Sheffield Hallam University

Identification of Rare Variants in Schizophrenia-relevant Pathways: A Computational Analysis

SLAY, Oliver Kieron

Available from the Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/35932/

A Sheffield Hallam University thesis

This thesis is protected by copyright which belongs to the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Please visit https://shura.shu.ac.uk/35932/ and http://shura.shu.ac.uk/information.html for further details about copyright and re-use permissions.

Identification of Rare Variants in Schizophrenia-relevant Pathways: A Computational Analysis

Oliver Kieron Slay

A thesis submitted in partial fulfilment of the requirements of Sheffield Hallam University for the Degree of Doctor of Philosophy

Faculty of Health and Wellbeing

June 2024

Candidate Declaration

I hereby declare that:

1. I have not been enrolled for another award of the University, or other academic or professional organisation, whilst undertaking my research degree.

2. None of the material contained in the thesis has been used in any other submission for an academic award.

3. I certify that this thesis is my own work. The use of all published or other sources of material consulted have been properly and fully acknowledged.

4. The work undertaken towards the thesis has been conducted in accordance with the SHU Principles of Integrity in Research and the SHU Research Ethics Policy, and ethics approval has been granted for all research studies in the thesis.

5. The word count of the thesis is …… 56,000

Name	Oliver Slay
Date of Submission	June 2024
Award	Doctor of Philosophy
Research Institute	Industry and Innovation
Director(s) of Studies	Dr Caroline Dalton

Acknowledgements

Firstly, I would like to express my gratitude to my supervisors, Dr Caroline Dalton, and Dr Alessandra Princivalle for all their supervision and encouragement during my PhD training. I appreciate the time and patience that has been extended to me, and their insightful feedback during the write-up of this work.

I would also like to thank Dr Adrian Hall for his guidance and support during my third year.

I would like to express my deepest gratitude to my lifelong partner, Dr Kirsten Bartlett, who has supported me through this time and enabled the opportunity to complete my studies. Also, to my mother for her patience and support during these past years.

I would like to thank Professor JZ Young, who inspired me throughout my school years in his study of the brain, following him into neuroscience in my later years.

Acknowledgement for the Common Fund (CF) Genotype-Tissue Expression Project (GTEx)

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses

described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v8.p2.

Acknowledgement for the Sweden-Schizophrenia Exome Sequencing Project

The datasets used for the analysis described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000473.v2.p2. Samples used for data analysis were provided by the Swedish Cohort Collection supported by the NIMH grant R01MH077139, the Sylvan C. Herman Foundation, the Stanley Medical Research Institute and The Swedish Research Council (grants 2009-4959 and 2011-4659). Support for the exome sequencing was provided by the NIMH Grand Opportunity grant RCMH089905, the Sylvan C. Herman Foundation, a grant from the Stanley Medical Research Institute and multiple gifts to the Stanley Center for Psychiatric Research at the Broad Institute of MIT and Harvard.

Abstract

Schizophrenia (SCZ) is a complex neuropsychiatric disorder with a strong genetic component. A comprehensive understanding of the genes and pathways involved remains elusive. The aim of this study was to identify rare genetic variants contributing to SCZ by leveraging a gene set with higher expression in brain regions relevant to SCZ and focusing on rare damaging variants identified in patients to refine the search for SCZ-associated genes.

Using RNA-Sequencing data from healthy donors obtained from the Genotype-Tissue Expression (GTEx) project, an interregional differential expression analysis was performed between two target regions (frontal cortex and hippocampus) robustly implicated in schizophrenia (SCZ) and two contrast regions (cerebellar hemisphere and substantia nigra) with less compelling evidence of involvement in SCZ based on existing literature. For each of the four regions, RNA-Seq data from 25 donors were analysed, totalling 100 datasets. This comprehensive approach identified 1076 genes with more specific expression in the frontal cortex and hippocampus compared to control regions.

Variant analysis was performed using whole-exome sequencing data from 3500 SCZ and 4800 control exomes from the Swedish-Schizophrenia Exomes project. Ultra-rare damaging variants (dURVs) were identified by excluding variants found in 125,748 exomes in gnomAD v2.1.1, and filtering to loss-of-function, or variants predicted to be damaging by five *in silico* prediction tools. Focussing exclusively on the identified 1076 highly expressed genes, 4,398 dURVs were identified exclusive to 2,444 SCZ cases and 1,066 SCZ cases had no dURVs. These findings implicate an oligogenic model of SCZ at the level of the individual, while highly polygenic at the level of the population.

Variant aggregation and pathway analysis found pathway heterogeneity between SCZ cases involving any of signal transduction, the complement system, histone modification, expression regulation, and WNT, RHO GTPase, or calcium signalling. The most common gene with dURVs was CSMD1 found in 1.2% of cases.

Table of Contents

Candidate Declaration i

Acknowledgements ii

Abstract iv

Table of Contents v

List of Tables xiii

List of Figures xv

Abbreviations xviii

1 Introduction 1

1.1 Schizophrenia 1

1.1.1 Epidemiology 1

1.1.2 Symptoms 1

1.1.3 Diagnosis 2

1.1.4 Causes 3

1.1.5 Genes and Environment 4

1.2 Searching for Genetic Causes of SCZ 4

1.2.1 Study Approaches 6

1.2.2 Family and Linkage Studies 6

1.2.3 Genome-wide Association Studies (GWAS) 8

1.2.4 Exome/Genome Sequencing Studies 11

1.3 Gap in the Literature 20

1.4 Overall Aim and Main objectives 21

2 Ranking of Brain Regions in Schizophrenia 22

2.1 Background 22

2.1.1 Schizophrenia 22

2.1.2 Symptoms 22

2.1.3 Brain Study Modalities in Schizophrenia 23

2.1.4 Genotype-Tissue Expression (GTEx) Project 25
2.1.5 Aim and Objectives 26
2.2 Methods 28
2.2.1 Selection of GTEx Brain Regions 28
2.2.2 Matching GTEx brain regions to brain atlas regions 30
2.2.3 Selection of SCZ brain studies from the literature 31
2.2.4 GTEx Brain Regions ranking from most affected to least affected 32
2.3 Results 33
2.3.1 Selection of GTEx Brain Regions 33
2.3.2 Anatomical location of the GTEx Brain Region sample sites 34
2.3.3 Selection of Brain Imaging Literature 46
2.3.4 Cortex thickness and subcortical volumes are affected by multiple factors 47
2.3.5 Ranking of the GTEx Brain Regions 48
2.3.6 Frontal Cortex and Hippocampus are the two most affected brain regions 49
2.3.7 Cerebellum and Substantia nigra as the two least affected brain regions. 53
2.4 Discussion 53
2.4.1 Most affected: Hippocampus and Frontal Cortex 53
2.4.2 Least affected: Cerebellar Hemisphere and Substantia nigra 54
3 RNA-Seq Differential Expression Analysis of Brain Regions 57
3.1 Background 57
3.1.1 Gene Expression 57
3.1.2 RNA-Sequencing 58
3.1.3 RNA-Seq Differential Expression Analysis 58
3.1.4 The Genotype-Tissue Expression (GTEx) project 59
3.1.5 Software 60

