in SCZ, which would not have been possible if we had relied on our dataset alone. In the case of the chromatin accessibility data, we integrated our ATAC-seq data with cell-specific Hi-C data and ChIP-seq data to predict E-P interactions. We also went a step further by isolating E-P interactions that overlapped with OCRs that contained SCZ GWAS SNPs thus enabling us to link genetic signal with disease-associated epigenetic signal and predict target genes through the E-P interaction. Findings such as these are only made possible through the generation and integration of multi-omic data. Detection of E-P interactions alone using the ABC model (Fulco et al., 2019; Nasser et al., 2021) as described in Chapter Four in

particular requires three types of omic data, including Hi-C which is often not as readily available though has been used to great effect in studies of other brain related disorders including Alzheimer’s disease (AD) (Nott et al., 2019; Bendl et al., 2022). Despite this, it is clear that new discoveries can be made through the use of multi-omic data, which is especially relevant now as many studies are converging on similar pathways and processes such as synaptic plasticity or ion channels (Trubetskoy et al., 2022; Ruzicka et al., 2022; Fromer et al., 2016).

A large majority of historical studies only focus on one data-type which is usually gene expression. However multi-omic studies are becoming more popular in recent years due to a number of reasons. Firstly, there is a growing understanding that different disease-specific changes can be identified in different omics data. For example, ATAC-Seq that profiles OCRs which are actively transcribed regions is not fully concordant with gene expression. A recent study in AD (Bendl et al., 2022) found that while gene expression and OCRs were highly correlated, they did not entirely match thus implying a need to study both gene expression and chromatin accessibility to fully characterise disease-associated changes. Secondly, many studies of both gene expression and genetic data have found a substantial amount of noncoding variation associated with neuropsychiatric disorders (Trubetskoy et al., 2022; Pardiñas et al., 2018; Gandal et al., 2018; Roussos et al., 2014). This suggests that studies of chromatin conformation and the epigenome must be integrated into this area of research in order to understand the change in genetic regulation and how it impacts coding portions of the genome. Thirdly, newer technologies that have been developed in recent years such as ATAC-seq (Buenrostro et al., 2013) and Hi-C (Belton et al., 2012) are becoming widely used as they become less expensive resulting in more studies profiling more than one molecular feature.

More recent multi-omic studies of neuropsychiatric disorders have provided new insights into the neurobiology of these conditions. The use of ChIP-seq data in a study researching chromatin domain alterations in SCZ and bipolar disorder (BD) (Girdhar et al., 2022) allowed for profiling of the histone modification landscapes of both conditions and was further enhanced by the use of Hi-C data to show histone peaks constrained by local chromosomal conformation, which goes beyond the traditional “peak by peak” based analysis. In doing so they showed altered organisation of the neuronal genome in SCZ and BD in conjunction with cis regulatory domains enriched for SCZ heritability. Moving beyond just isolated gene expression, or in this case histone modification changes, allows us to observe the consequence of a combination of genetic changes that will enable us to fill in the gaps in our knowledge between initial changes and onset of disease. Indeed, the initial changes in SCZ are believed to occur in neurodevelopment and this has been a prevailing hypothesis for many years (Weinberger, 1987; Owen et al., 2011; Owen and

O'Donovan, 2017). It has been shown that SCZ-associated variants overlap with variants from other neurodevelopmental conditions such as ASD (Rees et al., 2021). Findings from a recent multi-omic study of SCZ that integrated chromatin accessibility data from both the adult and fetal brain identified a trans-regulatory domain linked to glutamatergic neurons in early fetal development (Girdhar et al., 2023) and builds on previous studies (Bryois et al., 2018; Girdhar et al., 2022) through the use of these dual data-sets by mapping this change to a particular cell-type. Future multi-omic studies combining more data-types and timepoints will aid in truly filling the gaps in our knowledge regarding SCZ from neurodevelopment to adulthood.

Despite the major move towards incorporating any types of omic data in the study of neuropsychiatric disorders, there remains many limitation. Most obvious of all is sample size. Early GWAS studies of SCZ (O'Donovan et al., 2008; International Schizophrenia Consortium et al., 2009; Purcell et al., 2014) were hampered by this but have now have combined sample sizes of over 300,000. Other disorders such as BD still have much lower sample sizes but these too are increasing (Mullins et al., 2021). However functional studies lag behind this to a large extent with the largest studies containing samples of under 2,000 samples (Girdhar et al., 2023; Dong et al., 2022a; Hoffman et al., 2019; Wang et al., 2018) and cohorts of this size are the usually the exception rather than the rule. Yet, it is incredibly difficult to envisage and extremely unlikely that these studies will ever match the size of genetic studies due to the difficulty of obtaining postmortem brain tissue. They also share a major limitation with GWAS in that many studies lack genetic diversity and are predominantly comprised of white, European samples. This is changing however, with the latest GWAS containing individuals from African-American, Latinx and East Asian ancestry and other genetic studies containing more diverse populations (Trubetskoy et al., 2022; Liu et al., 2023). Additionally, the CommonMind Consortium also contains functional data from African-American and Latinx (Hoffman et al., 2019) as does the cohort used in Chapter Four of this thesis. Sample size issues have even impacted this thesis, with Chapter Four using deconvolution and expression imputation to enable us to conduct cell-type specific eQTL analysis in a larger cohort. Ideally, this would be conducted on a larger cell-type specific dataset rather than an imputed one, yet we must find ways to work around this limitation successfully for the time being.

A further limitation of multi-omic studies using postmortem brain tissue data is the postmortem tissue itself. Functional data is susceptible to batch effects and technical issues such as degradation and postmortem interval among others (Kangas et al., 2022). In addition, postmortem tissue will most likely only represent the effects of late-stage disease. Indeed, a recent study comparing gene expression between living and postmortem brain tissue found that expression levels in almost 80% of the genes analysed differed between

living and postmortem tissue from the PFC (Liharska et al., 2023). This raises questions as to the accuracy of findings from postmortem tissue samples. Yet this study also possesses limitations in that all of the living samples are from individuals with Parkinson's disease and contained no controls or samples from individuals with neuropsychiatric conditions, indicating the need for further work to ascertain the impact postmortem tissue may be having on these studies.

It is clear that multi-omic studies are extremely advantageous for probing the neurobiology of neuropsychiatric conditions. Incorporating more omic data in studies one and two of this thesis would be of great interest in future work as they only consider gene expression, which is only part of the picture in SCZ. In particular, Chapter Three which uses gene-sets derived from *MEF2C* knockout models could provide more insights into the impact of *MEF2C* dysregulation if chromatin accessibility or chromosomal conformation profiling were carried out in addition to RNA-seq. Furthermore, the integration of multi-omic data types using machine learning methods that was unfortunately unsuccessful in Chapter Four provide new ways of finding new subtypes of brain-related disorders as evidenced by (Ramaswami et al., 2020) in their study of ASD.

5.4 The Importance of Cell-type Specific Studies in Functional Genomics

An overall theme of this thesis is the exploration of cell-type specific changes in the context of SCZ. Chapters Two and Three aim to explore the contributions of different cell-types to SCZ and *MEF2C* knockouts respectively by integrating data derived from bulk tissue with single-nucleus expression data. Chapter Four identifies cell-type specific changes in SCZ using data derived from four specific cell-types to more robustly characterise these alterations. Functional data derived from bulk data has until recently been the main resource for studying functional changes in any disorder. Yet while it has been an invaluable resource for many reasons outlined in the introductory chapter to this thesis, it only presents the overall trends and more subtle changes can be missed. This is especially important in a disease context as these subtle changes may represent possible drug targets or other important insights into the etiology of a disorder. As such, it is important that we consider each cell-type individually when studying disorders such as SCZ. This can be achieved by integrating bulk data with cell-type specific information as illustrated in Chapters Two and Three or by generating cell-type specific datasets as shown in Chapter Four.

Using cell-type specific information to inform studies using bulk data has been largely

successful. Tools such as EWCE (Skene and Grant, 2016) and other methods have used cell-type markers or expression data to implicate cell-types such as excitatory neurons in SCZ (Gandal et al., 2018; Skene et al., 2018), which is consistent with cell-type specific studies (Arion et al., 2015; Ruzicka et al., 2022). Chapter Three of this thesis used this method to identify neuronal cell-types and OPCs as cells potentially affected by *MEF2C*. Cell-type deconvolution can also be used to gain cell-type specific insights from bulk data as evidenced by Chapters Two and Four. Cell-type information has also been leveraged to conduct cell-type informed QTL analysis in the human fetal brain (Wen et al., 2023). This analysis integrated snATAC-seq data with eGenes derived from bulk gene expression to map them to a cellular context which provides another way of leveraging cell-type information from bulk data.

Of course, the goal going forward should be to generate cell-type specific or single-cell/nucleus data when conducting functional genomics studies. Recent cell-type specific studies of SCZ have uncovered further differences between cell-types predominantly at the chromatin accessibility and gene expression level. Studies have included the development of a machine learning algorithm to identify transcribed enhancers in neuronal and non-neuronal cells using cell-specific transcriptomic and epigenomic data with neuronal-specific transcribed enhancers enriched for SCZ risk variants (Dong et al., 2022b). Chapter Four of this thesis also reports cell-type specific effects such as linking epigenetic dysregulation and GWAS to target genes in oligodendrocytes using ABC. This effect was only observed in this cell-type underscoring the need to profile multiple cell-types. Multiple cell-type specific studies (Dong et al., 2022b; Girdhar et al., 2022; Hauberg et al., 2020) have highlighted neuronal changes yet the non-neuronal dysregulation in Chapter Four of this thesis is clear.

Cell-type specific studies such as this are however limited by the number of cell-types they profile yet their sample sizes usually eclipse that of single-nucleus studies which can profile all cell-types present in a sample. There is currently only one snRNAseq of SCZ (Ruzicka et al., 2022), though there are many control-based single-nucleus atlases of the human brain from initial studies with very low cell numbers (Lake et al., 2017; Darmanis et al., 2015) to recent efforts that are extremely high-throughput (Siletti et al., 2023). One advantage of single-nucleus studies which use control data only is that we can profile cell-type specific gene expression across the lifespan as has been evidenced by a number of studies (Eze et al., 2021; Ament et al., 2023; Velmeshev et al., 2023; Zhu et al., 2023). snRNAseq of SCZ (Ruzicka et al., 2022) has found that excitatory neurons are the most affected cell-type which is consistent with prior work however it was also noted that nearly half of DEGs were dysregulated in a single cell type which suggests a high degree of cell-type specificity.

Additionally, studies of QTL detection in cell-type specific samples have been limited to date. The majority have occurred in sorted cell-type specific samples rather than single-nucleus samples but have still observed extensive cell-specificity (Zeng et al., 2023; Dong et al., 2022b) in both eQTL and caQTLs. A recent snRNA-seq study of the human brain using neurotypical samples (Bryois et al., 2022) detected cell-type specific eQTLs showing that almost half of detected eGenes show cell-type specific effects and eQTL colocalization with SCZ GWAS loci. This again shows the need for performing these analyses in cell-type specific samples. eQTL analysis conducted in Chapter Four of this thesis also observed cell-type specific effects and while our cell-types are more limited than (Bryois et al., 2022), we observed this colocalization using SCZ-derived data.

While single-cell studies are certainly the future of functional genomics, there are a number of inherent limitations currently. Single-cell/nucleus studies do not currently have the power to determine transcript level data. They cannot detect complex RNA isoforms as they are constrained to gene quantification proximal to the 3' end, resulting in the loss of full-length transcript isoform information (Arzalluz-Luque and Conesa, 2018). As such cell-type-specific transcript dysregulation remains largely unexplored. In Chapter Four of this thesis we exploited the advantage of our cell-specific dataset to uncover cell-type specific DETs that were not detected at gene-level. Among them was a transcript of *CACNA1C* which is noted SCZ risk gene. This together with the fact that DETs in SCZ are thought to have higher effect size (Gandal et al., 2018) suggests that despite the advantages of snRNAseq, we are still missing important information regarding alterations of the SCZ transcriptome. Long-read single-nucleus RNA isoform sequencing has the potential to further elucidate perturbed transcripts in SCZ in the future. However, it does suffer from low-read throughput (Gupta et al., 2018) but new technologies are being developed to overcome this (Shi et al., 2023). Studies of AD-associated changes in microglia have used long-read sequencing to great effect (Humphrey et al., 2023) identifying over 35,000 novel isoforms. Future studies such as this could be valuable in determining disease-associated transcripts at the cell-type specific level in SCZ.

Another limitation of current cell-type specific studies is that often only one brain region is analysed, usually the prefrontal cortex (PFC). There are no current single-nuclei analysis of multiple brain regions in SCZ yet there are a number of cell-specific studies which predominantly profiled neuron and non-neuronal cell across brain regions. One finding from a multi-omic study which profiled samples across 25 brain regions (Dong et al., 2022a) observed extensive changes including variation in alternative promoter-isoform usage and enhancer-promoter interactions across the brain regions. Other region specific changes have also been observed in multi-omic studies in AD (Bendl et al., 2022) and in control

only studies of chromatin accessibility (Hauberg et al., 2020) highlighting the need for future studies to profile multiple regions.

An important consideration in cell-type specific studies is the fact that changes observed in a particular cell-type may not be due to a direct disease-related change. Differential changes that are observed may be due to a disease-related change in another cell-type which in turn impacts other cells and regions in the brain. These types of changes may affect the analyses discussed in this thesis in a number of ways. For example, in Chapter Three the knockdown of *MEF2C* occurred in neuronal cells only but this change may affect the expression of genes in non-neuronal cell types. Similarly, in Chapter Four of this thesis we observed extensive dysregulation in oligodendrocytes despite glutamatergic neurons being the primary cell-type implicated in SCZ (Trubetskoy et al., 2022). This may suggest that these changes were secondary to changes in other cell-types. However, we did find that OCRs which were differentially accessible were enriched for common variants associated with SCZ which may suggest a direct effect in this case but some of the changes observed may be due to secondary effects. Therefore it is important going forward that we learn to disentangle what changes are directly related to disease and which are a downstream effect of those changes.

It is indeed evident that cell-type specific studies will provide much more detailed information in the context of SCZ neurobiology in the future. These will need to be carried out if we are to truly determine the etiology, identify drug targets and develop treatments for this disorder. Chapter Three of this thesis could be improved by generating cell-type specific gene expression profiles from the mouse models from which the gene-sets were derived to more robustly implicate cell-types. While a thorough follow-up to Chapter Four would involve a dual snRNAseq/scATACseq to profile a more diverse set of cell-types similar to a recent snRNAseq/scATACseq study of cortex tissue from control samples (Zhu et al., 2022).

5.5 Concluding Remarks

The work that is presented in this thesis aims to explore cell-type specific perturbations in SCZ using multi-omic data. My initial study aimed to leverage bulk expression data to explore cell-type proportion changes using cell-type deconvolution methods. While future work would be required to fully support the findings of this analysis, it provided a good basis for subsequent work presented in my thesis. This thesis also presents work which aimed to identify cell-types affected by genes under the control of *MEF2C* and identified a non-neuronal cell-type which was not implicated in previous studies. The final study presented here builds upon all of the knowledge that I have accumulated throughout this

process and overcomes some analytical challenges that I have encountered, particularly in relation to Chapter Two. It characterises the transcriptomic and chromatin accessibility landscape of SCZ and presents evidence of extensive disease associated dysregulation across both landscapes particularly in oligodendrocytes. This is important as many other studies have identified neuronal contributions but the contribution of non-neuronal cell-types to SCZ is more poorly understood. This study also illustrates the power of combining multi-omic data to provide new insights into the neurobiology of SCZ. The integration of multi-omic data enabled us to link SCZ-associated epigenetic change with SCZ GWAS signal and detect cell-type specific eQTLs. Gaining new insights into the neurobiology of the disorder in this way is vital in order to one day provide new drug targets and treatment options. Overall, this thesis demonstrates the importance of investigating cell-type specific effects and the power of using multi-omic data to do so.

Bibliography

- Adachi, M., Lin, P.Y., Pranav, H., Monteggia, L.M., 2016. Postnatal loss of *mef2c* results in dissociation of effects on synapse number and learning and memory. *Biol. Psychiatry* 80, 140–148.
- Aguet, F., Alasoo, K., Li, Y.I., Battle, A., Im, H.K., Montgomery, S.B., Lappalainen, T., 2023. Molecular quantitative trait loci. *Nature Reviews Methods Primers* 3, 1–22.
- Albrecht, A., Stork, O., 2012. Are NCAM deficient mice an animal model for schizophrenia? *Front. Behav. Neurosci.* 6, 43.
- Als, T.D., Kurki, M.I., Grove, J., Voloudakis, G., Therrien, K., Tasanko, E., Nielsen, T.T., Naamanka, J., Veerapen, K., Levey, D.F., Bendl, J., Bybjerg-Grauholm, J., Zeng, B., Demontis, D., Rosengren, A., Athanasiadis, G., Bækved-Hansen, M., Qvist, P., Bragi Walters, G., Thorgeirsson, T., Stefánsson, H., Musliner, K.L., Rajagopal, V.M., Farajzadeh, L., Thirstrup, J., Vilhjálmsson, B.J., McGrath, J.J., Mattheisen, M., Meier, S., Agerbo, E., Stefánsson, K., Nordentoft, M., Werge, T., Hougaard, D.M., Mortensen, P.B., Stein, M.B., Gelernter, J., Hovatta, I., Roussos, P., Daly, M.J., Mors, O., Palotie, A., Børghlum, A.D., 2023. Depression pathophysiology, risk prediction of recurrence and comorbid psychiatric disorders using genome-wide analyses. *Nat. Med.* 29, 1832–1844.
- Ament, S.A., Cortes-Gutierrez, M., Herb, B.R., Mocci, E., Colantuoni, C., McCarthy, M.M., 2023. A single-cell genomic atlas for maturation of the human cerebellum during early childhood. *Sci. Transl. Med.* 15, eade1283.
- American Psychiatric Association, 2022. *Diagnostic and statistical manual of mental disorders*. 5th ed., text rev. ed.
- Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M.L., Huber, W., Morgan, M., Gottardo, R., Hicks, S.C., 2020. Orchestrating single-cell analysis with bioconductor. *Nat. Methods* 17, 137–145.
- Castro-de Araujo, L.F., Machado, D.B., Barreto, M.L., Kanaan, R.A., 2020. Subtyping schizophrenia based on symptomatology and cognition using a data driven approach. *Psychiatry Res Neuroimaging* 304, 111136.