3.1.6 Aim and Objectives 65

3.2 Methods 66

3.2.1 Study Populations and RNA-Seq Samples 66

3.2.2 GTEx – Principal Components Analysis for all brain regions 70

3.2.3 Quality Control - Analysis of single dataset, SRR821690 - Hippocampus 71

3.2.4 Compare Aligner Software – STAR and Rsubread - Analysis of 20 BrainSeq samples 72

3.2.5 GTEx Frontal Cortex vs Cerebellar Hemisphere – Differential Expression Analysis 74

3.2.6 Main Differential Expression Analysis - Cortex/Hippocampus vs Cerebellar Hemisphere/Substantia Nigra 75

3.3 Results 80

3.3.1 Exploratory Methodological Findings 80

3.3.2 Main Comparative Analysis of Four Brain Regions – Cortex and Hippocampus vs Cerebellar Hemisphere and Substantia Nigra 86

3.3.3 Genes with higher expression in the Frontal Cortex compared to control region samples 88

3.3.4 Genes with higher expression in the Hippocampus compared to control region samples 92

3.3.5 Protein-coding genes with higher expression in the Frontal Cortex and Hippocampus 93

3.3.6 Validation against genes associated with SCZ in Open Targets 95

3.3.7 Mean expression of CO/HP-genes across all GTEx brain regions 97

3.3.8 Expression Weighted Celltype Enrichment in single-cell cortex and hippocampus 98

3.3.9 Gene set enrichment analysis for CO- and HP-genes 99

3.4 Discussion 103

3.4.1 Exploratory Methodological Findings 104

3.4.2 Main Differential Expression Analysis in Four Brain Regions 107
3.4.3 Validation against genes associated with SCZ in Open Targets 109
3.4.4 Differentially expressed genes showed greater expression across target regions 110
3.4.5 Genes showed enrichment in key cell types from cortex and hippocampus 111
3.4.6 CO/HP-genes were enriched in neuronal and metabolic pathways 112
3.5 Limitations and Further Work 112
4 Variant Filtering of SCZ Exome data 115
4.1 Overview 115
4.2 Background 115
4.2.1 Genomic Variants 115
4.2.2 Genome-Wide Association Studies in Schizophrenia 116
4.2.3 Whole Exome Sequencing of SCZ cases 116
4.2.4 Reference Population Databases 117
4.2.5 Pathogenicity / Variant Effects 118
4.2.6 Variant Aggregation in Pathways – systems biology 120
4.2.7 Aims and Objectives 121
4.3 Methods 122
4.3.1 Sample Population and Data 122
4.3.2 Subsampling of 100 subjects 126
4.3.3 Annotation of variants using BCFtools Consequences (`csq`) 127
4.3.4 Annotation of variants using gnomAD Exomes v2.1.1 128
4.3.5 Normalisation of variants 129
4.3.6 Filtering to rare variants 130
4.3.7 Pathogenicity of Remaining Variants – Ensembl Variant Effects Predictor (VEP) 130

4.3.8 Binning variants to Genes and Pathways 131
4.4 Results 132
4.4.1 Filtering of Swedish Exome data to CO/HP-genes in 100 SCZ patients 132
4.4.2 Annotation with Coding Consequences considering multinucleotide variants (MNVs) 135
4.4.3 Annotation with gnomAD Exomes and filtering to Rare Variants 138
4.4.4 Variant Effect Predictor 140
4.4.5 Aggregation to Gene Regions and Subregions 141
4.4.6 Aggregation in Pathways 144
4.5 Discussion 149
4.5.1 Variant Burden in the CO-/HP-genes in an example group of 100 cases with SCZ 150
4.5.2 Coding consequences 151
4.5.3 Identifying Rare Variants in CO-/HP-genes - not found in gnomAD Exomes 153
4.5.4 Variant Effect Prediction including SIFT and PolyPhen2 scores 154
4.5.5 Aggregation to Gene Regions and Subregions 155
4.5.6 Aggregation in Pathways 156
4.5.7 Limitations 158
4.5.8 Further work 159
5 Grouped Variant Case-Control Analysis 160
5.1 Introduction 160
5.1.1 Schizophrenia 160
5.1.2 Whole Exome Sequencing data 160
5.1.3 Main Aim and Objectives 160
5.2 Methods 161
5.2.1 Data 161

5.2.2 Sample selection and subsampling into groups of 100 subjects and controls 162
5.2.3 Annotation of variants using BCFtools Consequences (csq) 163
5.2.4 Annotation of variants using gnomAD v2.1.1 163
5.2.5 Normalisation of variants 164
5.2.6 Pathogenicity of Rare Variants – Ensembl Variant Effect Prediction 164
5.2.7 Pre-processing of Variants in R 167
5.2.8 Filter variants to SCZ-only, predicted-pathogenic variants 169
5.2.9 Comparison with Genome-Wide Association Studies (GWAS) 171
5.2.10 Variant Pathway Analysis – consistent pathways across groups 171
5.3 Results 172
5.3.1 Analysis of SCZ and CTL groups 172
5.3.2 Effect of filtering with gnomAD Exomes 173
5.3.3 Analysis of Consequences for all groups 177
5.3.4 Variant Effect Predictor 177
5.3.5 Reduction to ultra-rare, predicted damaging variants only affecting SCZ cases 183
5.3.6 GWAS risk loci overlap 187
5.3.7 Gene ontologies (GO) affecting SCZ groups 191
5.4 Discussion 206
5.4.1 Filtering and Analysing Variants in CO/HP-genes 207
5.4.2 Identifying Consistent Biological Pathways Enriched for Rare Variants 209
5.4.3 Overlapping GWAS loci and rare damaging variants (RDVs) 211
5.4.4 Potassium channels and shared associations with other disorders 212
5.4.5 Conclusion and Further Work 212
6 Discussion 213
6.1 Introduction 213

6.2 Increased exclusivity of genes to the frontal cortex and hippocampal gene sets compared to studies 214
6.2.1 Discovering disease-relevant gene sets 215
6.2.2 The selection of control regions 217
6.2.3 Focus on four brain regions 218
6.2.4 Use of bulk tissue RNA-Seq 218
6.2.5 Differential Region Expression finds more region-specific transcripts 219
6.2.6 Future work 219
6.3 Rare damaging variants exclusive to a cohort of schizophrenia cases 220
6.3.1 SCZ is oligogenic for rare damaging variants in highly expressed genes 221
6.3.2 Some important genes are not represented by this gene set 221
6.3.3 Combining scores from appropriate *in silico* prediction tools 223
6.4 Highlighted potential role of rare variants in common pathways 225
6.4.1 Signal transduction is the most consistently affected pathway 225
6.4.2 The complement system pathway 226
6.4.3 Epigenetic and expression regulatory pathways affect almost 10% of cases 226
6.4.4 Combinations of affected pathways may lead to different neuropsychiatric diseases 227
6.4.5 Future work 227
6.5 Conclusion 228
6.5.1 Personalised medicine is required for effective drug treatment 229
7 References 230
8 Appendix 286
8.1 Hardy Scale classification 286
8.2 Differential Expression Analysis Phenotype data – 100 samples 286
8.3 GO enrichment analyses for CO- and HP-genes 289

8.3.1 GO Biological Process enrichment 289
8.3.2 GO Cellular Component enrichment 292
8.3.3 GO Molecular Function enrichment 294
8.3.4 KEGG Pathways enrichment 296
8.3.5 Reactome 2022 Pathways enrichment 298
8.4 Deleterious SIFT and Damaging PolyPhen2 Rare Variants in SCZ000 299
8.5 Genes with Rare Damaging Variants in 3500 Schizophrenia Cases 301
8.5.1 Genes with RDVs in SCZ cases only 301
8.5.2 Genes in which SCZ cases have over twice the number of RDVs 302
8.5.3 Genes in which SCZ cases have more RDV but less than double 302
8.5.4 Genes have at least one RDV in one case but more in controls 302

List of Tables

Table 1.1 Key WES, GWAS studies on patients with SCZ 18

Table 2.1 Brain study modalities 24

Table 2.2 GTEx brain regions 29

Table 2.3 Anatomical location of GTEx sample sites in the brain 35

Table 2.4 Imaging studies reviewed 46

Table 3.1 RNA-Seq alignment software 62

Table 3.2 Differentially expressed genes for each contrast 89

Table 3.3 Functional Annotation Clustering of 943 CO-genes 100

Table 3.4 Functional Annotation Cluster of 536 HP-genes 102

Table 4.1 gnomAD Metrics used for annotating variants in each group 129

Table 4.2 Breakdown of variant types with consequence of missense or worse from the 100 SCZ samples 136

Table 4.3 Rare multinucleotide variant in EXTL1 in the 100 SCZ samples 137

Table 4.4 Rare multinucleotide variants in LPA in the 100 SCZ samples 137

Table 4.5 Top 100 genes ordered by counts of variants with worse consequences 138

Table 4.6 VEP summary data for SCZ000 141

Table 4.7 Top 20 Genes by count of rare SNVs 142

Table 4.8 Top 20 10Kb gene subregions by count of rare variants 143

Table 4.9 Top 20 20Kb gene subregions by count of rare variants 144

Table 4.10 Top 20 100 Kb gene subregions by count of rare variants 144

Table 4.11 Top 20 GO biological processes overrepresented for genes with RVs in SCZ000 146

Table 4.12 Top 20 GO-Slim biological processes overrepresented for genes with RVs in SCZ000 147

Table 4.13 Top 20 GO molecular functions overrepresented for genes with RVs in SCZ000 148

Table 4.14 PANTHER Pathways overrepresented for genes with RVs in SCZ000 149

Table 4.15 Reactome Pathways overrepresented for genes with RVs in SCZ000 149

Table 5.1 Subjects passing the Coverage criterion 162

Table 5.2 Plugins used from Ensembl VEP 165

Table 5.3 CO/HP-genes with rare damaging variants by allele count 185

Table 5.4 Variants overlapping GWAS loci 188

Table 5.5 Top 30 GWAS loci with variants in CO/HP-genes in SCZ groups 189

Table 5.6 Table showing CSMD1 variants surrounding GWAS loci 191

Table 5.7 KEGG enrichment analysis of 629 genes with RDVs involved in GO Protein Binding (MF) 202

Table 5.8 Top 10 genes with RDVs in cases enriched in KEGG pathways from 'protein binding' 203

Table 5.9 Reactome enrichment analysis of 104 genes with RDVs involved in GO Extracellular Spaces (CC) 205

Table 8.1 GO Biological Process terms enriched for CO/HP Genes 289

Table 8.2 GO Cellular Component terms enriched for CO- and HP-Genes 293

Table 8.3 GO Molecular Function enriched for CO/HP Genes 294

Table 8.4 KEGG 2021 Human pathways enriched for CO/HP-genes 297

Table 8.5 Reactome 2022 pathways enriched for CO/HP-genes 298

Table 8.6 Rare variants in SCZ000 - deleterious by SIFT or damaging by PolyPhen2 ... 300

List of Figures

Figure 2.1 Brain regions sampled in the GTEx Project 26

Figure 2.2 GTEx Regions of Interest - Left Hemisphere 30

Figure 2.3 Ventral Dorsolateral Prefrontal Cortex (BA9/BA46) (red, top right) 37

Figure 2.4 GTEx cortical regions 38

Figure 2.5 GTEx regions: Anterior Cingulate Cortex (BA24), Caudate nucleus, and Nucleus accumbens 39

Figure 2.6 GTEx regions: Putamen, Amygdala and Hypothalamus 41

Figure 2.7 GTEx regions: Hippocampus and Substantia nigra 43

Figure 2.8 Cerebellar lateral hemisphere cortical structure – Nissl-stained 45

Figure 2.9 Evidence matrix showing modalities reviewed 49

Figure 2.10 Haijma et al. (2013) effect sizes for subcortical structures 50

Figure 2.11 ENIGMA and COCORO: Changes in subcortical volumes between patients with SCZ and healthy controls 51

Figure 2.12 Longitudinal changes in subcortical volumes between patients with SCZ and healthy controls (Olabi) 52

Figure 3.1 Main steps involved in RNA-Seq Differential Expression Analysis 60

Figure 3.2 Summary of alignment strategies tested on BrainSeq samples 72

Figure 3.3 Principal Components Analysis of all GTEx Brain Regions 81

Figure 3.4 Quality scores across bases in each read pair for SRR821690 82

Figure 3.5 Mapping rates for all BrainSeq samples using STAR and Rsubread 84

Figure 3.6 PCA for 162 Frontal Cortex and Cerebellar Hemisphere samples 86

Figure 3.7 PCA plot for all 4 brain regions 88

Figure 3.8 Enhanced Volcano plot - expression in Frontal Cortex vs control regions 90

Figure 3.9 Bar chart of the division of biotypes in the brain region comparisons 91

Figure 3.10 Enhanced Volcano plot - expression in Hippocampus vs control regions ...93

Figure 3.11 Venn diagram of protein-coding genes from the 4 brain region contrasts.95

Figure 3.12 Venn diagram showing count of Open Targets platform SCZ-associated genes in the 4 brain region contrasts 96