- Arion, D., Corradi, J.P., Tang, S., Datta, D., Boothe, F., He, A., Cacace, A.M., Zaczek, R., Albright, C.F., Tseng, G., Lewis, D.A., 2015. Distinctive transcriptome alterations of prefrontal pyramidal neurons in schizophrenia and schizoaffective disorder. *Mol. Psychiatry* 20, 1397–1405.
- Arion, D., Huo, Z., Enwright, J.F., Corradi, J.P., Tseng, G., Lewis, D.A., 2017. Transcriptome alterations in prefrontal pyramidal cells distinguish schizophrenia from bipolar and major depressive disorders. *Biol. Psychiatry* 82, 594–600.
- Arnsten, A.F.T., Wang, M.J., Paspalas, C.D., 2012. Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses. *Neuron* 76, 223–239.
- Arzalluz-Luque, Á., Conesa, A., 2018. Single-cell RNAseq for the study of isoforms-how is that possible? *Genome Biol.* 19, 110.
- Assali, A., Harrington, A.J., Cowan, C.W., 2019. Emerging roles for MEF2 in brain development and mental disorders. *Curr. Opin. Neurobiol.* 59, 49–58.
- Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., De Preter, K., 2020. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* 11, 5650.
- Barbosa, A.C., Kim, M.S., Ertunc, M., Adachi, M., Nelson, E.D., McAnally, J., Richardson, J.A., Kavalali, E.T., Monteggia, L.M., Bassel-Duby, R., Olson, E.N., 2008. MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9391–9396.
- Bauer, D., Haroutunian, V., Meador-Woodruff, J.H., McCullumsmith, R.E., 2010. Abnormal glycosylation of EAAT1 and EAAT2 in prefrontal cortex of elderly patients with schizophrenia. *Schizophr. Res.* 117, 92–98.
- Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., Dekker, J., 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58, 268–276.
- Bendl, J., Hauberg, M.E., Girdhar, K., Im, E., Vicari, J.M., Rahman, S., Fernando, M.B., Townsley, K.G., Dong, P., Misir, R., Kleopoulos, S.P., Reach, S.M., Apontes, P., Zeng, B., Zhang, W., Voloudakis, G., Brennand, K.J., Nixon, R.A., Haroutunian, V., Hoffman, G.E., Fullard, J.F., Roussos, P., 2022. The three-dimensional landscape of cortical chromatin accessibility in alzheimer’s disease. *Nat. Neurosci.* 25, 1366–1378.
- Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S., Pirinen, M., 2016. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501.

- Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J.A., 2010. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045.
- Bhattacharya, A., Vo, D.D., Jops, C., Kim, M., Wen, C., Hervoso, J.L., Pasaniuc, B., Gandal, M.J., 2022. Isoform-level transcriptome-wide association uncovers extensive novel genetic risk mechanisms for neuropsychiatric disorders in the human brain.
- Birnbaum, R., Weinberger, D.R., 2017. Genetic insights into the neurodevelopmental origins of schizophrenia. *Nat. Rev. Neurosci.* 18, 727–740.
- Bleuler, E., Jung, C.G., 1908. Komplexe und krankheitsursachen bei dementia praecox. *Zentralblatt für Nervenheilkunde und Psychiatrie* 31, 220–227.
- Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120.
- Brown, A.S., 2011. Exposure to prenatal infection and risk of schizophrenia. *Front. Psychiatry* 2, 63.
- Bryois, J., Calini, D., Macnair, W., Foo, L., Urich, E., Ortmann, W., Iglesias, V.A., Selvaraj, S., Nutma, E., Marzin, M., Amor, S., Williams, A., Castelo-Branco, G., Menon, V., De Jager, P., Malhotra, D., 2022. 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.
- Bryois, J., Garrett, M.E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G.D., Shieh, A.W., Buil, A., Fullard, J.F., Roussos, P., Sklar, P., Akbarian, S., Haroutunian, V., Stockmeier, C.A., Wray, G.A., White, K.P., Liu, C., Reddy, T.E., Ashley-Koch, A., Sullivan, P.F., Crawford, G.E., 2018. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nat. Commun.* 9, 3121.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218.
- Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2015. ATAC-seq: A method for assaying chromatin accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* 109, 21.29.1–21.29.9.
- Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price,

- A.L., Neale, B.M., 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R., 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420.
- Byrne, M., Agerbo, E., Bennedsen, B., Eaton, W.W., Mortensen, P.B., 2007. Obstetric conditions and risk of first admission with schizophrenia: a danish national register based study. *Schizophr. Res.* 97, 51–59.
- Chailangkarn, T., Trujillo, C.A., Freitas, B.C., Hrvoj-Mihic, B., Herai, R.H., Yu, D.X., Brown, T.T., Marchetto, M.C., Bardy, C., McHenry, L., Stefanacci, L., Järvinen, A., Searcy, Y.M., DeWitt, M., Wong, W., Lai, P., Ard, M.C., Hanson, K.L., Romero, S., Jacobs, B., Dale, A.M., Dai, L., Korenberg, J.R., Gage, F.H., Bellugi, U., Halgren, E., Semendeferi, K., Muotri, A.R., 2016. A human neurodevelopmental model for williams syndrome. *Nature* 536, 338–343.
- Chand, G.B., Dwyer, D.B., Erus, G., Sotiras, A., Varol, E., Srinivasan, D., Doshi, J., Pomponio, R., Pigoni, A., Dazzan, P., Kahn, R.S., Schnack, H.G., Zanetti, M.V., Meisenzahl, E., Busatto, G.F., Crespo-Facorro, B., Pantelis, C., Wood, S.J., Zhuo, C., Shinohara, R.T., Shou, H., Fan, Y., Gur, R.C., Gur, R.E., Satterthwaite, T.D., Koutsouleris, N., Wolf, D.H., Davatzikos, C., 2020. Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. *Brain* 143, 1027–1038.
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga-science* 4, 7.
- Charlson, F.J., Ferrari, A.J., Santomauro, D.F., Diminic, S., Stockings, E., Scott, J.G., McGrath, J.J., Whiteford, H.A., 2018. Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophr. Bull.* 44, 1195.
- Cheli, V.T., Santiago González, D.A., Spreuer, V., Paez, P.M., 2015. Voltage-gated ca^{2+} entry promotes oligodendrocyte progenitor cell maturation and myelination in vitro. *Exp. Neurol.* 265, 69–83.
- Chen, Y.J.J., Johnson, M.A., Lieberman, M.D., Goodchild, R.E., Schobel, S., Lewandowski, N., Rosoklija, G., Liu, R.C., Gingrich, J.A., Small, S., Moore, H., Dwork, A.J., Talmage, D.A., Role, L.W., 2008. Type III neuregulin-1 is required for normal sensorimotor gating, memory-related behaviors, and corticostriatal circuit components. *J. Neurosci.* 28, 6872–6883.

- Choi, S.W., Mak, T.S.H., O'Reilly, P.F., 2020. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* 15, 2759–2772.
- Choi, S.W., O'Reilly, P.F., 2019. PRSice-2: Polygenic risk score software for biobank-scale data. *Gigascience* 8.
- Chou, I.J., Kuo, C.F., Huang, Y.S., Grainge, M.J., Valdes, A.M., See, L.C., Yu, K.H., Luo, S.F., Huang, L.S., Tseng, W.Y., Zhang, W., Doherty, M., 2017. Familial aggregation and heritability of schizophrenia and co-aggregation of psychiatric illnesses in affected families. *Schizophr. Bull.* 43, 1070–1078.
- Choudhury, Z., Lennox, B., 2021. Maternal immune activation and Schizophrenia-Evidence for an immune priming disorder. *Front. Psychiatry* 12, 585742.
- Clapcote, S.J., Lipina, T.V., Millar, J.K., Mackie, S., Christie, S., Ogawa, F., Lerch, J.P., Trimble, K., Uchiyama, M., Sakuraba, Y., Kaneda, H., Shiroishi, T., Houslay, M.D., Henkelman, R.M., Sled, J.G., Gondo, Y., Porteous, D.J., Roder, J.C., 2007. Behavioral phenotypes of *disc1* missense mutations in mice. *Neuron* 54, 387–402.
- Collado-Torres, L., Burke, E.E., Peterson, A., Shin, J., Straub, R.E., Rajpurohit, A., Semick, S.A., Ulrich, W.S., BrainSeq Consortium, Price, A.J., Valencia, C., Tao, R., Deep-Soboslay, A., Hyde, T.M., Kleinman, J.E., Weinberger, D.R., Jaffe, A.E., 2019. Regional heterogeneity in gene expression, regulation, and coherence in the frontal cortex and hippocampus across development and schizophrenia. *Neuron* 103, 203–216.e8.
- Contractor, A., Ethell, I.M., Portera-Cailliau, C., 2021. Cortical interneurons in autism. *Nat. Neurosci.* 24, 1648–1659.
- Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., Satpathy, A.T., Mumbach, M.R., Hoadley, K.A., Robertson, A.G., Sheffield, N.C., Felau, I., Castro, M.A.A., Berman, B.P., Staudt, L.M., Zenklusen, J.C., Laird, P.W., Curtis, C., Cancer Genome Atlas Analysis Network, Greenleaf, W.J., Chang, H.Y., 2018. The chromatin accessibility landscape of primary human cancers. *Science* 362.
- Correll, C.U., Martin, A., Patel, C., Benson, C., Goulding, R., Kern-Sliwa, J., Joshi, K., Schiller, E., Kim, E., 2022a. Systematic literature review of schizophrenia clinical practice guidelines on acute and maintenance management with antipsychotics. *Schizophrenia* 8, 1–10.
- Correll, C.U., Schooler, N.R., 2020. Negative symptoms in schizophrenia: A review and clinical guide for recognition, assessment, and treatment. *Neuropsychiatr. Dis. Treat.* 16, 519–534.

- Correll, C.U., Solmi, M., Croatto, G., Schneider, L.K., Rohani-Montez, S.C., Fairley, L., Smith, N., Bitter, I., Gorwood, P., Taipale, H., Tiihonen, J., 2022b. Mortality in people with schizophrenia: a systematic review and meta-analysis of relative risk and aggravating or attenuating factors. *World Psychiatry* 21, 248–271.
- Cosgrove, D., Whitton, L., Fahey, L., Ó Broin, P., Donohoe, G., Morris, D.W., 2021. Genes influenced by MEF2C contribute to neurodevelopmental disease via gene expression changes that affect multiple types of cortical excitatory neurons. *Hum. Mol. Genet.* 30, 961–970.
- Cropley, V.L., Klauser, P., Lenroot, R.K., Bruggemann, J., Sundram, S., Bousman, C., Pereira, A., Di Biase, M.A., Weickert, T.W., Weickert, C.S., Pantelis, C., Zalesky, A., 2017. Accelerated gray and white matter deterioration with age in schizophrenia. *Am. J. Psychiatry* 174, 286–295.
- Cumming, P., Abi-Dargham, A., Gründer, G., 2021. Molecular imaging of schizophrenia: Neurochemical findings in a heterogeneous and evolving disorder. *Behav. Brain Res.* 398, 113004.
- Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A., Quake, S.R., 2015. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* 112, 7285–7290.
- Davies, G., Welham, J., Chant, D., Torrey, E.F., McGrath, J., 2003. A systematic review and meta-analysis of northern hemisphere season of birth studies in schizophrenia. *Schizophr. Bull.* 29, 587–593.
- Dennison, C.A., Legge, S.E., Pardiñas, A.F., Walters, J.T.R., 2020. Genome-wide association studies in schizophrenia: Recent advances, challenges and future perspective. *Schizophr. Res.* 217, 4–12.
- Ding, J., Adiconis, X., Simmons, S.K., Kowalczyk, M.S., Hession, C.C., Marjanovic, N.D., Hughes, T.K., Wadsworth, M.H., Burks, T., Nguyen, L.T., Kwon, J.Y.H., Barak, B., Ge, W., Kedaigle, A.J., Carroll, S., Li, S., Hacohen, N., Rozenblatt-Rosen, O., Shalek, A.K., Villani, A.C., Regev, A., Levin, J.Z., 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* 38, 737–746.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.
- Dong, P., Bendl, J., Misir, R., Shao, Z., Edelstien, J., Davis, D.A., Haroutunian, V., Scott, W.K., Acker, S., Lawless, N., Hoffman, G.E., Fullard, J.F., Roussos, P., 2022a.

Transcriptome and chromatin accessibility landscapes across 25 distinct human brain regions expand the susceptibility gene set for neuropsychiatric disorders.

- Dong, P., Hoffman, G.E., Apontes, P., Bendl, J., Rahman, S., Fernando, M.B., Zeng, B., Vicari, J.M., Zhang, W., Girdhar, K., Townsley, K.G., Misir, R., CommonMind Consortium, Brennand, K.J., Haroutunian, V., Voloudakis, G., Fullard, J.F., Roussos, P., 2022b. Population-level variation in enhancer expression identifies disease mechanisms in the human brain. *Nat. Genet.* 54, 1493–1503.
- Dong, X., Liao, Z., Gritsch, D., Hadzhiev, Y., Bai, Y., Locascio, J.J., Guennewig, B., Liu, G., Blauwendraat, C., Wang, T., Adler, C.H., Hedreen, J.C., Faull, R.L.M., Frosch, M.P., Nelson, P.T., Rizzu, P., Cooper, A.A., Heutink, P., Beach, T.G., Mattick, J.S., Müller, F., Scherzer, C.R., 2018. Enhancers active in dopamine neurons are a primary link between genetic variation and neuropsychiatric disease. *Nat. Neurosci.* 21, 1482–1492.
- Dudbridge, F., Gusnanto, A., 2008. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* 32, 227.
- Duncan, L.E., Ostacher, M., Ballon, J., 2019. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology* 44, 1518–1523.
- Ellegood, J., Markx, S., Lerch, J.P., Steadman, P.E., Genç, C., Provenzano, F., Kushner, S.A., Henkelman, R.M., Karayiorgou, M., Gogos, J.A., 2014. Neuroanatomical phenotypes in a mouse model of the 22q11.2 microdeletion. *Mol. Psychiatry* 19, 99–107.
- Elloumi, F., Hu, Z., Li, Y., Parker, J.S., Gulley, M.L., Amos, K.D., Troester, M.A., 2011. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. *BMC Med. Genomics* 4, 54.
- Engmann, O., Hortobágyi, T., Pidsley, R., Troakes, C., Bernstein, H.G., Kreutz, M.R., Mill, J., Nikolic, M., Giese, K.P., 2011. Schizophrenia is associated with dysregulation of a cdk5 activator that regulates synaptic protein expression and cognition. *Brain* 134, 2408–2421.
- Enwright, J.F., Huo, Z., Arion, D., Corradi, J.P., Tseng, G., Lewis, D.A., 2018. Transcriptome alterations of prefrontal cortical parvalbumin neurons in schizophrenia. *Mol. Psychiatry* 23, 1606–1613.
- Erdeniz, B., Serin, E., İbadi, Y., Taş, C., 2017. Decreased functional connectivity in schizophrenia: The relationship between social functioning, social cognition and graph theoretical network measures. *Psychiatry Res Neuroimaging* 270, 22–31.

- Ermakov, E.A., Melamud, M.M., Buneva, V.N., Ivanova, S.A., 2022. Immune system abnormalities in schizophrenia: An integrative view and translational perspectives. *Front. Psychiatry* 13, 880568.
- Evins, A.E., Green, A.I., Kane, J.M., Murray, R.M., 2013. Does using marijuana increase the risk for developing schizophrenia? *J. Clin. Psychiatry* 74, e08.
- Eze, U.C., Bhaduri, A., Haeussler, M., Nowakowski, T.J., Kriegstein, A.R., 2021. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* 24, 584–594.
- Fahey, L., Ali, D., Donohoe, G., Ó Broin, P., Morris, D.W., 2023. Genes positively regulated by *mef2c* in cortical neurons are enriched for common genetic variation associated with IQ and educational attainment. *Hum. Mol. Genet.* 32, 3194–3203.
- Fan, C.C., McGrath, J.J., Appadurai, V., Buil, A., Gandal, M.J., Schork, A.J., Mortensen, P.B., Agerbo, E., Geschwind, S.A., Geschwind, D., Werge, T., Thompson, W.K., Pedersen, C.B., 2018. Spatial fine-mapping for gene-by-environment effects identifies risk hot spots for schizophrenia. *Nat. Commun.* 9, 5296.
- Fessel, J., 2022. Abnormal oligodendrocyte function in schizophrenia explains the long latent interval in some patients. *Transl. Psychiatry* 12, 120.
- Filice, F., Schwaller, B., Michel, T.M., Grünblatt, E., 2020. Profiling parvalbumin interneurons using iPSC: challenges and perspectives for autism spectrum disorder (ASD). *Mol. Autism* 11, 10.
- Filice, F., Vörckel, K.J., Sungur, A.Ö., Wöhr, M., Schwaller, B., 2016. Reduction in parvalbumin expression not loss of the parvalbumin-expressing GABA interneuron subpopulation in genetic parvalbumin and shank mouse models of autism. *Mol. Brain* 9, 10.
- Flavell, S.W., Kim, T.K., Gray, J.M., Harmin, D.A., Hemberg, M., Hong, E.J., Markenscoff-Papadimitriou, E., Bear, D.M., Greenberg, M.E., 2008. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron* 60, 1022–1038.
- French, L., Gray, C., Leonard, G., Perron, M., Pike, G.B., Richer, L., Séguin, J.R., Veillette, S., Evans, C.J., Artiges, E., Banaschewski, T., Bokde, A.W.L., Bromberg, U., Bruehl, R., Buchel, C., Cattrell, A., Conrod, P.J., Flor, H., Frouin, V., Gallinat, J., Garavan, H., Gowland, P., Heinz, A., Lemaitre, H., Martinot, J.L., Nees, F., Orfanos, D.P., Pangelinan, M.M., Poustka, L., Rietschel, M., Smolka, M.N., Walter, H., Whelan, R., Timpson, N.J., Schumann, G., Smith, G.D., Pausova, Z., Paus, T., 2015. Early

cannabis use, polygenic risk score for schizophrenia and brain maturation in adolescence. *JAMA Psychiatry* 72, 1002–1011.

Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M., Carrera, N., Humphreys, I., Johnson, J.S., Roussos, P., Barker, D.D., Banks, E., Milanova, V., Grant, S.G., Hannon, E., Rose, S.A., Chambert, K., Mahajan, M., Scolnick, E.M., Moran, J.L., Kirov, G., Palotie, A., McCarroll, S.A., Holmans, P., Sklar, P., Owen, M.J., Purcell, S.M., O'Donovan, M.C., 2014. De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506, 179–184.

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

Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., Nguyen, T.H., Kane, M., Perez, E.M., Durand, N.C., Lareau, C.A., Stamenova, E.K., Aiden, E.L., Lander, E.S., Engreitz, J.M., 2019. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* 51, 1664–1669.

Fullard, J.F., Hauberg, M.E., Bendl, J., Egervari, G., Cirnaru, M.D., Reach, S.M., Motl, J., Ehrlich, M.E., Hurd, Y.L., Roussos, P., 2018. An atlas of chromatin accessibility in the adult human brain. *Genome Res.* 28, 1243–1252.

Gallego, J.A., Blanco, E.A., Morell, C., Lencz, T., Malhotra, A.K., 2021. Complement component C4 levels in the cerebrospinal fluid and plasma of patients with schizophrenia. *Neuropsychopharmacology* 46, 1140–1144.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., Cox, N.J., Im, H.K., 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098.

- Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., Shieh, A.W., Haney, J., Parhami, S., Belmont, J., Kim, M., Moran Losada, P., Khan, Z., Mleczko, J., Xia, Y., Dai, R., Wang, D., Yang, Y.T., Xu, M., Fish, K., Hof, P.R., Warrell, J., Fitzgerald, D., White, K., Jaffe, A.E., PsychENCODE Consortium, Peters, M.A., Gerstein, M., Liu, C., Iakoucheva, L.M., Pinto, D., Geschwind, D.H., 2018. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 362.
- GBD 2016 Disease and Injury Incidence and Prevalence Collaborators, 2017. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet* 390, 1211–1259.
- van de Geijn, B., McVicker, G., Gilad, Y., Pritchard, J.K., 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12, 1061–1063.
- Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D., Diehn, M., West, R.B., Plevritis, S.K., Alizadeh, A.A., 2015. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* 21, 938–945.
- Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., Plagnol, V., 2014. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10, e1004383.
- Girdhar, K., Bendl, J., Baumgartner, A., Therrien, K., Venkatesh, S., Mathur, D., Dong, P., Rahman, S., Kleopoulos, S.P., Misir, R., Reach, S.M., Auluck, P.K., Marengo, S., Lewis, D.A., Haroutunian, V., Funk, C., Voloudakis, G., Hoffman, G.E., Fullard, J.F., Roussos, P., 2023. The neuronal chromatin landscape in adult schizophrenia brains is linked to early fetal development. *medRxiv* .
- Girdhar, K., Hoffman, G.E., Bendl, J., Rahman, S., Dong, P., Liao, W., Hauberg, M.E., Sloofman, L., Brown, L., Devillers, O., Kassim, B.S., Wiseman, J.R., Park, R., Zharovsky, E., Jacobov, R., Flatow, E., Kozlenkov, A., Gilgenast, T., Johnson, J.S., Couto, L., Peters, M.A., Phillips-Cremins, J.E., Hahn, C.G., Gur, R.E., Tamminga, C.A., Lewis, D.A., Haroutunian, V., PsychENCODE Consortium, Dracheva, S., Lipska, B.K., Marengo, S., Kundakovic, M., Fullard, J.F., Jiang, Y., Roussos, P., Akbarian, S., 2022. Chromatin domain alterations linked to 3D genome organization in a large cohort of schizophrenia and bipolar disorder brains. *Nat. Neurosci.* 25, 474–483.
- Girdhar, K., Hoffman, G.E., Jiang, Y., Brown, L., Kundakovic, M., Hauberg, M.E., Francoeur, N.J., Wang, Y.C., Shah, H., Kavanagh, D.H., Zharovsky, E., Jacobov, R.,

- Wiseman, J.R., Park, R., Johnson, J.S., Kassim, B.S., Sloofman, L., Mattei, E., Weng, Z., Sieberts, S.K., Peters, M.A., Harris, B.T., Lipska, B.K., Sklar, P., Roussos, P., Akbarian, S., 2018. Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome. *Nat. Neurosci.* 21, 1126–1136.
- Gong, T., Szustakowski, J.D., 2013. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* 29, 1083–1085.
- Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., Awashti, S., Belliveau, R., Bettella, F., Buxbaum, J.D., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Christensen, J.H., Churchhouse, C., Dellenvall, K., Demontis, D., De Rubeis, S., Devlin, B., Djurovic, S., Dumont, A.L., Goldstein, J.I., Hansen, C.S., Hauberg, M.E., Hollegaard, M.V., Hope, S., Howrigan, D.P., Huang, H., Hultman, C.M., Klei, L., Maller, J., Martin, J., Martin, A.R., Moran, J.L., Nyegaard, M., Nærland, T., Palmer, D.S., Palotie, A., Pedersen, C.B., Pedersen, M.G., dPoterba, T., Poulsen, J.B., Pourcain, B.S., Qvist, P., Rehnström, K., Reichenberg, A., Reichert, J., Robinson, E.B., Roeder, K., Roussos, P., Saemundsen, E., Sandin, S., Satterstrom, F.K., Davey Smith, G., Stefansson, H., Steinberg, S., Stevens, C.R., Sullivan, P.F., Turley, P., Walters, G.B., Xu, X., Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium, BUPGEN, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, 23andMe Research Team, Stefansson, K., Geschwind, D.H., Nordentoft, M., Hougaard, D.M., Werge, T., Mors, O., Mortensen, P.B., Neale, B.M., Daly, M.J., Børglum, A.D., 2019. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* 51, 431–444.
- Grzadzinski, R., Huerta, M., Lord, C., 2013. DSM-5 and autism spectrum disorders (ASDs): an opportunity for identifying ASD subtypes. *Mol. Autism* 4, 12.
- Gu, X., Fu, C., Lin, L., Liu, S., Su, X., Li, A., Wu, Q., Jia, C., Zhang, P., Chen, L., Zhu, X., Wang, X., 2018. mir-124 and mir-9 mediated downregulation of HDAC5 promotes neurite development through activating MEF2C-GPM6A pathway. *J. Cell. Physiol.* 233, 673–687.
- Gupta, I., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A.B., Sloan, S.A., Luo, W., Fedrigo, O., Ross, M.E., Tilgner, H.U., 2018. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* .
- Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Schizophrenia Working Group of the Psychiatric Genomics Consortium, McCarroll,

- S., Neale, B.M., Ophoff, R.A., O'Donovan, M.C., Crawford, G.E., Geschwind, D.H., Katsanis, N., Sullivan, P.F., Pasaniuc, B., Price, A.L., 2018. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* 50, 538–548.
- Han, B., Duong, D., Sul, J.H., de Bakker, P.I.W., Eskin, E., Raychaudhuri, S., 2016. A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum. Mol. Genet.* 25, 1857–1866.
- Han, B., Eskin, E., 2011. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* 88, 586–598.
- Harrington, A.J., Bridges, C.M., Berto, S., Blankenship, K., Cho, J.Y., Assali, A., Siemsen, B.M., Moore, H.W., Tsvetkov, E., Thielking, A., Konopka, G., Everman, D.B., Scofield, M.D., Skinner, S.A., Cowan, C.W., 2020. MEF2C hypofunction in neuronal and neuroimmune populations produces MEF2C haploinsufficiency syndrome-like behaviors in mice. *Biol. Psychiatry* 88, 488–499.
- Harrington, A.J., Raissi, A., Rajkovich, K., Berto, S., Kumar, J., Molinaro, G., Raduazzo, J., Guo, Y., Loerwald, K., Konopka, G., Huber, K.M., Cowan, C.W., 2016. MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *Elife* 5.
- Hauberg, M.E., Creus-Muncunill, J., Bendl, J., Kozlenkov, A., Zeng, B., Corwin, C., Chowdhury, S., Kranz, H., Hurd, Y.L., Wegner, M., Børghlum, A.D., Dracheva, S., Ehrlich, M.E., Fullard, J.F., Roussos, P., 2020. Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat. Commun.* 11, 5581.
- Henriksen, M.G., Nordgaard, J., Jansson, L.B., 2017. Genetics of Schizophrenia: Overview of Methods, Findings and Limitations. *Front. Hum. Neurosci.* 11, 250542.
- Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T.M., Nordentoft, M., Glenthøj, B., 2018. Heritability of schizophrenia and schizophrenia spectrum based on the nationwide danish twin register. *Biol. Psychiatry* 83, 492–498.
- Hoffman, G.E., Bendl, J., Voloudakis, G., Montgomery, K.S., Sloofman, L., Wang, Y.C., Shah, H.R., Hauberg, M.E., Johnson, J.S., Girdhar, K., Song, L., Fullard, J.F., Kramer, R., Hahn, C.G., Gur, R., Marengo, S., Lipska, B.K., Lewis, D.A., Haroutunian, V., Hemby, S., Sullivan, P., Akbarian, S., Chess, A., Buxbaum, J.D., Crawford, G.E., Domenici, E., Devlin, B., Sieberts, S.K., Peters, M.A., Roussos, P., 2019. CommonMind consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. *Sci Data* 6, 180.

- Hoffman, G.E., Ma, Y., Montgomery, K.S., Bendl, J., Jaiswal, M.K., Kozlenkov, A., Peters, M.A., Dracheva, S., Fullard, J.F., Chess, A., Devlin, B., Sieberts, S.K., Roussos, P., 2022. Sex differences in the human brain transcriptome of cases with schizophrenia. *Biol. Psychiatry* 91, 92–101.
- Hoffman, G.E., Roussos, P., 2021. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics* 37, 192–201.
- Hoffman, G.E., Schadt, E.E., 2016. variancepartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* 17, 483.
- Horváth, S., Janka, Z., Mirnics, K., 2011. Analyzing schizophrenia by DNA microarrays. *Biol. Psychiatry* 69, 157–162.
- Howes, O.D., Cummings, C., Chapman, G.E., Shatalina, E., 2023. Neuroimaging in schizophrenia: an overview of findings and their implications for synaptic changes. *Neuropsychopharmacology* 48, 151–167.
- Howrigan, D.P., Rose, S.A., Samocha, K.E., Fromer, M., Cerrato, F., Chen, W.J., Churchhouse, C., Chambert, K., Chandler, S.D., Daly, M.J., Dumont, A., Genovese, G., Hwu, H.G., Laird, N., Kosmicki, J.A., Moran, J.L., Roe, C., Singh, T., Wang, S.H., Faraone, S.V., Glatt, S.J., McCarroll, S.A., Tsuang, M., Neale, B.M., 2020. Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. *Nat. Neurosci.* 23, 185–193.
- Huang, D., Wang, Y., Thompson, J.W., Yin, T., Alexander, P.B., Qin, D., Mudgal, P., Wu, H., Liang, Y., Tan, L., Pan, C., Yuan, L., Wan, Y., Li, Q.J., Wang, X.F., 2022. Cancer-cell-derived GABA promotes β -catenin-mediated tumour growth and immunosuppression. *Nat. Cell Biol.* 24, 230–241.
- Humphrey, J., Brophy, E., Kosoy, R., Zeng, B., Coccia, E., Mattei, D., Ravi, A., Efthymiou, A.G., Navarro, E., Muller, B.Z., Snijders, G.J., Allan, A., Münch, A., Kitata, R.B., Kleopoulos, S.P., Argyriou, S., Shao, Z., Francoeur, N., Tsai, C.F., Gritsenko, M.A., Monroe, M.E., Paurus, V.L., Weitz, K.K., Shi, T., Sebra, R., Liu, T., de Witte, L.D., Goate, A.M., Bennett, D.A., Haroutunian, V., Hoffman, G.E., Fullard, J.F., Roussos, P., Raj, T., 2023. Long-read RNA-seq atlas of novel microglia isoforms elucidates disease-associated genetic regulation of splicing. medRxiv .
- Hunt, G.J., Freytag, S., Bahlo, M., Gagnon-Bartsch, J.A., 2019. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* 35, 2093–2099.
- Huxley, P., Krayner, A., Poole, R., Prendergast, L., Aryal, S., Warner, R., 2021. Schizophrenia outcomes in the 21st century: A systematic review. *Brain Behav.* 11, e02172.

- Hwang, B., Lee, J.H., Bang, D., 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50, 1–14.
- Hyde, C.L., Nagle, M.W., Tian, C., Chen, X., Paciga, S.A., Wendland, J.R., Tung, J.Y., Hinds, D.A., Perlis, R.H., Winslow, A.R., 2016. Identification of 15 genetic loci associated with risk of major depression in individuals of european descent. *Nat. Genet.* 48, 1031–1036.
- International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O’Donovan, M.C., Sullivan, P.F., Sklar, P., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752.
- Jaaro-Peled, H., 2009. Gene models of schizophrenia: DISC1 mouse models. *Prog. Brain Res.* 179, 75–86.
- Jääskeläinen, E., Haapea, M., Rautio, N., Juola, P., Penttilä, M., Nordström, T., Rissanen, I., Husa, A., Keskinen, E., Marttila, R., Filatova, S., Paaso, T.M., Koivukangas, J., Moilanen, K., Isohanni, M., Miettunen, J., 2015. Twenty years of schizophrenia research in the northern finland birth cohort 1966: A systematic review. *Schizophr. Res. Treatment* 2015, 524875.
- Jaffe, A.E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T.M., Weinberger, D.R., Kleinman, J.E., 2016. Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.* 19, 40–47.
- Jaffe, A.E., Hoepfner, D.J., Saito, T., Blanpain, L., Ukaigwe, J., Burke, E.E., Collado-Torres, L., Tao, R., Tajinda, K., Maynard, K.R., Tran, M.N., Martinowich, K., Deep-Soboslay, A., Shin, J.H., Kleinman, J.E., Weinberger, D.R., Matsumoto, M., Hyde, T.M., 2020. Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. *Nat. Neurosci.* 23, 510–519.
- Jaffe, A.E., Straub, R.E., Shin, J.H., Tao, R., Gao, Y., Collado-Torres, L., Kam-Thong, T., Xi, H.S., Quan, J., Chen, Q., Colantuoni, C., Ulrich, W.S., Maher, B.J., Deep-Soboslay, A., BrainSeq Consortium, Cross, A.J., Brandon, N.J., Leek, J.T., Hyde, T.M., Kleinman, J.E., Weinberger, D.R., 2018. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* 21, 1117–1125.
- Jahangir, M., Zhou, J.S., Lang, B., Wang, X.P., 2021. GABAergic system dysfunction and challenges in schizophrenia research. *Front Cell Dev Biol* 9, 663854.

- Jakovcevski, M., Akbarian, S., 2012. Epigenetic mechanisms in neurological disease. *Nat. Med.* 18, 1194–1204.
- Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witte, A., Stringer, S., Aarsland, D., Almdahl, I.S., Andersen, F., Bergh, S., Bettella, F., Bjornsson, S., Brækhus, A., Bråthen, G., de Leeuw, C., Desikan, R.S., Djurovic, S., Dumitrescu, L., Fladby, T., Hohman, T.J., Jonsson, P.V., Kiddle, S.J., Rongve, A., Saltvedt, I., Sando, S.B., Selbæk, G., Shoai, M., Skene, N.G., Snaedal, J., Stordal, E., Ulstein, I.D., Wang, Y., White, L.R., Hardy, J., Hjerling-Leffler, J., Sullivan, P.F., van der Flier, W.M., Dobson, R., Davis, L.K., Stefansson, H., Stefansson, K., Pedersen, N.L., Ripke, S., Andreassen, O.A., Posthuma, D., 2019. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer’s disease risk. *Nat. Genet.* 51, 404–413.
- Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., Halperin, E., 2020. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* 11, 1971.
- Johnson, E.C., Border, R., Melroy-Greif, W.E., de Leeuw, C.A., Ehringer, M.A., Keller, M.C., 2017. No evidence that schizophrenia candidate genes are more associated with schizophrenia than noncandidate genes. *Biol. Psychiatry* 82, 702–708.
- Jonas, K.G., Lencz, T., Li, K., Malhotra, A.K., Perlman, G., Fochtmann, L.J., Bromet, E.J., Kotov, R., 2019. Schizophrenia polygenic risk score and 20-year course of illness in psychotic disorders. *Transl. Psychiatry* 9, 1–8.
- Kahn, R.S., Sommer, I.E., Murray, R.M., Meyer-Lindenberg, A., Weinberger, D.R., Cannon, T.D., O’Donovan, M., Correll, C.U., Kane, J.M., van Os, J., Insel, T.R., 2015. Schizophrenia. *Nature Reviews Disease Primers* 1, 1–23.
- Kane, J.M., Agid, O., Baldwin, M.L., Howes, O., Lindenmayer, J.P., Marder, S., Olfson, M., Potkin, S.G., Correll, C.U., 2019. Clinical guidance on the identification and management of Treatment-Resistant schizophrenia. *J. Clin. Psychiatry* 80.
- Kangas, S.M., Teppo, J., Lahtinen, M.J., Suoranta, A., Ghimire, B., Mattila, P., Uusimaa, J., Varjosalo, M., Katisko, J., Hinttala, R., 2022. Analysis of human brain tissue derived from DBS surgery. *Transl. Neurodegener.* 11, 22.
- Kantrowitz, J.T., Correll, C.U., Jain, R., Cutler, A.J., 2023. New Developments in the Treatment of Schizophrenia: An Expert Roundtable. *Int. J. Neuropsychopharmacol.* 26, 322–330.