Figure 3.13 PCA plot coloured by mean expression of CO/HP-genes per GTEx sample 97

Figure 3.14 Expression Weighted Celltype Enrichment in Cortex and Hippocampus 99

Figure 3.15 GTExPortal: Expression PCA for all GTEx Cerebellar Hemisphere samples106

Figure 4.1 Workflow for selection of variants in the CO/HP-genes from 100 SCZ patients 133

Figure 4.2 SNVs per sample in SCZ000 134

Figure 4.3 INDELs per sample in SCZ000 134

Figure 4.4 Singletons per sample in SCZ000 135

Figure 4.5 Rare SNVs (AF<0.01 and AC<=5 in gnomAD) per sample in SCZ000 139

Figure 4.6 Rare INDELs (AF<0.01 and AC<=5 in gnomAD) per sample in SCZ000 139

Figure 4.7 Rare singletons (AF<0.01 and AC<=5 in gnomAD) per sample in SCZ000 ...140

Figure 4.8 Ensembl VEP summary of variants in group SCZ000 141

Figure 5.1 Plot of average counts of genotypes per subject across 35 SCZ groups and 48 CTL groups 173

Figure 5.2 Plot of average counts of genotypes for rare variants only across all groups 174

Figure 5.3 Box plot of SCZ and CTL groups in chromosomes 1-6.....175

Figure 5.4 Box plot of SCZ and CTL groups in chromosomes 7-12 176

Figure 5.5 Box plot of SCZ and CTL groups in chromosomes 13-18 176

Figure 5.6 Box plot of SCZ and CTL groups in chromosomes 19-22 and X 177

Figure 5.7 Count of pathogenic PolyPhen-2 HumDiv scores across all groups 179

Figure 5.8 Count of SIFT scores per group 180

Figure 5.9 Count of REVEL scores per group.....181

Figure 5.10 Count of dbscSNV AdaBoost scores per group 182

Figure 5.11 Count of dbscSNV Random Forest scores per group 182

Figure 5.12 Bar chart showing the distribution of rare damaging variants per SCZ group 184

Figure 5.13 Top 50 GO Biological Process terms containing RDVs 193

Figure 5.14 Genes with RDVs affecting G-protein coupled receptor pathways.....195

Figure 5.15 Genes with RDVs in Monoatomic ion transport processes 197

Figure 5.16 Genes with RDVs involved in Axon Guidance 199

Figure 5.17 Top 50 Molecular Function terms by count of variants.....200

Figure 5.18 Top 50 Cellular Components by count of variants.....204

Figure 6.1 Regional expression of AMPAR subunits 222

Abbreviations

ACC	Anterior cingulate cortex
AMPA	α-amino-3-hydoxy-5-methyl-isoxazolepropionic acid
ASD	Autism Spectrum Disorder
BPD	Bipolar disorder
BSS	Biospecimen Source Site
CDCV	Common Disease – Common Variant
CDRV	Common Disease – Rare Variant (also MRV)
CNV	Copy Number Variant
COCORO	Cognitive Genetics Collaborative Research Organisation
DALY	Disability-Adjusted Life Years
dbGaP	Database of Genotypes and Phenotypes
DNA	Deoxyribonucleic acid
DSM	Diagnostic and Statistical Manual of Mental Disorders
DTI	Diffusion Tensor Imaging
dURV	Damaging Ultra-Rare Variant
DWI	Diffusion Weighted Imaging
ENIGMA	Enhancing Neuro Imaging Genetics through Meta-Analysis
eQTL	Expression Quantitative Trait Loci
EWCE	Expression Weighted Celltype Enrichment
ExAC	Exome Aggregation Consortium (superseded by gnomAD)
fMRI	Functional Magnetic Resonance Imaging
fMRS	Functional Magnetic Resonance Spectroscopy
GBD	Global Burden of Disease

gnomAD	Genomic Aggregation Database
GO	Gene Ontology
GSMA	Genome Search Meta-Analysis method
GTEx	The Genotype-Tissue Expression Project
GWAS	Genome-wide Association Study
ISC	International Schizophrenia Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MBEB	Miami Brain Endowment Bank
MNV	Multinucleotide Variant
MRI	Magnetic Resonance Imaging
MRS	Magnetic Resonance Spectroscopy
MRV	Multiple Rare Variants
NAcc	Nucleus accumbens
NDD	Neurodevelopmental Disorders
NIH	National Institute of Health
NMDA	N-methyl-D-aspartate
PANSS	Positive and Negative Syndrome Scale
PET	Positron Emission Tomography
PGC	Psychiatric Genomics Consortium
PM	Post-mortem
PSD	Post-synaptic Density
RNA	Ribonucleic acid

SCHEMA	Schizophrenia Exome Sequencing Meta-Analysis consortium
SCZ	Schizophrenia
SGENE	Schizophrenia Genetics Consortium
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SPECT	Single-positron Emission Computed Tomography
sQTL	Splicing Quantitative Trait Loci
URV	Ultra-Rare Variant
WES	Whole Exome Sequencing
WGE	Whole Genome Sequencing

1 Introduction

1.1 Schizophrenia

1.1.1 Epidemiology

Schizophrenia (SCZ) is a severe and complex neuropsychiatric disorder with a significant global presence and a high economic burden that extends beyond the affected individual. The Global Burden of Disease (GBD) study in 2019 (Solmi et al., 2023) estimated an age-standardized prevalence of 287.41 per 100,000 individuals, translating to nearly 24 million cases worldwide. This highlights the significant global burden of the disorder, further emphasized by the estimated 184.15 Disability-Adjusted Life Years (DALY) per 100,000 individuals attributed to SCZ in 2019. Incidence of SCZ is highest in 15-25 age group and prevalence is greatest in the 35-45 age group, affecting individuals at the most productive era of life. SCZ significantly impairs daily functioning, with its impact extending far beyond the affected individual, often leading to difficulties maintaining employment, social relationships, and independent living. Families and caregivers often bear the significant burden of providing support and managing the daily challenges associated with the disease. Furthermore, SCZ contributes substantially to the economic burden on society through healthcare costs and lost productivity. In the United States, there was an estimated excess economic burden in 2019 of $343.2 billion in direct costs (health care and non-health care) and indirect costs, for example, caregiving, unemployment, and premature mortality (Kadakia et al., 2022; Solmi et al., 2023).

1.1.2 Symptoms

Schizophrenia symptoms traditionally fall into three categories: positive, negative, and cognitive. Also proposed, has been a five-factor model aligning with the Positive and Negative Syndrome Scale (PANSS): positive, negative, disorganised (often exchanged for cognitive), excited, and depressed, although these models are in a constant state of development and modification informed by advances in research (Giordano et al., 2022; McCutcheon et al., 2023; Wallwork et al., 2012). Positive symptoms describe phenotypes that are not found in healthy individuals, most commonly, delusions, hallucinations, feelings of grandiosity and having unusual thoughts. Negative

symptoms describe normal behaviours and functions that are reduced in SCZ patients, possibly from the early stages of the disease, and include: social withdrawal, affective flattening, poor thinking inferred from deficits in speech and language (alogia), anhedonia, and avolition. Cognitive (or 'disorganised') symptoms include deficits in attention, executive function and working memory (McCutcheon et al., 2023). In the five-factor model, the 'Excited' factor includes poor impulse control, excitement, uncooperativeness, and the 'Depressed' factor includes symptoms of anxiety, guilt, and depression posturing (Wallwork et al., 2012). Patients with SCZ exhibit some but not all the symptoms which makes diagnosis challenging.

1.1.3 Diagnosis

The diagnosis of SCZ is a complex process that requires a comprehensive understanding of the individual's symptoms and their impact on daily life. The process typically involves a comprehensive psychiatric evaluation, medical history analysis, and physical examination. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5), published by the American Psychiatric Association, provides the standard criteria for diagnosing SCZ. According to DSM-5, the presence of two or more of the symptoms for a significant portion of time during a 1-month period indicates a diagnosis of SCZ and continuation of symptoms for at least 6 months: delusions, hallucinations, disorganized speech, disorganized or catatonic behaviour, and negative symptoms, with at least one symptom from the first three (Tandon et al., 2013). However, these symptoms alone are not sufficient for a diagnosis. Clinicians must also rule out other mental health disorders and determine whether substance abuse or a medical condition might be causing the symptoms. Furthermore, the symptoms must cause significant distress or difficulty in functioning in work, relationships, or self-care, and at least some symptoms must have been present for six months. Furthermore, the diagnostic process can be complicated by the fact that SCZ symptoms can overlap with those of other psychiatric disorders, and many patients with SCZ also have co-occurring mental or behavioural health disorders. Therefore, a thorough and comprehensive evaluation is crucial for an accurate diagnosis (Jauhar et al., 2022). In recent years, there has been a growing interest in the potential of neuroimaging techniques, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), to aid in the diagnosis of SCZ. While these techniques are promising,

they are currently used primarily for research purposes in SCZ and are not yet a standard part of the diagnostic process (Samson et al., 2022; Varambally et al., 2012).

The diagnosis of SCZ is not only critical for treatment planning but also has significant implications for patients and their families. It can affect many aspects of their lives, including educational and vocational opportunities, relationships, and self-perception. Therefore, accurate diagnosis and early intervention are crucial for improving the long-term outcomes of individuals with SCZ (Correll, Schooler, & Robinson, 2020). Functional impairment is evaluated to ascertain the impact of symptoms on daily life. This structured approach aims to differentiate SCZ from other psychiatric conditions and facilitate accurate diagnosis and subsequent treatment planning (Jauhar et al., 2022; Tandon et al., 2013).

1.1.4 Causes

Understanding the specific neurobiological mechanisms underlying SCZ has been a continuous pursuit, with several hypotheses emerging over time, for example, the dopamine, glutamate, neurodevelopmental hypotheses, and more recently, the synaptic hypothesis (Howes & Onwordi, 2023; Jauhar et al., 2022; Nakamura & Takata, 2023). The dopamine hypothesis proposed that an overactivity of the dopamine system, particularly in the mesolimbic pathway, contributed to the positive symptoms of SCZ, such as hallucinations and delusions. This hypothesis arose from observations that certain drugs (for example, amphetamines) that increase dopamine levels can induce psychosis-like symptoms, and that antipsychotic medications often target dopamine receptors. As research progressed, the limitations of the dopamine hypothesis became apparent (Weinberger & Laruelle, 2001), as it could not fully explain the diverse range of symptoms in SCZ, including cognitive deficits and negative symptoms. The glutamate hypothesis emerged, proposing that dysfunction within the glutamatergic system, the brain's primary excitatory neurotransmitter system, plays a crucial role in the disorder. This hypothesis is supported by evidence of altered glutamate signalling in various brain regions of individuals with SCZ, and the ability of certain N-methyl-D-aspartate (NMDA) receptor antagonists to induce some psychotic symptoms (Jauhar et al., 2022).

The search for an underlying cause has increasingly focussed on the synapse, the junction where neurons communicate. Emerging evidence suggests a potential link

between abnormal synaptic function and genetic risk factors. Signals from one neuron pass across synapses from the presynaptic axon to a postsynaptic neuron at any of the thousands of dendritic spines, structures on the dendrites of neurons. The synaptic density is the space between the pre- and postsynaptic membranes. It is a highly regulated interface, served by multiple proteins that help organise and maintain the structure. Synapse formation, maintenance and elimination is a continually active process throughout life. Studies have observed reduced numbers of dendritic spines – crucial structures for receiving information – and a role for microglia, immune cells in the brain, in pruning these connections (Breitmeyer et al., 2023). Furthermore, studies have identified genes associated with the complement system, a network involved in immune function, whose activity is elevated in SCZ (Frankle et al., 2003; Howes & Onwordi, 2023; Woo et al., 2020). This growing body of evidence highlights the potential role of synaptic dysfunction and genetics in the development of SCZ.

1.1.5 Genes and Environment

SCZ's origins are not rooted in a single cause for all cases, but rather a complex interplay between genetic and environmental factors. Twin, adoption, and family studies have been instrumental in revealing this intricate relationship. The increased prevalence of SCZ among individuals with affected family members points towards a significant genetic influence. Heritability, a statistical measure, estimates the proportion of a trait or complex disease like SCZ that can be attributed to genetic differences in a population. Studies suggest heritability estimates for SCZ range from 60-80% (Sullivan et al., 2003). While genetics undoubtedly play a crucial role, the remaining 20-40% of risk cannot be solely attributed to them, highlighting the substantial impact of environmental factors. Examples of such environmental elements include prenatal complications and childhood trauma (Paquin et al., 2021), exposure to viruses or toxins during development (Cheslack-Postava & Brown, 2022), urbanicity (Haddad et al., 2015), and substance abuse (Stilo & Murray, 2019). These factors can interact with an individual's genetic predisposition, potentially increasing their vulnerability to developing SCZ.

1.2 Searching for Genetic Causes of SCZ

With its highly heritable component and complexity of symptoms, unravelling the specific genetic underpinnings of SCZ has been a significant challenge. Over the last 30

years, numerous studies have identified an abundance of genetic loci associated with an increased risk for SCZ. This genetic heterogeneity depicts a highly diverse genetic landscape, suggesting that SCZ does not stem from one single cause but rather from a multitude of genetic variations interacting together or with environmental factors (Nakamura & Takata, 2023; Owen et al., 2023; Tandon et al., 2024). Unlike Mendelian diseases with a single causative gene, SCZ has proven to be a complex genetic disorder. This genetic heterogeneity, where multiple genetic factors likely contribute to the risk of developing SCZ can be further categorised into different types of heterogeneity:

- Locus heterogeneity: Different genes in various chromosomal locations might contribute to SCZ in different individuals.
- Allelic heterogeneity: Different mutations within the same gene can lead to the same or similar phenotypes.
- Model heterogeneity: How these genetic factors interact (gene-gene interactions) with each other, or with the environment (gene-environment) may vary between individuals with SCZ.