- Karlsson Linnér, R., Biroli, P., Kong, E., Meddens, S.F.W., Wedow, R., Fontana, M.A., Lebreton, M., Tino, S.P., Abdellaoui, A., Hammerschlag, A.R., Nivard, M.G., Okbay, A., Rietveld, C.A., Timshel, P.N., Trzaskowski, M., Vlaming, R.d., Zünd, C.L., Bao, Y., Buzdugan, L., Caplin, A.H., Chen, C.Y., Eibich, P., Fontanillas, P., Gonzalez, J.R., Joshi, P.K., Karhunen, V., Kleinman, A., Levin, R.Z., Lill, C.M., Meddens, G.A., Muntané, G., Sanchez-Roige, S., van Rooij, F.J., Taskesen, E., Wu, Y., Zhang, F., 23and Me Research Team, eQTLgen Consortium, International Cannabis Consortium, Social Science Genetic Association Consortium, Auton, A., Boardman, J.D., Clark, D.W., Conlin, A., Dolan, C.C., Fischbacher, U., Groenen, P.J.F., Harris, K.M., Hasler, G., Hofman, A., Ikram, M.A., Jain, S., Karlsson, R., Kessler, R.C., Kooyman, M., MacKillop, J., Männikkö, M., Morcillo-Suarez, C., McQueen, M.B., Schmidt, K.M., Smart, M.C., Sutter, M., Thurik, A.R., Uitterlinden, A.G., White, J., Wit, H.d., Yang, J., Bertram, L., Boomsma, D.I., Esko, T., Fehr, E., Hinds, D.A., Johannesson, M., Kumari, M., Laibson, D., Magnusson, P.K.E., Meyer, M.N., Navarro, A., Palmer, A.A., Pers, T.H., Posthuma, D., Schunk, D., Stein, M.B., Svento, R., Tiemeier, H., Timmers, P.R.H.J., Turley, P., Ursano, R.J., Wagner, G.G., Wilson, J.F., Gratten, J., Lee, J.J., Cesarini, D., Benjamin, D.J., Koellinger, P.D., Beauchamp, J.P., 2019. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* 51, 245–257.
- Kawada, J.I., Takeuchi, S., Imai, H., Okumura, T., Horiba, K., Suzuki, T., Torii, Y., Yasuda, K., Imanaka-Yoshida, K., Ito, Y., 2021. Immune cell infiltration landscapes in pediatric acute myocarditis analyzed by CIBERSORT. *J. Cardiol.* 77, 174–178.
- Kesby, J.P., Eyles, D.W., McGrath, J.J., Scott, J.G., 2018. Dopamine, psychosis and schizophrenia: the widening gap between basic and clinical neuroscience. *Transl. Psychiatry* 8, 30.
- Khachadourian, V., Zaks, N., Lin, E., Reichenberg, A., Janecka, M., 2021. Advanced paternal age and risk of schizophrenia in offspring - review of epidemiological findings and potential mechanisms. *Schizophr. Res.* 233, 72–79.
- Khandaker, G.M., Zimbron, J., Dalman, C., Lewis, G., Jones, P.B., 2012. Childhood infection and adult schizophrenia: a meta-analysis of population-based studies. *Schizophr. Res.* 139, 161–168.
- Kharchenko, P.V., 2021. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* 18, 723–732.
- Kister, A., Kister, I., 2022. Overview of myelin, major myelin lipids, and myelin-associated proteins. *Front Chem* 10, 1041961.

- Koopmans, F., van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijssouw, T., Coba, M.P., Cornelisse, L.N., Farrell, R.J., Goldschmidt, H.L., Howrigan, D.P., Hussain, N.K., Imig, C., de Jong, A.P.H., Jung, H., Kohansalnódehi, M., Kramarz, B., Lipstein, N., Lovering, R.C., MacGillavry, H., Mariano, V., Mi, H., Ninov, M., Osumi-Sutherland, D., Pielot, R., Smalla, K.H., Tang, H., Tashman, K., Toonen, R.F.G., Verpelli, C., Reig-Viader, R., Watanabe, K., van Weering, J., Achsel, T., Ashrafi, G., Asi, N., Brown, T.C., De Camilli, P., Feuermann, M., Foulger, R.E., Gaudet, P., Joglekar, A., Kanellopoulos, A., Malenka, R., Nicoll, R.A., Pulido, C., de Juan-Sanz, J., Sheng, M., Südhof, T.C., Tilgner, H.U., Bagni, C., Bayés, À., Biederer, T., Brose, N., Chua, J.J.E., Dieterich, D.C., Gundelfinger, E.D., Hoogenraad, C., Hugarir, R.L., Jahn, R., Kaeser, P.S., Kim, E., Kreutz, M.R., McPherson, P.S., Neale, B.M., O'Connor, V., Posthuma, D., Ryan, T.A., Sala, C., Feng, G., Hyman, S.E., Thomas, P.D., Smit, A.B., Verhage, M., 2019. SynGO: An Evidence-Based, Expert-Curated knowledge base for the synapse. *Neuron* 103, 217–234.e4.
- Kosoy, R., Fullard, J.F., Zeng, B., Bendl, J., Dong, P., Rahman, S., Kleopoulos, S.P., Shao, Z., Girdhar, K., Humphrey, J., de Paiva Lopes, K., Charney, A.W., Kopell, B.H., Raj, T., Bennett, D., Kellner, C.P., Haroutunian, V., Hoffman, G.E., Roussos, P., 2022. Genetics of the human microglia regulome refines alzheimer's disease risk loci. *Nat. Genet.* 54, 1145–1154.
- Krueger, D.D., Howell, J.L., Hebert, B.F., Olausson, P., Taylor, J.R., Nairn, A.C., 2006. Assessment of cognitive function in the heterozygous reeler mouse. *Psychopharmacology* 189, 95–104.
- Kruse, A.O., Bustillo, J.R., 2022. Glutamatergic dysfunction in schizophrenia. *Transl. Psychiatry* 12, 500.
- Kuo, S.S., Pogue-Geile, M.F., 2019. Variation in fourteen brain structure volumes in schizophrenia: A comprehensive meta-analysis of 246 studies. *Neurosci. Biobehav. Rev.* 98, 85–94.
- Lake, B.B., Codeluppi, S., Yung, Y.C., Gao, D., Chun, J., Kharchenko, P.V., Linnarsson, S., Zhang, K., 2017. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.* 7, 6031.
- Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., Russo, G., Thorton-Wells, T.A., Jones, N., Smith, A.V., Chouraki, V., Thomas, C., Ikram, M.A., Zelenika, D., Vardarajan, B.N., Kamatani, Y., Lin, C.F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M.L., Ruiz, A., Bihoreau, M.T., Choi, S.H., Reitz, C., Pasquier, F., Cruchaga,

C., Craig, D., Amin, N., Berr, C., Lopez, O.L., De Jager, P.L., Deramecourt, V., Johnston, J.A., Evans, D., Lovestone, S., Letenneur, L., Morón, F.J., Rubinsztein, D.C., Eiriksdottir, G., Sleegers, K., Goate, A.M., Fiévet, N., Huentelman, M.W., Gill, M., Brown, K., Kamboh, M.I., Keller, L., Barberger-Gateau, P., McGuinness, B., Larson, E.B., Green, R., Myers, A.J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D.W., Yu, L., Tsolaki, M., Bossù, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N.C., Hardy, J., Deniz Naranjo, M.C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer's Disease, Alzheimer's Disease Genetic Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J.R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M.M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S.G., Coto, E., Hamilton-Nelson, K.L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M.J., Faber, K.M., Jonsson, P.V., Combarros, O., O'Donovan, M.C., Cantwell, L.B., Soininen, H., Blacker, D., Mead, S., Mosley, Jr, T.H., Bennett, D.A., Harris, T.B., Fratiglioni, L., Holmes, C., de Bruijn, R.F., Passmore, P., Montine, T.J., Bettens, K., Rotter, J.I., Brice, A., Morgan, K., Foroud, T.M., Kukull, W.A., Hannequin, D., Powell, J.F., Nalls, M.A., Ritchie, K., Lunetta, K.L., Kauwe, J.S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltunen, M., Martin, E.R., Schmidt, R., Rujescu, D., Wang, L.S., Dartigues, J.F., Mayeux, R., Tzourio, C., Hofman, A., Nöthen, M.M., Graff, C., Psaty, B.M., Jones, L., Haines, J.L., Holmans, P.A., Lathrop, M., Pericak-Vance, M.A., Launer, L.J., Farrer, L.A., van Duijn, C.M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G.D., Amouyel, P., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat. Genet.* 45, 1452–1458.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9, e1003118.

Le Meur, N., Holder-Espinasse, M., Jaillard, S., Goldenberg, A., Joriot, S., Amati-Bonneau, P., Guichet, A., Barth, M., Charollais, A., Journal, H., Auvin, S., Boucher, C., Kerckaert, J.P., David, V., Manouvrier-Hanu, S., Saugier-veber, P., Frébourg, T., Dubourg, C., Andrieux, J., Bonneau, D., 2010. MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental

- retardation with stereotypic movements, epilepsy and/or cerebral malformations. *J. Med. Genet.* 47, 22–29.
- Lee, C.H., Eskin, E., Han, B., 2017. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* 33, i379–i388.
- Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M.A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P.N., Walters, R.K., Willoughby, E.A., Yengo, L., 23andMe Research Team, COGENT (Cognitive Genomics Consortium), Social Science Genetic Association Consortium, Alver, M., Bao, Y., Clark, D.W., Day, F.R., Furlotte, N.A., Joshi, P.K., Kemper, K.E., Kleinman, A., Langenberg, C., Mägi, R., Trampush, J.W., Verma, S.S., Wu, Y., Lam, M., Zhao, J.H., Zheng, Z., Boardman, J.D., Campbell, H., Freese, J., Harris, K.M., Hayward, C., Herd, P., Kumari, M., Lencz, T., Luan, J., Malhotra, A.K., Metspalu, A., Milani, L., Ong, K.K., Perry, J.R.B., Porteous, D.J., Ritchie, M.D., Smart, M.C., Smith, B.H., Tung, J.Y., Wareham, N.J., Wilson, J.F., Beauchamp, J.P., Conley, D.C., Esko, T., Lehrer, S.F., Magnusson, P.K.E., Oskarsson, S., Pers, T.H., Robinson, M.R., Thom, K., Watson, C., Chabris, C.F., Meyer, M.N., Laibson, D.I., Yang, J., Johannesson, M., Koellinger, P.D., Turley, P., Visscher, P.M., Benjamin, D.J., Cesarini, D., 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* 50, 1112–1121.
- de Leeuw, C.A., Mooij, J.M., Heskes, T., Posthuma, D., 2015. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* 11, e1004219.
- Legge, S.E., Cardno, A.G., Allardyce, J., Dennison, C., Hubbard, L., Pardiñas, A.F., Richards, A., Rees, E., Di Florio, A., Escott-Price, V., Zammit, S., Holmans, P., Owen, M.J., O'Donovan, M.C., Walters, J.T.R., 2021. Associations between schizophrenia polygenic liability, symptom dimensions, and cognitive ability in schizophrenia. *JAMA Psychiatry* 78, 1143–1151.
- Leifer, D., Li, Y.L., Wehr, K., 1997. Myocyte-specific enhancer binding factor 2C expression in fetal mouse brain development. *J. Mol. Neurosci.* 8, 131–143.
- Levinson, D.F., Levinson, M.D., Segurado, R., Lewis, C.M., 2003. Genome scan meta-analysis of schizophrenia and bipolar disorder, part i: Methods and power analysis. *Am. J. Hum. Genet.* 73, 17–33.
- Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323.

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Li, H., Radford, J.C., Ragusa, M.J., Shea, K.L., McKercher, S.R., Zaremba, J.D., Soussou, W., Nie, Z., Kang, Y.J., Nakanishi, N., Okamoto, S.I., Roberts, A.J., Schwarz, J.J., Lipton, S.A., 2008. Transcription factor MEF2C influences neural stem/progenitor cell differentiation and maturation in vivo. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9397–9402.
- Liang, D., Elwell, A.L., Aygün, N., Krupa, O., Wolter, J.M., Kyere, F.A., Lafferty, M.J., Cheek, K.E., Courtney, K.P., Yusupova, M., Garrett, M.E., Ashley-Koch, A., Crawford, G.E., Love, M.I., de la Torre-Ubieta, L., Geschwind, D.H., Stein, J.L., 2021. Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. *Nat. Neurosci.* 24, 941–953.
- Liharska, L.E., Park, Y.J., Ziafat, K., Wilkins, L., Silk, H., Linares, L.M., Vornholt, E., Sullivan, B., Cohen, V., Kota, P., Feng, C., Cheng, E., Moya, E., Thompson, R.C., Johnson, J.S., Rieder, M.K., Huang, J., Scarpa, J., Hashemi, A., Polanco, J., Levin, M.A., Nadkarni, G.N., Sebra, R., Crary, J., Schadt, E.E., Beckmann, N.D., Kopell, B.H., Charney, A.W., 2023. A study of gene expression in the living human brain. *medRxiv* .
- Lin, D.Y., Sullivan, P.F., 2009. Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.* 85, 862–872.
- Lin, Q., Schwarz, J., Bucana, C., Olson, E.N., 1997. Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. *Science* 276, 1404–1407.
- Lipska, B.K., Weinberger, D.R., 2000. To model a psychiatric disorder in animals: schizophrenia as a reality test. *Neuropsychopharmacology* 23, 223–239.
- Liu, C., Kanazawa, T., Tian, Y., Mohamed Saini, S., Mancuso, S., Mostaid, M.S., Takahashi, A., Zhang, D., Zhang, F., Yu, H., Doo Shin, H., Sub Cheong, H., Ikeda, M., Kubo, M., Iwata, N., Woo, S.I., Yue, W., Kamatani, Y., Shi, Y., Li, Z., Everall, I., Pantelis, C., Bousman, C., 2019. The schizophrenia genetics knowledgebase: a comprehensive update of findings from candidate gene studies. *Transl. Psychiatry* 9, 205.
- Liu, D., Meyer, D., Fennessy, B., Feng, C., Cheng, E., Johnson, J.S., Park, Y.J., Rieder, M.K., Ascolillo, S., de Pins, A., Dobbyn, A., Lebovitch, D., Moya, E., Nguyen, T.H., Wilkins, L., Hassan, A., Psychiatric Genomics Consortium Phase 3 Targeted Sequencing of Schizophrenia Study Team, Burdick, K.E., Buxbaum, J.D., Domenici, E., Frangou, S., Hartmann, A.M., Laurent-Levinson, C., Malhotra, D., Pato, C.N., Pato, M.T.,

- Ressler, K., Roussos, P., Rujescu, D., Arango, C., Bertolino, A., Blasi, G., Bocchio-Chiavetto, L., Campion, D., Carr, V., Fullerton, J.M., Gennarelli, M., González-Peñas, J., Levinson, D.F., Mowry, B., Nimgaokar, V.L., Pergola, G., Rampino, A., Cervilla, J.A., Rivera, M., Schwab, S.G., Wildenauer, D.B., Daly, M., Neale, B., Singh, T., O'Donovan, M.C., Owen, M.J., Walters, J.T., Ayub, M., Malhotra, A.K., Lencz, T., Sullivan, P.F., Sklar, P., Stahl, E.A., Huckins, L.M., Charney, A.W., 2023. Schizophrenia risk conferred by rare protein-truncating variants is conserved across diverse human populations. *Nat. Genet.* 55, 369–376.
- Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., Abedian, S., Cheon, J.H., Cho, J., Dayani, N.E., Franke, L., Fuyuno, Y., Hart, A., Juyal, R.C., Juyal, G., Kim, W.H., Morris, A.P., Poustchi, H., Newman, W.G., Midha, V., Orchard, T.R., Vahedi, H., Sood, A., Sung, J.Y., Malekzadeh, R., Westra, H.J., Yamazaki, K., Yang, S.K., International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, Barrett, J.C., Alizadeh, B.Z., Parkes, M., Bk, T., Daly, M.J., Kubo, M., Anderson, C.A., Weersma, R.K., 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986.
- Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., Croteau-Chonka, D.C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J.C., Winkler, T.W., Wood, A.R., Workalemahu, T., Faul, J.D., Smith, J.A., Zhao, J.H., Zhao, W., Chen, J., Fehrmann, R., Hedman, Å.K., Karjalainen, J., Schmidt, E.M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J.L., Bragg-Gresham, J.L., Buyske, S., Demirkan, A., Deng, G., Ehret, G.B., Feenstra, B., Feitosa, M.F., Fischer, K., Goel, A., Gong, J., Jackson, A.U., Kanoni, S., Kleber, M.E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Leach, I.M., Medina-Gomez, C., Medland, S.E., Nalls, M.A., Palmer, C.D., Pasko, D., Pechlivanis, S., Peters, M.J., Prokopenko, I., Shungin, D., Stančáková, A., Strawbridge, R.J., Sung, Y.J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S.W., van Setten, J., Van Vliet-Ostaptchouk, J.V., Wang, Z., Yengo, L., Zhang, W., Isaacs, A., Albrecht, E., Ärnlöv, J., Arscott, G.M., Attwood, A.P., Bandinelli, S., Barrett, A., Bas, I.N., Bellis, C., Bennett, A.J., Berne, C., Blagieva, R., Blüher, M., Böhringer, S., Bonnycastle, L.L., Böttcher, Y., Boyd, H.A., Bruinenberg, M., Caspersen, I.H., Chen, Y.D.I., Clarke, R., Daw, E.W., de Craen, A.J.M., Delgado, G., Dimitriou, M., Doney, A.S.F., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R.M., Garcia, M.E., Geller, F., Giedraitis, V., Gigante, B., Go, A.S., Golay, A., Goodall, A.H., Gordon, S.D., Gorski, M., Grabe, H.J., Grallert, H., Grammer, T.B., Gräßler, J., Grönberg, H., Groves, C.J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C.A., Hassinen, M., Hayward, C., Heard-Costa, N.L., Helmer,

Q., Hengstenberg, C., Holmen, O., Hottenga, J.J., James, A.L., Jeff, J.M., Johansson, Å., Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W., Laitinen, J., Lamina, C., Leander, K., Lee, N.R., Lichtner, P., Lind, L., Lindström, J., Lo, K.S., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P.K.E., Mahajan, A., McArdle, W.L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K.L., Morken, M.A., Mulas, A., Müller, G., Müller-Nurasyid, M., Musk, A.W., Nagaraja, R., Nöthen, M.M., Nolte, I.M., Pilz, S., Rayner, N.W., Renstrom, F., Rettig, R., Ried, J.S., Ripke, S., Robertson, N.R., Rose, L.M., Sanna, S., Schernagl, H., Scholtens, S., Schumacher, F.R., Scott, W.R., Seufferlein, T., Shi, J., Smith, A.V., Smolonska, J., Stanton, A.V., Steinthorsdottir, V., Stirrups, K., Stringham, H.M., Sundström, J., Swertz, M.A., Swift, A.J., Syvänen, A.C., Tan, S.T., Tayo, B.O., Thorand, B., Thorleifsson, G., Tyrer, J.P., Uh, H.W., Vandenput, L., Verhulst, F.C., Vermeulen, S.H., Verweij, N., Vonk, J.M., Waite, L.L., Warren, H.R., Waterworth, D., Weedon, M.N., Wilkens, L.R., Willenborg, C., Wilsgaard, T., Wojczynski, M.K., Wong, A., Wright, A.F., Zhang, Q., LifeLines Cohort Study, Brennan, E.P., Choi, M., Dastani, Z., Drong, A.W., Eriksson, P., Franco-Cereceda, A., Gådin, J.R., Gharavi, A.G., Goddard, M.E., Handsaker, R.E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J.Y., Liang, L., Lifton, R.P., Ma, B., McCarroll, S.A., McKnight, A.J., Min, J.L., Moffatt, M.F., Montgomery, G.W., Murabito, J.M., Nicholson, G., Nyholt, D.R., Okada, Y., Perry, J.R.B., Dorajoo, R., Reinmaa, E., Salem, R.M., Sandholm, N., Scott, R.A., Stolk, L., Takahashi, A., Tanaka, T., van 't Hooft, F.M., Vinkhuyzen, A.A.E., Westra, H.J., Zheng, W., Zondervan, K.T., ADIPOGen Consortium, AGEN-BMI Working Group, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GLGC, ICBP, MAGIC Investigators, MuTHER Consortium, MI-Gen Consortium, PAGE Consortium, ReproGen Consortium, GENIE Consortium, International Endogene Consortium, Heath, A.C., Arveiler, D., Bakker, S.J.L., Beilby, J., Bergman, R.N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M.J., Cesana, G., Chakravarti, A., Chasman, D.I., Chines, P.S., Collins, F.S., Crawford, D.C., Cupples, L.A., Cusi, D., Danesh, J., de Faire, U., den Ruijter, H.M., Dominiczak, A.F., Erbel, R., Erdmann, J., Eriksson, J.G., Farrall, M., Felix, S.B., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N.G., Forrester, T., Franco, O.H., Gansevoort, R.T., Gejman, P.V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U., Hall, A.S., Harris, T.B., Hattersley, A.T., Hicks, A.A., Hindorf, L.A., Hingorani, A.D., Hofman, A., Homuth, G., Hovingh, G.K., Humphries, S.E., Hunt, S.C., Hyppönen, E., Illig, T., Jacobs, K.B., Jarvelin, M.R., Jöckel, K.H., Johansen, B., Jousilahti, P., Jukema, J.W., Jula, A.M., Kaprio, J., Kastelein, J.J.P., Keinanen-Kiukkaanniemi, S.M., Kiemeny, L.A., Knekt, P., Kooner, J.S., Kooperberg, C., Kovacs, P., Kraja, A.T., Kumari, M., Kuusisto, J., Lakka, T.A., Langenberg, C., Marchand, L.L., Lehtimäki, T., Lyssenko, V., Männistö, S., Marette, A., Matise, T.C., McKenzie, C.A., McKnight, B., Moll, F.L., Morris, A.D.,

Morris, A.P., Murray, J.C., Nelis, M., Ohlsson, C., Oldehinkel, A.J., Ong, K.K., Madden, P.A.F., Pasterkamp, G., Peden, J.F., Peters, A., Postma, D.S., Pramstaller, P.P., Price, J.F., Qi, L., Raitakari, O.T., Rankinen, T., Rao, D.C., Rice, T.K., Ridker, P.M., Rioux, J.D., Ritchie, M.D., Rudan, I., Salomaa, V., Samani, N.J., Saramies, J., Sarzynski, M.A., Schunkert, H., Schwarz, P.E.H., Sever, P., Shuldiner, A.R., Sinisalo, J., Stolk, R.P., Strauch, K., Tönjes, A., Trégouët, D.A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.C., Völker, U., Waeber, G., Willemsen, G., Wittteman, J.C., Zillikens, M.C., Adair, L.S., Amouyel, P., Asselbergs, F.W., Assimes, T.L., Bochud, M., Boehm, B.O., Boerwinkle, E., Bornstein, S.R., Bottinger, E.P., Bouchard, C., Cauchi, S., Chambers, J.C., Chanock, S.J., Cooper, R.S., de Bakker, P.I.W., Dedoussis, G., Ferrucci, L., Franks, P.W., Froguel, P., Groop, L.C., Haiman, C.A., Hamsten, A., Hui, J., Hunter, D.J., Hveem, K., Kaplan, R.C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N.G., März, W., Melbye, M., Metspalu, A., Moebus, S., Munroe, P.B., Njølstad, I., Oostra, B.A., Palmer, C.N.A., Pedersen, N.L., Perola, M., Pérusse, L., Peters, U., Power, C., Quertermous, T., Rauramaa, R., Rivadeneira, F., Saaristo, T.E., Saleheen, D., Sattar, N., Schadt, E.E., Schlessinger, D., Slagboom, P.E., Snieder, H., Spector, T.D., Thorsteinsdottir, U., Stumvoll, M., Tuomilehto, J., Uitterlinden, A.G., Uusitupa, M., van der Harst, P., Walker, M., Wallaschofski, H., Wareham, N.J., Watkins, H., Weir, D.R., Wichmann, H.E., Wilson, J.F., Zanen, P., Borecki, I.B., Deloukas, P., Fox, C.S., Heid, I.M., O'Connell, J.R., Strachan, D.P., Stefansson, K., van Duijn, C.M., Abecasis, G.R., Franke, L., Frayling, T.M., McCarthy, M.I., Visscher, P.M., Scherag, A., Willer, C.J., Boehnke, M., Mohlke, K.L., Lindgren, C.M., Beckmann, J.S., Barroso, I., North, K.E., Ingelsson, E., Hirschhorn, J.N., Loos, R.J.F., Speliotes, E.K., 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, Harold, Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M.,

- Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalín, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F., 2013. The Genotype-Tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585.
- Lyons, G.E., Micales, B.K., Schwarz, J., Martin, J.F., Olson, E.N., 1995. Expression of *mef2* genes in the mouse central nervous system suggests a role in neuronal maturation. *J. Neurosci.* 15, 5727–5738.
- Ma, Q., Telese, F., 2015. Genome-wide epigenetic analysis of MEF2A and MEF2C transcription factors in mouse cortical neurons. *Commun. Integr. Biol.* 8, e1087624.
- Maas, D.A., Martens, M.B., Priovoulos, N., Zuure, W.A., Homberg, J.R., Nait-Oumesmar, B., Martens, G.J.M., 2020. Key role for lipids in cognitive symptoms of schizophrenia. *Transl. Psychiatry* 10, 399.
- Mah, S., Nelson, M.R., Delisi, L.E., Reneland, R.H., Markward, N., James, M.R., Nyholt, D.R., Hayward, N., Handoko, H., Mowry, B., Kammerer, S., Braun, A., 2006. Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. *Mol. Psychiatry* 11, 471–478.
- Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., Gujral, M., Brandler, W.M., Malhotra, D., Wang, Z., Fajardo, K.V.F., Maile, M.S., Ripke, S., Agartz, I., Albus, M., Alexander, M., Amin, F., Atkins, J., Bacanu, S.A., Belliveau, Jr, R.A., Bergen, S.E., Bertalan, M., Bevilacqua, E., Bigdeli, T.B., Black, D.W., Bruggeman, R., Buccola, N.G., Buckner, R.L., Bulik-Sullivan, B., Byerley, W., Cahn, W., Cai, G., Cairns, M.J., Champion, D., Cantor, R.M., Carr, V.J., Carrera, N., Catts, S.V., Chambert, K.D., Cheng, W., Cloninger, C.R., Cohen, D., Cormican, P., Craddock, N., Crespo-Facorro, B., Crowley, J.J., Curtis, D., Davidson, M., Davis, K.L., Degenhardt, F., Del Favero, J., DeLisi, L.E., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A.H., Farh, K.H., Farrell, M.S., Frank, J., Franke, L., Freedman, R., Freimer, N.B., Friedman, J.I., Forstner, A.J., Fromer, M., Genovese, G., Georgieva, L., Gershon, E.S., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J.I., Gratten, J., de Haan, L., Hamshere, M.L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A.M., Henskens, F.A., Herms, S., Hirschhorn, J.N., Hoffmann, P., Hofman, A., Huang, H., Ikeda, M., Joa, I., Kähler, A.K., Kahn, R.S., Kalaydjieva,

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

Marttinen, M., Kurkinen, K.M., Soininen, H., Haapasalo, A., Hiltunen, M., 2015. Synaptic dysfunction and septin protein family members in neurodegenerative diseases. *Mol. Neurodegener.* 10, 16.

Maury, E.A., Sherman, M.A., Genovese, G., Gilgenast, T.G., Kamath, T., Burris, S.J., Rajarajan, P., Flaherty, E., Akbarian, S., Chess, A., McCarroll, S.A., Loh, P.R., Phillips-Cremins, J.E., Brennand, K.J., Macosko, E.Z., Walters, J.T.R., O'Donovan, M., Sullivan, P., Psychiatric Genomic Consortium Schizophrenia and CNV workgroup, Brain Somatic Mosaicism Network, Sebat, J., Lee, E.A., Walsh, C.A., 2023.

- Schizophrenia-associated somatic copy-number variants from 12,834 cases reveal recurrent NRXN1 and ABCB11 disruptions. *Cell Genom* 3, 100356.
- McCutcheon, R.A., Keefe, R.S.E., McGuire, P.K., 2023. Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. *Mol. Psychiatry* 28, 1902–1918.
- McCutcheon, R.A., Krystal, J.H., Howes, O.D., 2020. Dopamine and glutamate in schizophrenia: biology, symptoms and treatment. *World Psychiatry* 19, 15–33.
- McGuire, J.L., Depasquale, E.A., Funk, A.J., O'Donnovan, S.M., Hasselfeld, K., Marwaha, S., Hammond, J.H., Hartounian, V., Meador-Woodruff, J.H., Meller, J., McCullumsmith, R.E., 2017. Abnormalities of signal transduction networks in chronic schizophrenia. *NPJ Schizophr* 3, 30.
- McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* 28, 495–501.
- McNamara, N.B., Munro, D.A.D., Bestard-Cuche, N., Uyeda, A., Bogie, J.F.J., Hoffmann, A., Holloway, R.K., Molina-Gonzalez, I., Askew, K.E., Mitchell, S., Mungall, W., Dodds, M., Dittmayer, C., Moss, J., Rose, J., Szymkowiak, S., Amann, L., McColl, B.W., Prinz, M., Spires-Jones, T.L., Stenzel, W., Horsburgh, K., Hendriks, J.J.A., Pridans, C., Muramatsu, R., Williams, A., Priller, J., Miron, V.E., 2023. Microglia regulate central nervous system myelin growth and integrity. *Nature* 613, 120–129.
- Mealer, R.G., Williams, S.E., Noel, M., Yang, B., D'Souza, A.K., Nakata, T., Graham, D.B., Creasey, E.A., Cetinbas, M., Sadreyev, R.I., Scolnick, E.M., Woo, C.M., Smoller, J.W., Xavier, R.J., Cummings, R.D., 2022. The schizophrenia-associated variant in SLC39A8 alters protein glycosylation in the mouse brain. *Mol. Psychiatry* 27, 1405–1415.
- Mendizabal, I., Berto, S., Usui, N., Toriumi, K., Chatterjee, P., Douglas, C., Huh, I., Jeong, H., Layman, T., Tamminga, C.A., Preuss, T.M., Konopka, G., Yi, S.V., 2019. Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biol.* 20, 135.
- Middleton, F.A., Mirnics, K., Pierri, J.N., Lewis, D.A., Levitt, P., 2002. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J. Neurosci.* 22, 2718–2729.
- Milite, C., Feoli, A., Viviano, M., Rescigno, D., Cianciulli, A., Balzano, A.L., Mai, A., Castellano, S., Sbardella, G., 2016. The emerging role of lysine methyltransferase SETD8 in human diseases. *Clin. Epigenetics* 8, 102.

- Millar, J.K., Wilson-Annan, J.C., Anderson, S., Christie, S., Taylor, M.S., Semple, C.A., Devon, R.S., St Clair, D.M., Muir, W.J., Blackwood, D.H., Porteous, D.J., 2000. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* 9, 1415–1423.
- Mirnics, K., Levitt, P., Lewis, D.A., 2006. Critical appraisal of DNA microarrays in psychiatric genomics. *Biol. Psychiatry* 60, 163–176.
- Mirnics, K., Middleton, F.A., Marquez, A., Lewis, D.A., Levitt, P., 2000. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* 28, 53–67.
- Mirnics, K., Pevsner, J., 2004. Progress in the use of microarray technology to study the neurobiology of disease. *Nat. Neurosci.* 7, 434–439.
- Mistry, M., Gillis, J., Pavlidis, P., 2013. Genome-wide expression profiling of schizophrenia using a large combined cohort. *Mol. Psychiatry* 18, 215–225.
- Mitchell, A.C., Javidfar, B., Pothula, V., Ibi, D., Shen, E.Y., Peter, C.J., Bicks, L.K., Fehr, T., Jiang, Y., Brennand, K.J., Neve, R.L., Gonzalez-Maeso, J., Akbarian, S., 2018. MEF2C transcription factor is associated with the genetic and epigenetic risk architecture of schizophrenia and improves cognition in mice. *Mol. Psychiatry* 23, 123–132.
- Momtazmanesh, S., Zare-Shahabadi, A., Rezaei, N., 2019. Cytokine alterations in schizophrenia: An updated review. *Front. Psychiatry* 10, 892.
- Morris, J.A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D.A., Hao, S., Mimitou, E.P., Smibert, P., Roeder, K., Katsevich, E., Lappalainen, T., Sanjana, N.E., 2023. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* 380, eadh7699.
- Mueller, T.M., Haroutunian, V., Meador-Woodruff, J.H., 2014. N-Glycosylation of GABAA receptor subunits is altered in schizophrenia. *Neuropsychopharmacology* 39, 528–537.
- Mullins, N., Forstner, A.J., O’Connell, K.S., Coombes, B., Coleman, J.R.I., Qiao, Z., Als, T.D., Bigdeli, T.B., Børte, S., Bryois, J., Charney, A.W., Drange, O.K., Gandal, M.J., Hagenaars, S.P., Ikeda, M., Kamitaki, N., Kim, M., Krebs, K., Panagiotaropoulou, G., Schilder, B.M., Sloofman, L.G., Steinberg, S., Trubetskoy, V., Winsvold, B.S., Won, H.H., Abramova, L., Adorjan, K., Agerbo, E., Al Eissa, M., Albani, D., Alliey-Rodriguez, N., Anjorin, A., Antilla, V., Antoniou, A., Awasthi, S., Baek, J.H., Bækvad-Hansen, M., Bass, N., Bauer, M., Beins, E.C., Bergen, S.E., Birner,

A., Bøcker Pedersen, C., Bøen, E., Boks, M.P., Bosch, R., Brum, M., Brumpton, B.M., Brunkhorst-Kanaan, N., Budde, M., Bybjerg-Grauholm, J., Byerley, W., Cairns, M., Casas, M., Cervantes, P., Clarke, T.K., Cruceanu, C., Cuellar-Barboza, A., Cunningham, J., Curtis, D., Czerski, P.M., Dale, A.M., Dalkner, N., David, F.S., Degenhardt, F., Djurovic, S., Dobbyn, A.L., Douzenis, A., Elvsåshagen, T., Escott-Price, V., Ferrer, I.N., Fiorentino, A., Foroud, T.M., Forty, L., Frank, J., Frei, O., Freimer, N.B., Frisé, L., Gade, K., Garnham, J., Gelernter, J., Giørtz Pedersen, M., Gizer, I.R., Gordon, S.D., Gordon-Smith, K., Greenwood, T.A., Grove, J., Guzman-Parra, J., Ha, K., Haraldsson, M., Hautzinger, M., Heilbronner, U., Hellgren, D., Herms, S., Hoffmann, P., Holmans, P.A., Huckins, L., Jamain, S., Johnson, J.S., Kalman, J.L., Kamatani, Y., Kennedy, J.L., Kittel-Schneider, S., Knowles, J.A., Kogevinas, M., Koromina, M., Kranz, T.M., Kranzler, H.R., Kubo, M., Kupka, R., Kushner, S.A., Lavebratt, C., Lawrence, J., Leber, M., Lee, H.J., Lee, P.H., Levy, S.E., Lewis, C., Liao, C., Lucae, S., Lundberg, M., MacIntyre, D.J., Magnusson, S.H., Maier, W., Maihofer, A., Malaspina, D., Maratou, E., Martinsson, L., Mattheisen, M., McCarroll, S.A., McGregor, N.W., McGuffin, P., McKay, J.D., Medeiros, H., Medland, S.E., Millischer, V., Montgomery, G.W., Moran, J.L., Morris, D.W., Mühleisen, T.W., O'Brien, N., O'Donovan, C., Olde Loohuis, L.M., Oruc, L., Papiol, S., Pardiñas, A.F., Perry, A., Pfennig, A., Porichi, E., Potash, J.B., Quedt, D., Raj, T., Rapaport, M.H., DePaulo, J.R., Regeer, E.J., Rice, J.P., Rivas, F., Rivera, M., Roth, J., Roussos, P., Ruderfer, D.M., Sánchez-Mora, C., Schulte, E.C., Senner, F., Sharp, S., Shilling, P.D., Sigurdsson, E., Sirignano, L., Slaney, C., Smeland, O.B., Smith, D.J., Sobell, J.L., Söholm Hansen, C., Soler Artigas, M., Spijker, A.T., Stein, D.J., Strauss, J.S., Światkowska, B., Terao, C., Thorgeirsson, T.E., Toma, C., Tooney, P., Tsermpini, E.E., Vawter, M.P., Vedder, H., Walters, J.T.R., Witt, S.H., Xi, S., Xu, W., Yang, J.M.K., Young, A.H., Young, H., Zandi, P.P., Zhou, H., Zillich, L., HUNT All-In Psychiatry, Adolfsson, R., Agartz, I., Alda, M., Alfredsson, L., Babadjanova, G., Backlund, L., Baune, B.T., Bellivier, F., Bengesser, S., Berrettini, W.H., Blackwood, D.H.R., Boehnke, M., Børghlum, A.D., Breen, G., Carr, V.J., Catts, S., Corvin, A., Craddock, N., Dannlowski, U., Dikeos, D., Esko, T., Etain, B., Ferentinos, P., Frye, M., Fullerton, J.M., Gawlik, M., Gershon, E.S., Goes, F.S., Green, M.J., Grigoriou-Serbanescu, M., Hauser, J., Henskens, F., Hillert, J., Hong, K.S., Hougaard, D.M., Hultman, C.M., Hveem, K., Iwata, N., Jablensky, A.V., Jones, I., Jones, L.A., Kahn, R.S., Kelsoe, J.R., Kirov, G., Landén, M., Leboyer, M., Lewis, C.M., Li, Q.S., Lissowska, J., Lochner, C., Loughland, C., Martin, N.G., Mathews, C.A., Mayoral, F., McElroy, S.L., McIntosh, A.M., McMahan, F.J., Melle, I., Michie, P., Milani, L., Mitchell, P.B., Morken, G., Mors, O., Mortensen, P.B., Mowry, B., Müller-Myhsok, B., Myers, R.M., Neale, B.M., Nievergelt, C.M., Nordentoft, M., Nöthen, M.M., O'Donovan, M.C., Oedegaard, K.J., Olsson, T., Owen, M.J., Paciga, S.A., Pantelis, C., Pato, C., Pato, M.T., Patrinos, G.P., Perlis, R.H., Posthuma, D.,