In the latter, model heterogeneity, there are two models that have been used as a basis for the search for, and to explain the genetic causes in SCZ: The Common Disease, Common Variant (CDCV) model is a hypothesis that suggests a common disease, like SCZ, is a result of or modified by common genetic variants within a population. According to this model, genetic variants with relatively low penetrance but appreciable frequency in the general population contribute significantly to the genetic susceptibility to common diseases. In other words, the CDCV model proposes that common diseases are influenced by a combination of several common genetic variants, each with a small to moderate effect on disease risk (Lander, 1996). This stems from the observation that there are frequently haplotypic variations (alleles) commonly found in humans at different levels in different populations, for example, in genes that encode aldehyde dehydrogenase (ALDH) and alcohol dehydrogenase (ADH) which affect the metabolism of alcohol (Edenberg, 2007). The suggestion is that there may be combinations of common variants that may be causal, in a similar manner, for neuropsychiatric disorders. Genome-wide association studies (GWAS) were developed based on the CDCV model (Corvin et al., 2010).

The Common Disease, Rare Variant (CDRV) model, also known as the Multiple Rare Variants (MRV) model, on the other hand, proposes that one or more rare genetic variants within an individual can by itself cause disease, or collectively contribute to the risk of developing a common disease. According to this model, each rare variant may have a moderate to high penetrance but is found at a low frequency in the general population. The CDRV model suggests that these rare variants can act independently, or in concert with other genetic and environmental factors, to influence disease risk (Dickson et al., 2010; McClellan et al., 2007).

1.2.1 Study Approaches

The search for underlying genetic variants can be divided into two main approaches. On the one hand there are several research groups who have focussed on family pedigrees, using study designs like linkage analysis and, later, whole exome sequencing to locate genetic elements that co-segregate with the disease in these large family pedigrees (Cantor, 2019; Millar et al., 2000; Myles-Worsley et al., 2013; Ng, M. Y. M. et al., 2009). In line with the locus heterogeneity of SCZ, findings from these single families did not always replicate in other family pedigrees (Iwata et al., 2004). Highlighting the difficulty of isolating the causal element in SCZ in a single family, is that many of these families have continued to be studied over several decades using more advanced techniques, from early karyotyping through microarray and finally to whole exome sequencing. In contrast to this, larger population studies have attempted to identify common variants between unrelated individuals, either using association methods in genome-wide association studies (GWAS) (Corvin et al., 2010; Marshall et al., 2017; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Trubetskoy et al., 2022; Visscher et al., 2012; Wray et al., 2018) or via large-scale sequencing studies (Genovese et al., 2016; Jaffe, A. E. et al., 2018; Purcell et al., 2014; Singh et al., 2016; UK10K Consortium et al., 2015). These are covered in the following sections.

1.2.2 Family and Linkage Studies

Linkage analysis was one of the first genome-wide methods employed to identify causal genes co-segregating with disease in families with high disease burden (Borkowska et al., 2016; Millar et al., 2000). It relies on the principle of genetic markers being close enough to the disease-causing locus that their co-inheritance from parent

to offspring can be matched between affected members and absent from unaffected members of the family (Cantor, 2019). Linkage analysis played a key role in the early stages of identifying potential genetic regions associated with SCZ. Linkage studies primarily identified broad susceptibility loci spanning large chromosomal regions. However, they were particularly valuable in families with rare, highly penetrant variants like chromosomal rearrangements, leading to the discovery of important genes like DISC1. A translocation between chromosomes 1 and 11 - t(1;11) – in a male patient with SCZ, lead to the discovery of a Scottish family with multiple members affected by a range of mental and/or behavioural disorders (St Clair et al., 1990). Psychiatric diagnoses were found more frequently along with the translocation than without, but there were still affected family members who did not carry the translocation. The translocation breakpoint on chromosome 1 disrupted a gene that later became known as disrupted in schizophrenia 1 (DISC1). The Scottish family have been the focus of study in SCZ for over three decades since. In a more recent study, Ryan et al. (2018) studied both breakpoints in the translocation at base level resolution using whole genome sequencing and modern linkage techniques (for example, two-point linkage), discovering rare variants in genes, PDE4D on chromosome 5 and CNTN5 and GRM5 on chromosome 11 that may modify the risk in some family members. The latter two may have co-segregated with the translocation more in affected members.

Replication of findings from linkage studies were not always successful, but this does not imply the original findings were false. A large meta-analysis of linkage studies (GSMA) covering 3255 pedigrees and 7,413 genotyped SCZ or related cases suggested evidence for linkage on chromosomes 1, 2q, 3q, 4q, 5q, 8p, and 10q, providing valuable starting points for further investigation (Lewis et al., 2003; Ng, M. Y. M. et al., 2009). Replicability is a high standard of scientific research. However, in the context of SCZ, the mutational target is known to be highly heterogeneous, with numerous genes contributing to the disorder. As linkage studies typically focus on a single rare variant in a single family, replication of findings can be challenging due to the diverse genetic factors involved in SCZ. For example, if there are 100 genes throughout the genome conferring risk for SCZ and rare, damaging variants are required in only one, then variants private to a single family are unlikely to be replicated in any family with a variant in the other 99 genes (assuming they are in different chromosomal regions).

Therefore, one would expect replication of findings to be negative in a proportion of attempts (Ott, 2004).

The development of linkage markers from 2 to approximately 10 million markers enabled a much higher resolution in linkage (Cantor, 2019). Furthermore, with the advent of whole exome sequencing (WES) and whole genome sequencing (WGS), as described for the Scottish pedigree (Ryan et al., 2018), the ability to identify rare variants co-segregating with disease is one the most powerful methods for identifying genetic causes that affect single families (Ott et al., 2015; Teare & Santibanez Koref, 2014).