- Ramos-Quiroga, J.A., Reif, A., Reininghaus, E.Z., Ribasés, M., Rietschel, M., Ripke, S., Rouleau, G.A., Saito, T., Schall, U., Schalling, M., Schofield, P.R., Schulze, T.G., Scott, L.J., Scott, R.J., Serretti, A., Shannon Weickert, C., Smoller, J.W., Stefansson, H., Stefansson, K., Stordal, E., Streit, F., Sullivan, P.F., Turecki, G., Vaaler, A.E., Vieta, E., Vincent, J.B., Waldman, I.D., Weickert, T.W., Werge, T., Wray, N.R., Zwart, J.A., Biernacka, J.M., Nurnberger, J.I., Cichon, S., Edenberg, H.J., Stahl, E.A., McQuillin, A., Di Florio, A., Ophoff, R.A., Andreassen, O.A., 2021. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* 53, 817–829.
- Nakamura, T., Takata, A., 2023. The molecular pathology of schizophrenia: an overview of existing knowledge and new directions for future research. *Mol. Psychiatry* 28, 1868–1889.
- Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., von Coelln, R., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlstrøm, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Botia, J.A., Martinez, M., Corvol, J.C., Lesage, S., Jankovic, J., Shulman, L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher, P.M., Graham, R.R., Singleton, A.B., 23andMe Research Team, System Genomics of Parkinson’s Disease Consortium, International Parkinson’s Disease Genomics Consortium, 2019. 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.
- Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., Mualim, K., Natri, H.M., Weeks, E.M., Munson, G., Kane, M., Kang, H.Y., Cui, A., Ray, J.P., Eisenhaure, T.M., Collins, R.L., Dey, K., Pfister, H., Price, A.L., Epstein, C.B., Kundaje, A., Xavier, R.J., Daly, M.J., Huang, H., Finucane, H.K., Hacohen, N., Lander, E.S., Engreitz, J.M., 2021. Genome-wide enhancer maps link risk variants to disease genes. *Nature* 593, 238–243.
- Newcomer, J.W., Farber, N.B., Jevtovic-Todorovic, V., Selke, G., Melson, A.K., Hershey, T., Craft, S., Olney, J.W., 1999. Ketamine-induced NMDA receptor hypofunction as a model of memory impairment and psychosis. *Neuropsychopharmacology* 20, 106–118.
- Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., Diehn, M., Alizadeh, A.A.,

2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 37, 773–782.
- Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D.A., Mostafavi, S., De Jager, P.L., 2017. An xQTL map integrates the genetic architecture of the human brain’s transcriptome and epigenome. *Nat. Neurosci.* 20, 1418–1426.
- Ng, M.Y.M., Levinson, D.F., Faraone, S.V., Suarez, B.K., DeLisi, L.E., Arinami, T., Riley, B., Paunio, T., Pulver, A.E., Irmansyah, Holmans, P.A., Escamilla, M., Wildenauer, D.B., Williams, N.M., Laurent, C., Mowry, B.J., Brzustowicz, L.M., Maziade, M., Sklar, P., Garver, D.L., Abecasis, G.R., Lerer, B., Fallin, M.D., Gurling, H.M.D., Gejman, P.V., Lindholm, E., Moises, H.W., Byerley, W., Wijsman, E.M., Forabosco, P., Tsuang, M.T., Hwu, H.G., Okazaki, Y., Kendler, K.S., Wormley, B., Fanous, A., Walsh, D., O’Neill, F.A., Peltonen, L., Nestadt, G., Lasseter, V.K., Liang, K.Y., Papadimitriou, G.M., Dikeos, D.G., Schwab, S.G., Owen, M.J., O’Donovan, M.C., Norton, N., Hare, E., Raventos, H., Nicolini, H., Albus, M., Maier, W., Nimgaonkar, V.L., Terenius, L., Mallet, J., Jay, M., Godard, S., Nertney, D., Alexander, M., Crowe, R.R., Silverman, J.M., Bassett, A.S., Roy, M.A., Mérette, C., Pato, C.N., Pato, M.T., Roos, J.L., Kohn, Y., Amann-Zalcenstein, D., Kalsi, G., McQuillin, A., Curtis, D., Brynjolfson, J., Sigmundsson, T., Petursson, H., Sanders, A.R., Duan, J., Jazin, E., Myles-Worsley, M., Karayiorgou, M., Lewis, C.M., 2009. Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol. Psychiatry* 14, 774–785.
- Ni, G., Zeng, J., Revez, J.A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D.R., Coleman, J.R.I., Smoller, J.W., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Yang, J., Visscher, P.M., Wray, N.R., 2021. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol. Psychiatry* 90, 611–620.
- Notaras, M., Lodhi, A., Dündar, F., Collier, P., Sayles, N.M., Tilgner, H., Greening, D., Colak, D., 2022. Schizophrenia is defined by cell-specific neuropathology and multiple neurodevelopmental mechanisms in patient-derived cerebral organoids. *Mol. Psychiatry* 27, 1416–1434.
- Nott, A., Holtman, I.R., Coufal, N.G., Schlachetzki, J.C.M., Yu, M., Hu, R., Han, C.Z., Pena, M., Xiao, J., Wu, Y., Keulen, Z., Pasillas, M.P., O’Connor, C., Nickl, C.K., Schafer, S.T., Shen, Z., Rissman, R.A., Brewer, J.B., Gosselin, D., Gonda, D.D., Levy, M.L., Rosenfeld, M.G., McVicker, G., Gage, F.H., Ren, B., Glass, C.K., 2019. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* 366, 1134–1139.

- Nyegaard, M., Demontis, D., Foldager, L., Hedemand, A., Flint, T.J., Sørensen, K.M., Andersen, P.S., Nordentoft, M., Werge, T., Pedersen, C.B., Hougaard, D.M., Mortensen, P.B., Mors, O., Børglum, A.D., 2010. CACNA1C (rs1006737) is associated with schizophrenia. *Mol. Psychiatry* 15, 119–121.
- Ochoa, S., Usall, J., Cobo, J., Labad, X., Kulkarni, J., 2012. Gender differences in schizophrenia and first-episode psychosis: a comprehensive literature review. *Schizophr. Res. Treatment* 2012, 916198.
- O'Donovan, M.C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., Nikolov, I., Hamshere, M., Carroll, L., Georgieva, L., Dwyer, S., Holmans, P., Marchini, J.L., Spencer, C.C.A., Howie, B., Leung, H.T., Hartmann, A.M., Möller, H.J., Morris, D.W., Shi, Y., Feng, G., Hoffmann, P., Propping, P., Vasilescu, C., Maier, W., Rietschel, M., Zammit, S., Schumacher, J., Quinn, E.M., Schulze, T.G., Williams, N.M., Giegling, I., Iwata, N., Ikeda, M., Darvasi, A., Shifman, S., He, L., Duan, J., Sanders, A.R., Levinson, D.F., Gejman, P.V., Cichon, S., Nöthen, M.M., Gill, M., Corvin, A., Rujescu, D., Kirov, G., Owen, M.J., Buccola, N.G., Mowry, B.J., Freedman, R., Amin, F., Black, D.W., Silverman, J.M., Byerley, W.F., Cloninger, C.R., Molecular Genetics of Schizophrenia Collaboration, 2008. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat. Genet.* 40, 1053–1055.
- Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S.M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., Jiang, Y., Hicks, B., Tian, C., Hinds, D.A., Ahlskog, R., Magnusson, P.K.E., Oskarsson, S., Hayward, C., Campbell, A., Porteous, D.J., Freese, J., Herd, P., 23andMe Research Team, Social Science Genetic Association Consortium, Watson, C., Jala, J., Conley, D., Koellinger, P.D., Johannesson, M., Laibson, D., Meyer, M.N., Lee, J.J., Kong, A., Yengo, L., Cesarini, D., Turley, P., Visscher, P.M., Beauchamp, J.P., Benjamin, D.J., Young, A.I., 2022. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* 54, 437–449.
- Oluwadare, O., Highsmith, M., Cheng, J., 2019. An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biol. Proced. Online* 21, 7.
- Orrico-Sá, A., Ló-Lacort, M., Muñ-Quiles, C., Sanfé-Gimeno, G., Dí-Domingo, J., 2020. Epidemiology of schizophrenia and its management over 8-years period using real-world data in Spain. *BMC Psychiatry* 20, 1–9.
- Owen, M.J., O'Donovan, M.C., 2017. Schizophrenia and the neurodevelopmental continuum:evidence from genomics. *World Psychiatry* 16, 227–235.

- Owen, M.J., O'Donovan, M.C., Thapar, A., Craddock, N., 2011. Neurodevelopmental hypothesis of schizophrenia. *Br. J. Psychiatry* 198, 173–175.
- Owen, M.J., Sawa, A., Mortensen, P.B., 2016. Schizophrenia. *Lancet* 388, 86–97.
- Pain, O., Glanville, K.P., Hagenaars, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., Coleman, J.R.I., Rimfeld, K., Breen, G., Plomin, R., Folkersen, L., Lewis, C.M., 2021. Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* 17, e1009021.
- Pang, K., Wang, L., Wang, W., Zhou, J., Cheng, C., Han, K., Zoghbi, H.Y., Liu, Z., 2020. Coexpression enrichment analysis at the single-cell level reveals convergent defects in neural progenitor cells and their cell-type transitions in neurodevelopmental disorders. *Genome Res.* 30, 835–848.
- Pardiñas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S.E., Bishop, S., Cameron, D., Hamshere, M.L., Han, J., Hubbard, L., Lynham, A., Mantripragada, K., Rees, E., MacCabe, J.H., McCarroll, S.A., Baune, B.T., Breen, G., Byrne, E.M., Dannlowski, U., Eley, T.C., Hayward, C., Martin, N.G., McIntosh, A.M., Plomin, R., Porteous, D.J., Wray, N.R., Caballero, A., Geschwind, D.H., Huckins, L.M., Ruderfer, D.M., Santiago, E., Sklar, P., Stahl, E.A., Won, H., Agerbo, E., Als, T.D., Andreassen, O.A., Bækvad-Hansen, M., Mortensen, P.B., Pedersen, C.B., Børghlum, A.D., Bybjerg-Grauholm, J., Djurovic, S., Durmishi, N., Pedersen, M.G., Golimbet, V., Grove, J., Hougaard, D.M., Mattheisen, M., Molden, E., Mors, O., Nordentoft, M., Pejovic-Milovancevic, M., Sigurdsson, E., Silagadze, T., Hansen, C.S., Stefansson, K., Stefansson, H., Steinberg, S., Tosato, S., Werge, T., GERAD1 Consortium, CRESTAR Consortium, Collier, D.A., Rujescu, D., Kirov, G., Owen, M.J., O'Donovan, M.C., Walters, J.T.R., 2018. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* 50, 381–389.
- Parellada, E., Gassó, P., 2021. Glutamate and microglia activation as a driver of dendritic apoptosis: a core pathophysiological mechanism to understand schizophrenia. *Transl. Psychiatry* 11, 271.
- Phan, B.N., Bohlen, J.F., Davis, B.A., Ye, Z., Chen, H.Y., Mayfield, B., Sripathy, S.R., Cerceo Page, S., Campbell, M.N., Smith, H.L., Gallop, D., Kim, H., Thaxton, C.L., Simon, J.M., Burke, E.E., Shin, J.H., Kennedy, A.J., Sweatt, J.D., Philpot, B.D., Jaffe, A.E., Maher, B.J., 2020. A myelin-related transcriptomic profile is shared by Pitt-Hopkins syndrome models and human autism spectrum disorder. *Nat. Neurosci.* 23, 375–385.

- Pieraccioli, M., Nicolai, S., Antonov, A., Somers, J., Malewicz, M., Melino, G., Raschellà, G., 2016. ZNF281 contributes to the DNA damage response by controlling the expression of XRCC2 and XRCC4. *Oncogene* 35, 2592–2601.
- Pitman, K.A., Ricci, R., Gasperini, R., Beasley, S., Pavez, M., Charlesworth, J., Foa, L., Young, K.M., 2020. The voltage-gated calcium channel CaV1.2 promotes adult oligodendrocyte progenitor cell survival in the mouse corpus callosum but not motor cortex. *Glia* 68, 376–392.
- Poletti, M., Gebhardt, E., Pelizza, L., Preti, A., Raballo, A., 2020. Looking at intergenerational risk factors in schizophrenia spectrum disorders: New frontiers for early vulnerability identification? *Front. Psychiatry* 11, 566683.
- PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J.A., Vaccarino, F.M., Farnham, P.J., Crawford, G.E., Jaffe, A.E., Pinto, D., Dracheva, S., Geschwind, D.H., Mill, J., Nairn, A.C., Abyzov, A., Pochareddy, S., Prabhakar, S., Weissman, S., Sullivan, P.F., State, M.W., Weng, Z., Peters, M.A., White, K.P., Gerstein, M.B., Amiri, A., Armoskus, C., Ashley-Koch, A.E., Bae, T., Beckel-Mitchener, A., Berman, B.P., Coetzee, G.A., Coppola, G., Francoeur, N., Fromer, M., Gao, R., Grennan, K., Herstein, J., Kavanagh, D.H., Ivanov, N.A., Jiang, Y., Kitchen, R.R., Kozlenkov, A., Kundakovic, M., Li, M., Li, Z., Liu, S., Mangravite, L.M., Mattei, E., Markenscoff-Papadimitriou, E., Navarro, F.C.P., North, N., Omberg, L., Panchision, D., Parikshak, N., Poschmann, J., Price, A.J., Purcaro, M., Reddy, T.E., Roussos, P., Schreiner, S., Scuderi, S., Sebra, R., Shibata, M., Shieh, A.W., Skarica, M., Sun, W., Swarup, V., Thomas, A., Tsuji, J., van Bakel, H., Wang, D., Wang, Y., Wang, K., Werling, D.M., Willsey, A.J., Witt, H., Won, H., Wong, C.C.Y., Wray, G.A., Wu, E.Y., Xu, X., Yao, L., Senthil, G., Lehner, T., Sklar, P., Sestan, N., 2015. The PsychENCODE project. *Nat. Neurosci.* 18, 1707–1712.
- Psychiatric GWAS Consortium Coordinating Committee, Cichon, S., Craddock, N., Daly, M., Faraone, S.V., Gejman, P.V., Kelsoe, J., Lehner, T., Levinson, D.F., Moran, A., Sklar, P., Sullivan, P.F., 2009. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am. J. Psychiatry* 166, 540–556.
- Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O’Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., Duncan, L., Stahl, E., Genovese, G., Fernández, E., Collins, M.O., Komiyama, N.H., Choudhary, J.S., Magnusson, P.K.E., Banks, E., Shakir, K., Garimella, K., Fennell, T., DePristo, M., Grant, S.G.N., Haggarty, S.J., Gabriel, S., Scolnick, E.M., Lander, E.S., Hultman, C.M., Sullivan, P.F., McCarroll, S.A., Sklar, P., 2014. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185–190.

- Rajarajan, P., Borrman, T., Liao, W., Schrode, N., Flaherty, E., Casiño, C., Powell, S., Yashaswini, C., LaMarca, E.A., Kassim, B., Javidfar, B., Espeso-Gil, S., Li, A., Won, H., Geschwind, D.H., Ho, S.M., MacDonald, M., Hoffman, G.E., Roussos, P., Zhang, B., Hahn, C.G., Weng, Z., Brennand, K.J., Akbarian, S., 2018. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* 362.
- Ramaswami, G., Won, H., Gandal, M.J., Haney, J., Wang, J.C., Wong, C.C.Y., Sun, W., Prabhakar, S., Mill, J., Geschwind, D.H., 2020. Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism. *Nat. Commun.* 11, 4873.
- Rees, E., Creeth, H.D.J., Hwu, H.G., Chen, W.J., Tsuang, M., Glatt, S.J., Rey, R., Kirov, G., Walters, J.T.R., Holmans, P., Owen, M.J., O'Donovan, M.C., 2021. Schizophrenia, autism spectrum disorders and developmental disorders share specific disruptive coding mutations. *Nat. Commun.* 12, 5353.
- Rees, E., Moskvina, V., Owen, M.J., O'Donovan, M.C., Kirov, G., 2011. De novo rates and selection of schizophrenia-associated copy number variants. *Biol. Psychiatry* 70, 1109–1114.
- van Rheenen, W., van der Spek, R.A.A., Bakker, M.K., van Vugt, J.J.F.A., Hop, P.J., Zwamborn, R.A.J., de Klein, N., Westra, H.J., Bakker, O.B., Deelen, P., Shireby, G., Hannon, E., Moisse, M., Baird, D., Restuadi, R., Dolzhenko, E., Dekker, A.M., Gawor, K., Westeneng, H.J., Tazelaar, G.H.P., van Eijk, K.R., Kooyman, M., Byrne, R.P., Doherty, M., Heverin, M., Al Khleifat, A., Iacoangeli, A., Shatunov, A., Ticozzi, N., Cooper-Knock, J., Smith, B.N., Gromicho, M., Chandran, S., Pal, S., Morrison, K.E., Shaw, P.J., Hardy, J., Orrell, R.W., Sendtner, M., Meyer, T., Başak, N., van der Kooi, A.J., Ratti, A., Fogh, I., Gellera, C., Lauria, G., Corti, S., Cereda, C., Sproviero, D., D'Alfonso, S., Sorarù, G., Siciliano, G., Filosto, M., Padovani, A., Chiò, A., Calvo, A., Moglia, C., Brunetti, M., Canosa, A., Grassano, M., Beghi, E., Pupillo, E., Logroscino, G., Nefussy, B., Osmanovic, A., Nordin, A., Lerner, Y., Zabari, M., Gotkine, M., Baloh, R.H., Bell, S., Vourc'h, P., Corcia, P., Couratier, P., Millicamps, S., Meininger, V., Salachas, F., Mora Pardina, J.S., Assialioui, A., Rojas-García, R., Dion, P.A., Ross, J.P., Ludolph, A.C., Weishaupt, J.H., Brenner, D., Freischmidt, A., Bensimon, G., Brice, A., Durr, A., Payan, C.A.M., Saker-Delye, S., Wood, N.W., Topp, S., Rademakers, R., Tittmann, L., Lieb, W., Franke, A., Ripke, S., Braun, A., Kraft, J., Whiteman, D.C., Olsen, C.M., Uitterlinden, A.G., Hofman, A., Rietschel, M., Cichon, S., Nöthen, M.M., Amouyel, P., SLALOM Consortium, PARALS Consortium, SLAGEN Consortium, SLAP Consortium, Traynor, B.J., Singleton, A.B., Mitne Neto, M., Cauchi, R.J., Ophoff, R.A., Wiedau-Pazos, M., Lomen-Hoerth, C., van Deerlin, V.M., Grosskreutz,

J., Roediger, A., Gaur, N., Jörk, A., Barthel, T., Theele, E., Ilse, B., Stubendorff, B., Witte, O.W., Steinbach, R., Hübner, C.A., Graff, C., Brylev, L., Fominykh, V., Demeshonok, V., Ataulina, A., Rogelj, B., Koritnik, B., Zidar, J., Ravnik-Glavač, M., Glavač, D., Stević, Z., Drory, V., Povedano, M., Blair, I.P., Kiernan, M.C., Benyamin, B., Henderson, R.D., Furlong, S., Mathers, S., McCombe, P.A., Needham, M., Ngo, S.T., Nicholson, G.A., Pamphlett, R., Rowe, D.B., Steyn, F.J., Williams, K.L., Mather, K.A., Sachdev, P.S., Henders, A.K., Wallace, L., de Carvalho, M., Pinto, S., Petri, S., Weber, M., Rouleau, G.A., Silani, V., Curtis, C.J., Breen, G., Glass, J.D., Brown, Jr, R.H., Landers, J.E., Shaw, C.E., Andersen, P.M., Groen, E.J.N., van Es, M.A., Pasterkamp, R.J., Fan, D., Garton, F.C., McRae, A.F., Davey Smith, G., Gaunt, T.R., Eberle, M.A., Mill, J., McLaughlin, R.L., Hardiman, O., Kenna, K.P., Wray, N.R., Tsai, E., Runz, H., Franke, L., Al-Chalabi, A., Van Damme, P., van den Berg, L.H., Veldink, J.H., 2021. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat. Genet.* 53, 1636–1648.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.C., Pfening, A.R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R.A., Shores, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R.D., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Hansen, R.S., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.H., Feizi, S., Karlic, R., Kim, A.R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthal, K.T., Sinnott-Armstrong, N.A., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.A., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J.M., Li, W., Marra, M.A., McManus, M.T., Sunyaev, S., Thomson, J.A., Tlsty, T.D., Tsai, L.H., Wang, W., Waterland, R.A., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.A., Wang, T., Kellis, M., 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330.

Robinson, N., Bergen, S.E., 2021. Environmental risk factors for schizophrenia and bipolar disorder and their relationship to genetic risk: Current knowledge and future directions. *Front. Genet.* 12, 686666.

- Roussos, P., Katsel, P., Davis, K.L., Siever, L.J., Haroutunian, V., 2012. A system-level transcriptomic analysis of schizophrenia using postmortem brain tissue samples. *Arch. Gen. Psychiatry* 69, 1205–1213.
- Roussos, P., Mitchell, A.C., Voloudakis, G., Fullard, J.F., Pothula, V.M., Tsang, J., Stahl, E.A., Georgakopoulos, A., Ruderfer, D.M., Charney, A., Okada, Y., Siminovitch, K.A., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P.K., Plenge, R.M., Raychaudhuri, S., Fromer, M., Purcell, S.M., Brennand, K.J., Robakis, N.K., Schadt, E.E., Akbarian, S., Sklar, P., 2014. A role for noncoding variation in schizophrenia. *Cell Rep.* 9, 1417–1429.
- Ruzicka, W.B., Mohammadi, S., Fullard, J.F., Davila-Velderrain, J., Subburaju, S., Tso, D.R., Hourihan, M., Jiang, S., Lee, H.C., Bendl, J., PsychENCODE Consortium, Voloudakis, G., Haroutunian, V., Hoffman, G.E., Roussos, P., Kellis, M., 2022. Single-cell multi-cohort dissection of the schizophrenia transcriptome. *medRxiv*, 10.1101/2022.08.31.22279406.
- Savage, J.E., Jansen, P.R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C.A., Nagel, M., Awasthi, S., Barr, P.B., Coleman, J.R.I., Grasby, K.L., Hammerschlag, A.R., Kaminski, J.A., Karlsson, R., Krapohl, E., Lam, M., Nygaard, M., Reynolds, C.A., Trampush, J.W., Young, H., Zabaneh, D., Hägg, S., Hansell, N.K., Karlsson, I.K., Linnarsson, S., Montgomery, G.W., Muñoz-Manchado, A.B., Quinlan, E.B., Schumann, G., Skene, N.G., Webb, B.T., White, T., Arking, D.E., Avramopoulos, D., Bilder, R.M., Bitsios, P., Burdick, K.E., Cannon, T.D., Chiba-Falek, O., Christoforou, A., Cirulli, E.T., Congdon, E., Corvin, A., Davies, G., Deary, I.J., DeRosse, P., Dickinson, D., Djurovic, S., Donohoe, G., Conley, E.D., Eriksson, J.G., Espeseth, T., Freimer, N.A., Giakoumaki, S., Giegling, I., Gill, M., Glahn, D.C., Hariri, A.R., Hatzimanolis, A., Keller, M.C., Knowles, E., Koltai, D., Konte, B., Lahti, J., Le Hellard, S., Lencz, T., Liewald, D.C., London, E., Lundervold, A.J., Malhotra, A.K., Melle, I., Morris, D., Need, A.C., Ollier, W., Palotie, A., Payton, A., Pendleton, N., Poldrack, R.A., Rääkkönen, K., Reinvang, I., Roussos, P., Rujescu, D., Sabb, F.W., Scult, M.A., Smealand, O.B., Smyrnis, N., Starr, J.M., Steen, V.M., Stefanis, N.C., Straub, R.E., Sundet, K., Tiemeier, H., Voineskos, A.N., Weinberger, D.R., Widen, E., Yu, J., Abecasis, G., Andreassen, O.A., Breen, G., Christiansen, L., Debrabant, B., Dick, D.M., Heinz, A., Hjerling-Leffler, J., Ikram, M.A., Kendler, K.S., Martin, N.G., Medland, S.E., Pedersen, N.L., Plomin, R., Polderman, T.J.C., Ripke, S., van der Sluis, S., Sullivan, P.F., Vrieze, S.I., Wright, M.J., Posthuma, D., 2018. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* 50, 912–919.

- Schaid, D.J., Chen, W., Larson, N.B., 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19, 491–504.
- Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* 43, 969–976.
- Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427.
- Shan, Y., Zhao, J., Zheng, Y., Guo, S., Schrodi, S.J., He, D., 2023. Understanding the function of the GABAergic system and its potential role in rheumatoid arthritis. *Front. Immunol.* 14, 1114350.
- Shen-Orr, S.S., Tibshirani, R., Khatry, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., Butte, A.J., 2010. Cell type-specific gene expression differences in complex tissues. *Nat. Methods* 7, 287–289.
- Shi, Z.X., Chen, Z.C., Zhong, J.Y., Hu, K.H., Zheng, Y.F., Chen, Y., Xie, S.Q., Bo, X.C., Luo, F., Tang, C., Xiao, C.L., Liu, Y.Z., 2023. High-throughput and high-accuracy single-cell RNA isoform analysis using PacBio circular consensus sequencing. *Nat. Commun.* 14, 2631.
- Shishido, R., Kunii, Y., Hino, M., Izumi, R., Nagaoka, A., Hayashi, H., Kakita, A., Tomita, H., Yabe, H., 2023. Evidence for increased DNA damage repair in the post-mortem brain of the high stress-response group of schizophrenia. *Front. Psychiatry* 14, 1183696.
- Shiwaku, H., Katayama, S., Kondo, K., Nakano, Y., Tanaka, H., Yoshioka, Y., Fujita, K., Tamaki, H., Takebayashi, H., Terasaki, O., Nagase, Y., Nagase, T., Kubota, T., Ishikawa, K., Okazawa, H., Takahashi, H., 2022. Autoantibodies against NCAM1 from patients with schizophrenia cause schizophrenia-related behavior and changes in synapses in mice. *Cell Rep Med* 3, 100597.
- Sieberts, S.K., Perumal, T.M., Carrasquillo, M.M., Allen, M., Reddy, J.S., Hoffman, G.E., Dang, K.K., Calley, J., Ebert, P.J., Eddy, J., Wang, X., Greenwood, A.K., Mostafavi, S., CommonMind Consortium (CMC), The AMP-AD Consortium, Omberg, L., Peters, M.A., Logsdon, B.A., De Jager, P.L., Ertekin-Taner, N., Mangravite, L.M., 2020. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Sci Data* 7, 340.
- Siletti, K., Hodge, R., Mossi Albiach, A., Lee, K.W., Ding, S.L., Hu, L., Lönnerberg, P., Bakken, T., Casper, T., Clark, M., Dee, N., Gloe, J., Hirschstein, D., Shapovalova,

N.V., Keene, C.D., Nyhus, J., Tung, H., Yanny, A.M., Arenas, E., Lein, E.S., Linnarsson, S., 2023. Transcriptomic diversity of cell types across the adult human brain. *Science* 382, eadd7046.

Singh, T., Kurki, M.I., Curtis, D., Purcell, S.M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., Pietiläinen, O., Gerety, S.S., Ayub, M., Blyth, M., Cole, T., Collier, D., Coomber, E.L., Craddock, N., Daly, M.J., Danesh, J., DiForti, M., Foster, A., Freimer, N.B., Geschwind, D., Johnstone, M., Joss, S., Kirov, G., Körkkö, J., Kuismin, O., Holmans, P., Hultman, C.M., Iyegbe, C., Lönnqvist, J., Männikkö, M., McCarroll, S.A., McGuffin, P., McIntosh, A.M., McQuillin, A., Moilanen, J.S., Moore, C., Murray, R.M., Newbury-Ecob, R., Ouwehand, W., Paunio, T., Prigmore, E., Rees, E., Roberts, D., Sambrook, J., Sklar, P., St Clair, D., Veijola, J., Walters, J.T.R., Williams, H., Swedish Schizophrenia Study, INTERVAL Study, DDD Study, UK10 K Consortium, Sullivan, P.F., Hurles, M.E., O'Donovan, M.C., Palotie, A., Owen, M.J., Barrett, J.C., 2016. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* 19, 571–577.

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

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., Huang, J., Du, Y., Lin, L., Dong, Z., Yang, L.,

- Liu, X., Jiang, H., Xu, X., Wang, J., Yang, H., Bolund, L., Mardinoglu, A., Zhang, C., von Feilitzen, K., Lindskog, C., Pontén, F., Luo, Y., Hökfelt, T., Uhlén, M., Mulder, J., 2020. An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* 367.
- Skene, N.G., Bryois, J., Bakken, T.E., Breen, G., Crowley, J.J., Gaspar, H.A., Giusti-Rodriguez, P., Hodge, R.D., Miller, J.A., Muñoz-Manchado, A.B., O'Donovan, M.C., Owen, M.J., Pardiñas, A.F., Ryge, J., Walters, J.T.R., Linnarsson, S., Lein, E.S., Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Sullivan, P.F., Hjerling-Leffler, J., 2018. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* 50, 825–833.
- Skene, N.G., Grant, S.G.N., 2016. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* 10, 16.
- Slyper, M., Porter, C.B.M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., Vigneau, S., Jané-Valbuena, J., Tickle, T.L., Napolitano, S., Su, M.J., Patel, A.G., Karlstrom, A., Gritsch, S., Nomura, M., Waghray, A., Gohil, S.H., Tsankov, A.M., Jerby-Arnon, L., Cohen, O., Klughammer, J., Rosen, Y., Gould, J., Nguyen, L., Hofree, M., Tramontozzi, P.J., Li, B., Wu, C.J., Izar, B., Haq, R., Hodi, F.S., Yoon, C.H., Hata, A.N., Baker, S.J., Suvà, M.L., Bueno, R., Stover, E.H., Clay, M.R., Dyer, M.A., Collins, N.B., Matulonis, U.A., Wagle, N., Johnson, B.E., Rotem, A., Rozenblatt-Rosen, O., Regev, A., 2020. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* 26, 792–802.
- Smedler, E., Louhivuori, L., Romanov, R.A., Masini, D., Dehnisch Ellström, I., Wang, C., Caramia, M., West, Z., Zhang, S., Rebellato, P., Malmersjö, S., Brusini, I., Kanatani, S., Fisone, G., Harkany, T., Uhlén, P., 2022. Disrupted *Cacna1c* gene expression perturbs spontaneous Ca²⁺ activity causing abnormal brain development and increased anxiety. *Proc. Natl. Acad. Sci. U. S. A.* 119.
- Sos, B.C., Fung, H.L., Gao, D.R., Osothprarop, T.F., Kia, A., He, M.M., Zhang, K., 2016. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol.* 17, 20.
- Stauffer, E.M., Bethlehem, R.A.I., Warriier, V., Murray, G.K., Romero-Garcia, R., Seidnitz, J., Bullmore, E.T., 2021. Grey and white matter microstructure is associated with polygenic risk for schizophrenia. *Mol. Psychiatry* 26, 7709–7718.

- Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O.P.H., Mors, O., Mortensen, P.B., Sigurdsson, E., Gustafsson, O., Nyegaard, M., Tuulio-Henriksson, A., Ingason, A., Hansen, T., Suvisaari, J., Lonnqvist, J., Paunio, T., Børglum, A.D., Hartmann, A., Fink-Jensen, A., Nordentoft, M., Hougaard, D., Norgaard-Pedersen, B., Böttcher, Y., Olesen, J., Breuer, R., Möller, H.J., Giegling, I., Rasmussen, H.B., Timm, S., Mattheisen, M., Bitter, I., Réthelyi, J.M., Magnusdottir, B.B., Sigmundsson, T., Olason, P., Masson, G., Gulcher, J.R., Haraldsson, M., Fossdal, R., Thorgeirsson, T.E., Thorsteinsdottir, U., Ruggeri, M., Tosato, S., Franke, B., Strengman, E., Kiemenev, L.A., Genetic Risk and Outcome in Psychosis (GROUP), Melle, I., Djurovic, S., Abramova, L., Kaleda, V., Sanjuan, J., de Frutos, R., Bramon, E., Vassos, E., Fraser, G., Etinger, U., Picchioni, M., Walker, N., Toulopoulou, T., Need, A.C., Ge, D., Yoon, J.L., Shianna, K.V., Freimer, N.B., Cantor, R.M., Murray, R., Kong, A., Golimbet, V., Carracedo, A., Arango, C., Costas, J., Jönsson, E.G., Terenius, L., Agartz, I., Petursson, H., Nöthen, M.M., Rietschel, M., Matthews, P.M., Muglia, P., Peltonen, L., St Clair, D., Goldstein, D.B., Stefansson, K., Collier, D.A., 2009. Common variants conferring risk of schizophrenia. *Nature* 460, 744–747.
- Stegle, O., Parts, L., Durbin, R., Winn, J., 2010. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* 6, e1000770.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550.
- Sullivan, P.F., 2013. Questions about DISC1 as a genetic risk factor for schizophrenia. *Mol. Psychiatry* 18, 1050–1052.
- Sullivan, P.F., Agrawal, A., Bulik, C.M., Andreassen, O.A., Børglum, A.D., Breen, G., Cichon, S., Edenberg, H.J., Faraone, S.V., Gelernter, J., Mathews, C.A., Nievergelt, C.M., Smoller, J.W., O'Donovan, M.C., Psychiatric Genomics Consortium, 2018. Psychiatric genomics: An update and an agenda. *Am. J. Psychiatry* 175, 15–27.
- Sullivan, P.F., Kendler, K.S., Neale, M.C., 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* 60, 1187–1192.
- Sun, Z., Zhao, L., Bo, Q., Mao, Z., He, Y., Jiang, T., Li, Y., Wang, C., Li, R., 2021. Brain-Specific oxysterols and risk of schizophrenia in clinical High-Risk subjects and patients with schizophrenia. *Front. Psychiatry* 12, 711734.