1.2.3 Genome-wide Association Studies (GWAS)

GWAS began with modest sizes and few results, but with increased cohort sizes they have emerged as a robust method for identifying regions of the genome that associate with SCZ. The development of large SNP arrays, or 'SNP-chips' (Affymetrix 300K/500K/5.0/6.0, Illumina 550K / Human 610Quad) allowed researchers to genotype more people at a greater genomic resolution. GWAS focus on unrelated case-control models rather than family-based models because co-inheritance artificially inflates the statistic at the SNPs. Some GWAS studies use unrelated trios to control for population stratification, since the trios act as their own control. Most SNPs chosen for the SNP-chip are chosen carefully so that a) they are roughly, equally spaced along the chromosomes and b) the alternative allele has a minor allele frequency (MAF) that is generally larger than 5% (Wang, M. H. et al., 2019). The genotypes of all cases and controls are read at each SNP and a logistic regression is calculated with case status as the dependent variable (case = 1; control = 0) and the count of minor alleles as the independent variable (homozygous for the minor allele = 2, heterozygous = 1, homozygous for the major allele = 0). As there are almost a million SNPs being tested, the chance of a false positive is very large. The p value is adjusted for multiple testing across the whole genome by dividing by 1 million, to give a genome-wide significance threshold of 5×10^{-8} (Corvin et al., 2010).

While it is possible to test millions of common SNPs, the existence of linkage disequilibrium (LD) blocks and common haplotypes means that the results for neighbouring SNPs would be nearly identical. Studies therefore select a smaller number of 'tag SNPs' that tag multiple haplotypes within the cohort allowing

neighbouring common SNPs to be 'imputed' from the tag SNPs (Corvin et al., 2010). The tag SNPs thus act as proxies for a larger locus. If two significant loci overlap or fall outside of 50 kilobases from each other, then they combined into a single risk locus. An association with a SNP in GWAS is therefore an association with loci that can be as large as 2 Mb in size, although influences over larger distances (up to 10 Mb) have been demonstrated via simulated GWAS (Dickson et al., 2010).

GWAS for SCZ were initially created with small cohort sizes and modest results but have aggregated and grown into very large cohorts with many genome-wide significant results. The first study of SCZ with 158 cases and 178 controls (Lencz et al., 2007). A single locus near the gene CSF2RA was identified but did not reach genome-wide significance. Further GWAS produced similarly negligible results, until cohort sizes increased. It has been estimated that approximately one extra risk locus is identified with a genome-wide significant association for SCZ for every additional 1000 cases. The Schizophrenia Working Group of the Psychiatric Genomics Consortium combined previous genotype data, pooling previous studies from the International Schizophrenia Consortium (ISC), SCZ Genetics Consortium (SGENE) and Clinical Trials of Intervention Effectiveness (CATIE) into one large study (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). In the most recent study (PGC3) (Trubetskoy et al., 2022), the PGC used a combined set of 76,755 individuals with SCZ and 243,649 controls. Of 624 significant loci ($p < 1 \times 10^{-6}$), 287 risk loci were identified with a genome-wide significant ($p < 5 \times 10^{-8}$) association with SCZ, 179 more than the 108 genome-wide significant loci identified in PGC2 (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). 120 putative genes were identified using a statistical technique called 'fine-mapping' to reveal the most likely causal common variants within each risk locus associated with SCZ. Fine-mapping evaluates all SNPs simultaneously in the locus that are in LD with the tag SNP, assessing the probability that any SNP is causal (Benner et al., 2016). This helps to narrow down large risk loci to a smaller locus. The 120 genes were also enriched for rare disruptive coding variants in people with SCZ, significantly in glutamate receptor subunit (GRIN2A), the transcription factor SP4, cohesin subunit SA-1 (STAG1), and constitutive coactivator of PPAR-gamma-like protein 1 (FAM120A). Other genes prioritised included ACE and E3

ubiquitin ligases (Trubetskoy et al., 2022). The most significant tag SNP in PGC3 was on chromosome 6 with a p-value of 1.205×10^{-39} and an OR of 1.228. This tag SNP was in LD ($r^2 > 0.1$) with a block 5.5 Mb in size. The associated genes from PGC3 showed an enrichment in mainly postsynaptic genes and in a gene set derived from a single nucleus RNA-Seq (snRNA-Seq) study in pyramidal CA1 and CA3 neurons of the hippocampus, excitatory glutamatergic neurons, and inhibitory GABAergic interneurons in the cerebral cortices (Trubetskoy et al., 2022).

The Psychiatric Genomics Consortium's (PGC) Copy Number Group also conducted a copy number variant GWAS to explore the association between CNVs and schizophrenia (SCZ) risk (Marshall et al., 2017). Analysing 21,094 SCZ cases and 20,227 controls, the study identified 8 loci with a genome-wide significant association with SCZ, including 1q21.1, 22q11.2, proximal and distal 16p11.2, 3q29, 7q11.2, 15q33.3. Rare deletions in activity-regulated cytoskeleton-associated protein (ARC) complex genes, as well as losses in adhesion molecule (NRXN1) and postsynaptic scaffolding genes (DLG1 and DLG2). This GWAS demonstrated that variants with stronger effects can be detected through the investigation of CNVs, which are less common but have more significant impacts on SCZ risk compared to SNPs (Marshall et al., 2017).

From the very early GWAS studies it was known that rare variants could never be used as markers in GWAS studies. The cohorts of each genotype would be too small to have enough power to find any statistical significance (Corvin et al., 2010). A critique of GWAS has been that instead of the disease association being with the common tag SNP, that there are rare variants nearby that form synthetic associations with the tag SNP and thus with SCZ (Dickson et al., 2010; Mitchell & Porteous, 2009). This is one reason behind the development of the fine-mapping statistical analysis being used to supplement GWAS (Broekema et al., 2020), to further localise where the associated loci may be harbouring causal variants. Dickson et al. (2010) show through simulations of a GWAS for sickle cell anaemia with a known causal rare variant, that synthetic associations can be made between the disease and common variants up to 10 Mb from the rare causal variant. Dickson et al. (2010) also highlight that if rare causal variants were distributed among cases equally among genotypes, then no association would be identified, and furthermore that only a small proportion of cases need to harbour rare variants for the locus to be significant. In the latest GWAS (PGC3), as noted above,

many genes identified through fine-mapping were also enriched in rare variants (Trubetskoy et al., 2022), but this still potentially leaves many rare variants in loci that do not reach genome-wide significance unidentified (Broekema et al., 2020).

1.2.4 Exome/Genome Sequencing Studies

With the development of high-throughput sequencing technologies, it became possible to rapidly sequence a complete genome to single nucleotide level. Since sequencing the whole genome (WGS) was prohibitively expensive and because disease-causing variants were proposed to be those that code for proteins, projects initially focussed on whole exome sequencing (WES). The exome consists of all coding regions – exons – (and several bases of flanking DNA from non-coding introns) that make up protein-coding genes accounting for 1-2% of the human genome. Sequencing projects have been used in several approaches, for top-down variant prioritisation (Genovese et al., 2016; Purcell et al., 2014; UK10K Consortium et al., 2015), candidate gene sequencing (Frank et al., 2011; Lin et al., 2023), RNA-Sequencing (Sartorius et al., 2006), case-control studies (Singh et al., 2016), and weighted burden analysis (Curtis et al., 2018). As in GWAS over the last decade studies have been aggregated with increasing cohort sizes (and increasing lists of authors).