- Sutton, G.J., Poppe, D., Simmons, R.K., Walsh, K., Nawaz, U., Lister, R., Gagnon-Bartsch, J.A., Voineagu, I., 2022. Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nat. Commun.* 13, 1358.
- Sytnyk, V., Leshchyn'ska, I., Schachner, M., 2017. Neural cell adhesion molecules of the immunoglobulin superfamily regulate synapse formation, maintenance, and function. *Trends Neurosci.* 40, 295–308.
- Tanaka, S., 2006. Dopaminergic control of working memory and its relevance to schizophrenia: a circuit dynamics perspective. *Neuroscience* 139, 153–171.
- Tang, X., Jaenisch, R., Sur, M., 2021. The role of GABAergic signalling in neurodevelopmental disorders. *Nat. Rev. Neurosci.* 22, 290–307.
- Taylor, M., Jauhar, S., 2019. Are we getting any better at staying better? The long view on relapse and recovery in first episode nonaffective psychosis and schizophrenia. *Therapeutic Advances in Psychopharmacology* 9, 2045125319870033.
- Teare, M.D., Santibañez Koref, M.F., 2014. Linkage analysis and the study of mendelian disease in the era of whole exome and genome sequencing. *Brief. Funct. Genomics* 13, 378–383.
- The ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.
- The International HapMap Consortium, 2003. The international HapMap project. *Nature* 426, 789–796.
- Toker, L., Nido, G.S., Tzoulis, C., 2023. Not every estimate counts - evaluation of cell composition estimation approaches in brain bulk tissue data. *Genome Med.* 15, 41.
- Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.Y., Dennison, C.A., Hall, L.S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J.C., Koopmans, F., Magnusson, S., Richards, A.L., Sidorenko, J., Wu, Y., Zeng, J., Grove, J., Kim, M., Li, Z., Voloudakis, G., Zhang, W., Adams, M., Agartz, I., Atkinson, E.G., Agerbo, E., Al Eissa, M., Albus, M., Alexander, M., Alizadeh, B.Z., Alptekin, K., Als, T.D., Amin, F., Arolt, V., Arrojo, M., Athanasiu, L., Azevedo, M.H., Bacanu, S.A., Bass, N.J., Begemann, M., Belliveau, R.A., Bene, J., Benyamin, B., Bergen, S.E., Blasi, G., Bobes, J., Bonassi, S., Braun, A., Bressan, R.A., Bromet, E.J., Bruggeman, R., Buckley, P.F., Buckner, R.L., Bybjerg-Grauholm, J., Cahn, W., Cairns, M.J., Calkins, M.E., Carr, V.J., Castle, D., Catts, S.V., Chamberbert, K.D., Chan, R.C.K., Chaumette, B., Cheng, W., Cheung, E.F.C., Chong, S.A., Cohen, D., Consoli, A., Cordeiro, Q., Costas, J., Curtis, C., Davidson, M., Davis, K.L.,

de Haan, L., Degenhardt, F., DeLisi, L.E., Demontis, D., Dickerson, F., Dikeos, D., Dinan, T., Djurovic, S., Duan, J., Ducci, G., Dudbridge, F., Eriksson, J.G., Fañanás, L., Faraone, S.V., Fiorentino, A., Forstner, A., Frank, J., Freimer, N.B., Fromer, M., Frustaci, A., Gadelha, A., Genovese, G., Gershon, E.S., Giannitelli, M., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J.I., González Peñas, J., González-Pinto, A., Gopal, S., Gratten, J., Green, M.F., Greenwood, T.A., Guillin, O., Gülöksüz, S., Gur, R.E., Gur, R.C., Gutiérrez, B., Hahn, E., Hakonarson, H., Haroutunian, V., Hartmann, A.M., Harvey, C., Hayward, C., Henskens, F.A., Herms, S., Hoffmann, P., Howrigan, D.P., Ikeda, M., Iyegbe, C., Joa, I., Julià, A., Kähler, A.K., Kam-Thong, T., Kamatani, Y., Karachanak-Yankova, S., Kebir, O., Keller, M.C., Kelly, B.J., Khrunin, A., Kim, S.W., Klovin, J., Kondratiev, N., Konte, B., Kraft, J., Kubo, M., Kučinskas, V., Kučinskiene, Z.A., Kusumawardhani, A., Kuzelova-Ptackova, H., Landi, S., Lazzeroni, L.C., Lee, P.H., Legge, S.E., Lehrer, D.S., Lencer, R., Lerer, B., Li, M., Lieberman, J., Light, G.A., Limborska, S., Liu, C.M., Lönnqvist, J., Loughland, C.M., Lubinski, J., Luykx, J.J., Lynham, A., Macek, Jr, M., Mackinnon, A., Magnusson, P.K.E., Maher, B.S., Maier, W., Malaspina, D., Mallet, J., Marder, S.R., Marsal, S., Martin, A.R., Martorell, L., Mattheisen, M., McCarley, R.W., McDonald, C., McGrath, J.J., Medeiros, H., Meier, S., Meleg, B., Melle, I., Meshulam-Gately, R.I., Metspalu, A., Michie, P.T., Milani, L., Milanova, V., Mitjans, M., Molden, E., Molina, E., Molto, M.D., Mondelli, V., Moreno, C., Morley, C.P., Muntané, G., Murphy, K.C., Myin-Germeys, I., Nenadić, I., Nestadt, G., Nikitina-Zake, L., Noto, C., Nuechterlein, K.H., O'Brien, N.L., O'Neill, F.A., Oh, S.Y., Olincy, A., Ota, V.K., Pantelis, C., Papadimitriou, G.N., Parellada, M., Paunio, T., Pellegrino, R., Periyasamy, S., Perkins, D.O., Pfulmann, B., Pietiläinen, O., Pimm, J., Porteous, D., Powell, J., Quattrone, D., Quested, D., Radant, A.D., Rampino, A., Rapaport, M.H., Rautanen, A., Reichenberg, A., Roe, C., Roffman, J.L., Roth, J., Rothermundt, M., Rutten, B.P.F., Saker-Delye, S., Salomaa, V., Sanjuan, J., Santoro, M.L., Savitz, A., Schall, U., Scott, R.J., Seidman, L.J., Sharp, S.I., Shi, J., Siever, L.J., Sigurdsson, E., Sim, K., Skarabis, N., Slominsky, P., So, H.C., Sobell, J.L., Söderman, E., Stain, H.J., Steen, N.E., Steixner-Kumar, A.A., Stögmann, E., Stone, W.S., Straub, R.E., Streit, F., Strengman, E., Stroup, T.S., Subramaniam, M., Sugar, C.A., Suvisaari, J., Svrakic, D.M., Swerdlow, N.R., Szatkiewicz, J.P., Ta, T.M.T., Takahashi, A., Terao, C., Thibaut, F., Toncheva, D., Tooney, P.A., Torretta, S., Tosato, S., Tura, G.B., Turetsky, B.I., Üçok, A., Vaaler, A., van Amelsvoort, T., van Winkel, R., Veijola, J., Waddington, J., Walter, H., Waterreus, A., Webb, B.T., Weiser, M., Williams, N.M., Witt, S.H., Wormley, B.K., Wu, J.Q., Xu, Z., Yolken, R., Zai, C.C., Zhou, W., Zhu, F., Zimprich, F., Atbaşoğlu, E.C., Ayub, M., Benner, C., Bertolino, A., Black, D.W., Bray, N.J., Breen, G., Buccola, N.G., Byerley, W.F., Chen, W.J., Cloninger, C.R., Crespo-Facorro, B., Donohoe, G., Freedman, R., Galletly, C., Gandal, M.J., Gennarelli, M., Hougaard, D.M., Hwu, H.G., Jablensky, A.V., Mc-

Carroll, S.A., Moran, J.L., Mors, O., Mortensen, P.B., Müller-Myhsok, B., Neil, A.L., Nordentoft, M., Pato, M.T., Petryshen, T.L., Pirinen, M., Pulver, A.E., Schulze, T.G., Silverman, J.M., Smoller, J.W., Stahl, E.A., Tsuang, D.W., Vilella, E., Wang, S.H., Xu, S., Indonesia Schizophrenia Consortium, PsychENCODE, Psychosis Endophenotypes International Consortium, SynGO Consortium, Adolfsson, R., Arango, C., Baune, B.T., Belangero, S.I., Børghlum, A.D., Braff, D., Bramon, E., Buxbaum, J.D., Campion, D., Cervilla, J.A., Cichon, S., Collier, D.A., Corvin, A., Curtis, D., Forti, M.D., Domenici, E., Ehrenreich, H., Escott-Price, V., Esko, T., Fanous, A.H., Gareeva, A., Gawlik, M., Gejman, P.V., Gill, M., Glatt, S.J., Golimbet, V., Hong, K.S., Hultman, C.M., Hyman, S.E., Iwata, N., Jönsson, E.G., Kahn, R.S., Kennedy, J.L., Khusnutdinova, E., Kirov, G., Knowles, J.A., Krebs, M.O., Laurent-Levinson, C., Lee, J., Lencz, T., Levinson, D.F., Li, Q.S., Liu, J., Malhotra, A.K., Malhotra, D., McIntosh, A., McQuillin, A., Menezes, P.R., Morgan, V.A., Morris, D.W., Mowry, B.J., Murray, R.M., Nimgaonkar, V., Nöthen, M.M., Ophoff, R.A., Paciga, S.A., Palotie, A., Pato, C.N., Qin, S., Riettschel, M., Riley, B.P., Rivera, M., Rujescu, D., Saka, M.C., Sanders, A.R., Schwab, S.G., Serretti, A., Sham, P.C., Shi, Y., St Clair, D., Stefánsson, H., Stefansson, K., Tsuang, M.T., van Os, J., Vawter, M.P., Weinberger, D.R., Werge, T., Wildenauer, D.B., Yu, X., Yue, W., Holmans, P.A., Pocklington, A.J., Roussos, P., Vassos, E., Verhage, M., Visscher, P.M., Yang, J., Posthuma, D., Andreassen, O.A., Kendler, K.S., Owen, M.J., Wray, N.R., Daly, M.J., Huang, H., Neale, B.M., Sullivan, P.F., Ripke, S., Walters, J.T.R., O'Donovan, M.C., Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2022. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* 604, 502–508.

Tu, S., Akhtar, M.W., Escorihuela, R.M., Amador-Arjona, A., Swarup, V., Parker, J., Zaremba, J.D., Holland, T., Bansal, N., Holohan, D.R., Lopez, K., Ryan, S.D., Chan, S.F., Yan, L., Zhang, X., Huang, X., Sultan, A., McKercher, S.R., Ambasadhan, R., Xu, H., Wang, Y., Geschwind, D.H., Roberts, A.J., Terskikh, A.V., Rissman, R.A., Masliah, E., Lipton, S.A., Nakanishi, N., 2017. NitroSynapsin therapy for a mouse MEF2C haploinsufficiency model of human autism. *Nat. Commun.* 8, 1488.

Tucholski, J., Simmons, M.S., Pinner, A.L., Haroutunian, V., McCullumsmith, R.E., Meador-Woodruff, J.H., 2013. Abnormal n-linked glycosylation of cortical AMPA receptor subunits in schizophrenia. *Schizophr. Res.* 146, 177–183.

Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., 2021. Genome-wide association studies. *Nature Reviews Methods Primers* 1, 1–21.

Uliana, D.L., Zhu, X., Gomes, F.V., Grace, A.A., 2022. Using animal models for the

- studies of schizophrenia and depression: The value of translational models for treatment and prevention. *Front. Behav. Neurosci.* 16, 935320.
- Ursini, G., Punzi, G., Chen, Q., Marenco, S., Robinson, J.F., Porcelli, A., Hamilton, E.G., Mitjans, M., Maddalena, G., Begemann, M., Seidel, J., Yanamori, H., Jaffe, A.E., Berman, K.F., Egan, M.F., Straub, R.E., Colantuoni, C., Blasi, G., Hashimoto, R., Rujescu, D., Ehrenreich, H., Bertolino, A., Weinberger, D.R., 2018. Convergence of placenta biology and genetic risk for schizophrenia. *Nat. Med.* 24, 792–801.
- Varese, F., Smeets, F., Drukker, M., Lieverse, R., Lataster, T., Viechtbauer, W., Read, J., van Os, J., Bentall, R.P., 2012. Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective- and cross-sectional cohort studies. *Schizophr. Bull.* 38, 661–671.
- Vaucher, J., Keating, B.J., Lasserre, A.M., Gan, W., Lyall, D.M., Ward, J., Smith, D.J., Pell, J.P., Sattar, N., Paré, G., Holmes, M.V., 2018. Cannabis use and risk of schizophrenia: a mendelian randomization study. *Mol. Psychiatry* 23, 1287–1292.
- Velmeshev, D., Perez, Y., Yan, Z., Valencia, J.E., Castaneda-Castellanos, D.R., Wang, L., Schirmer, L., Mayer, S., Wick, B., Wang, S., Nowakowski, T.J., Paredes, M., Huang, E.J., Kriegstein, A.R., 2023. Single-cell analysis of prenatal and postnatal human cortical development. *Science* 382, eadf0834.
- Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J., 2017. 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22.
- Wallace, C., 2020. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* 16, e1008720.
- Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haike-Kains, B., Goldenberg, A., 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337.
- Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., Xu, M., Gandal, M.J., Lou, S., Zhang, J., Park, J.J., Yan, C., Rhie, S.K., Manakongtreecheep, K., Zhou, H., Nathan, A., Peters, M., Mattei, E., Fitzgerald, D., Brunetti, T., Moore, J., Jiang, Y., Girdhar, K., Hoffman, G.E., Kalayci, S., Gümüş, Z.H., Crawford, G.E., PsychENCODE Consortium, Roussos, P., Akbarian, S., Jaffe, A.E., White, K.P., Weng, Z., Sestan, N., Geschwind, D.H., Knowles, J.A., Gerstein, M.B., 2018. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362.

- Wang, J., Devlin, B., Roeder, K., 2020. Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression. *Bioinformatics* 36, 782–788.
- Wang, J., Roeder, K., Devlin, B., 2021. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.* 31, 1807–1818.
- Wang, X., Park, J., Susztak, K., Zhang, N.R., Li, M., 2019. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* 10, 380.
- Wang, Z., Li, P., Chi, D., Wu, T., Mei, Z., Cui, G., 2017. Association between c-reactive protein and risk of schizophrenia: An updated meta-analysis. *Oncotarget* 8, 75445–75454.
- Wei, J., Hemmings, G.P., 2000. The NOTCH4 locus is associated with susceptibility to schizophrenia. *Nat. Genet.* 25, 376–377.
- Weinberger, D.R., 1987. Implications of normal brain development for the pathogenesis of schizophrenia. *Arch. Gen. Psychiatry* 44, 660–669.
- Wen, C., Margolis, M., Dai, R., Zhang, P., Przytycki, P.F., Vo, D.D., Bhattacharya, A., Matoba, N., Jiao, C., Kim, M., Tsai, E., Hoh, C., Aygün, N., Walker, R.L., Chatzinakos, C., Clarke, D., Pratt, H., Consortium, P., Peters, M.A., Gerstein, M., Daskalakis, N.P., Weng, Z., Jaffe, A.E., Kleinman, J.E., Hyde, T.M., Weinberger, D.R., Bray, N.J., Sestan, N., Geschwind, D.H., Roeder, K., Gusev, A., Pasaniuc, B., Stein, J.L., Love, M.I., Pollard, K.S., Liu, C., Gandal, M.J., 2023. Cross-ancestry, cell-type-informed atlas of gene, isoform, and splicing regulation in the developing human brain. *medRxiv* .
- Woolway, G.E., Smart, S.E., Lynham, A.J., Lloyd, J.L., Owen, M.J., Jones, I.R., Walters, J.T.R., Legge, S.E., 2022. Schizophrenia polygenic risk and experiences of childhood adversity: A systematic review and meta-analysis. *Schizophr. Bull.* 48, 967–980.
- Xue, Y., Tong, L., LiuAnwei Liu, F., Liu, A., Zeng, S., Xiong, Q., Yang, Z., He, X., Sun, Y., Xu, C., 2019. Tumor-infiltrating M2 macrophages driven by specific genomic alterations are associated with prognosis in bladder cancer. *Oncol. Rep.* 42, 581–594.
- Yu, G., Wang, L.G., He, Q.Y., 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31, 2382–2383.
- Zeng, B., Bendl, J., Deng, C., Lee, D., Misir, R., Reach, S.M., Kleopoulos, S.P., Auluck, P., Marengo, S., Lewis, D.A., Haroutunian, V., Ahituv, N., Fullard, J.F., Hoffman, G.E., Roussos, P., 2023. Genetic regulation of cell-type specific chromatin accessibility shapes the etiology of brain diseases.

- Zeng, B., Bendl, J., Kosoy, R., Fullard, J.F., Hoffman, G.E., Roussos, P., 2022. Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat. Genet.* 54, 161–169.
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137.
- Zhang, Z., Zhao, Y., 2022a. Progress on the roles of MEF2C in neuropsychiatric diseases. *Mol. Brain* 15, 8.
- Zhang, Z., Zhao, Y., 2022b. Progress on the roles of MEF2C in neuropsychiatric diseases. *Mol. Brain* 15, 8.
- Zheng, F., Cui, Y., Yan, H., Liu, B., Jiang, T., 2016. The effects of a genome-wide supported variant in the CACNA1C gene on cortical morphology in schizophrenia patients and healthy subjects. *Sci. Rep.* 6, 34298.
- Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., Xu, X., Tang, F., Zhang, J., Qiao, J., Wang, X., 2018. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 555, 524–528.
- Zhou, R., Zhang, J., Zeng, D., Sun, H., Rong, X., Shi, M., Bin, J., Liao, Y., Liao, W., 2019. Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I-III colon cancer. *Cancer Immunol. Immunother.* 68, 433–442.
- Zhou, X., Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824.
- Zhu, K., Bendl, J., Rahman, S., Vicari, J.M., Coleman, C., Clarence, T., Latouche, O., Tsankova, N.M., Li, A., Brennand, K.J., Lee, D., Yuan, G.C., Fullard, J.F., Roussos, P., 2022. Multi-omic profiling of the developing human cerebral cortex at the single cell level.
- Zhu, K., Bendl, J., Rahman, S., Vicari, J.M., Coleman, C., Clarence, T., Latouche, O., Tsankova, N.M., Li, A., Brennand, K.J., Lee, D., Yuan, G.C., Fullard, J.F., Roussos, P., 2023. Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Sci Adv* 9, eadg3754.

Appendices

List of Awards

- Oral Presentation Award, Galway Neuroscience Centre Annual Research Day, December 2022
- Beckman Fund Scholarship, University of Galway, 2022
- Finalist, Early Career Investigator Program - Oral Presentation, World Congress of Psychiatric Genetics, October 2021
- Young Investigator Award for Best Postgraduate Oral Presentation, Irish Society of Human Genetics, September 2021

List of Presentations

Oral Presentations

- Irish Computational Biology and Genomics Symposium, University of Galway, December 2023
- Galway Neuroscience Centre Annual Research Day, University of Galway, December 2022
- Irish Computational Biology and Genomics Symposium, Online, February 2022
- Galway Neuroscience Centre Annual Research Day,, University of Galway, December 2021
- World Congress of Psychiatric Genetics, Online, October 2021
- Irish Society of Human Genetics, Online, September 2021

Poster Presentation

- Irish Computational Biology and Genomics Symposium, University of Galway, December 2023
- World Congress of Psychiatric Genetics, Montreal, October 2023
- Irish Society of Human Genetics, Galway, September 2023
- Biology of Genomes, Cold Spring Harbour, May 2022
- American Society of Human Genetics, Online, October 2021
- Irish Computational Biology and Genomics Symposium, Online, December 2020

Manuscripts arising from this thesis

Two manuscripts are currently being finalised based on the work discussed in Chapters 3 and 4 of this thesis.