1.2.4.1 Targeted Sequencing Studies

Hypothesis-driven targeted sequencing of candidate genes focussed on groups of genes from families linked to synaptic transmission were informative but yielded low returns. Frank et al. (2011) sequenced the exons from 10 glutamatergic signalling related genes from 1,034 cases and controls. The genes included 3 subunits of the ionotropic NMDA receptor (GRIN1, GRIN2A, GRIN2B), 2 subunits of the AMPA (a-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) receptor (GRIA1, GRIA2), one subunit of the metabotropic (GRM1) glutamate receptor and associated scaffolding partners (DLG1/SAP97, DLG2/PSD93, DLG3/SAP102, and DLG4/PSD95). GRM1 had a small cluster of 5 rare variants in the cysteine-rich domain exclusive to patients.

1.2.4.2 Early Sequencing Findings

Like GWAS, WES studies started out with small cohorts and grew as the cost of sequencing decreased. Myles-Worsley et al. (2013) used the approach of sequencing the whole genome to a depth of 30X for a single proband to locate the breakpoints of a hemi-deletion in SLC1A1 after which they could develop primers to verify whether

other carriers within the family had the same breakpoints. Also in family studies, a team in Australia sequenced the whole genomes of 91 high-density multiplex families and applied CNV algorithms to analyse for deletions and duplications (Khan et al., 2018). This revealed many rare CNVs in 300 individuals, a third of whom were unaffected by SCZ or any other neuropsychiatric diagnosis, especially in intronic and intergenic regions, and duplications in the exonic regions. This large-scale study revealed a burden of rare duplications and deletions in exonic and intronic regions in SCZ (Khan et al., 2018). These CNVs demonstrated high locus heterogeneity and overlapped some known SCZ genes (including either duplication or deletion of ABCA13, PKD1L1, AUTS, CALN1, CNTNAP2, CNTNAP4, CTTN, PPFIA2, SHANK2, CSMD1, DRD5, SLC2A9, NRG1, SATB2, ST6GAL2, TNR, and TOP3B).

1.2.4.3 Large Whole Exome Sequencing Studies

At the same time, large sequencing projects were created aim to sequence the exomes of 1000s of cases with the aim of identifying rare and low-frequency variants with large effect sizes in SCZ and other diseases. Two such projects were the UK10K Project which was started in 2010 (UK10K Consortium et al., 2015) and the Swedish Schizophrenia Exomes project (Genovese et al., 2016; Purcell et al., 2014). Within the 'Exomes' arm of the UK10K, 6000 people in three disease groups, shown to have a high heritability and driven by rare variants with a large impact (including schizophrenia and autism (ASD), rare disease, and severe obesity), had their exomes sequenced. In the approximately 3000 SCZ and ASD samples that passed quality control, there was a high burden of rare variants, and most were very rare (UK10K Consortium et al., 2015). Whilst there was some success at identifying disease-causing rare variants in the rare diseases group, no replicating disease-causing rare variants were identified in the severe obesity or SCZ/ASD groups. This suggests that the mutational target of SCZ and ASD is very large and supports the large amount of allelic and locus heterogeneity.

1.2.4.4 The Swedish Schizophrenia Exomes Project

The Swedish Schizophrenia Exomes project was completed in two phases. In the first phase, the exomes of 2,536 SCZ cases and 2,543 matched controls from a large group of 11,000 cases and controls were sequenced (Purcell et al., 2014). The authors identified a large burden of rare variants among both cases and controls. They implemented a top-down approach and started by filtering the variants using a set of

2,546 genes derived from recent GWAS risk loci – including CNV and de novo SNV results – as these had a higher likelihood of being associated with SCZ. These genes harboured a higher burden of ultra rare (minor allele frequency (MAF) < 0.1%), rare (MAF < 0.5%) disruptive coding variants, and singletons in cases compared to controls. The authors used 5 *in silico* prediction tools to estimate pathogenicity and categorised the rare missense variants as non-synonymous-strict if all 5 tools predicted pathogenicity. This class of variants as a group was also found to be enriched in the cases compared to non-synonymous variants predicted to be pathogenic by only one tool. A further grouping of variants whose coding consequence would be disruptive, for example, those causing a frameshift, gain of a stop codon, or disrupting a splice site, was analysed between case and controls in genes identified in other studies (Fromer et al., 2014; Xu et al., 2012) as carrying *de novo* mutations. Disruptive variants were enriched in genes involved in synaptic maintenance and organisation (PSD-related genes and the post-synaptic scaffolding-related, activity-regulated cytoskeleton-associated (ARC) complex) (Purcell et al., 2014). In a study published at the same time, Fromer et al. (2014) had sequenced the exomes of 623 SCZ trios to identify *de novo* mutations in SCZ probands. As well as a burden of inherited rare risk alleles in an enrichment analysis using a SCZ-relevant gene set, the authors also identified two rare *de novo* mutations in the TFIID transcription initiation complex subunit, TAF13, which reached genome-wide significance. Further enrichment of *de novo* non-synonymous mutations was found in genes relating to the ARC complex and PSD-related genes, similarly to Purcell et al. (2014). When the UK10K and Swedish Exomes phase I studies were combined in a case-control analysis (Singh et al., 2016), ten of the total 4,264 individuals were identified with loss-of-function (LoF) variants in the variant-intolerant SETD1A, a histone methyltransferase, but not in 45,376 controls. This is considered one of the first robust findings of a common variant found between unrelated variants. It also leaves the causal variants in the rest of the cases yet to be determined and would require further study to identify them.

By 2016, the Swedish Exomes project had expanded to the whole exomes of 12,332 individuals including 4,877 individuals with SCZ (Genovese et al., 2016). The group used more refined SCZ-relevant gene sets for enrichment analysis and leveraging variant frequencies in 45,376 healthy individuals in a large cohort from the Exome

Aggregation Consortium (ExAC) (Lek et al., 2016) to assess allelic frequencies in a wider population. 244,246 coding sequence and splice-site ultra-rare variants (URVs) were identified in a total of 11,080 SCZ cases and controls. Each of these URVs was only seen in one individual in the whole cohort and was absent from ExAC. Importantly, the increased size of the cohort compared to Purcell et al. (2014) meant that two thirds of the variants called as singletons in the earlier study were found to be not as rare as first thought. After assessing ultra-rare variants as disruptive and damaging (dURVs), by combining the results from several *in silico* prediction tools, Genovese et al. found that there was an average background of 4.0 dURVs per control and a small but significant ($P = 1.4 \times 10^{-10}$) increase of 0.25 dURVs per individual among cases. Rare variants are best studied in aggregation using a common property, for example, within a gene, set of genes, with specific physiochemical properties. The authors used a variety of enrichment analyses in gene sets derived from various sources. One source was a set of 2,647 genes specifically expressed in brain tissue analysed using RNA-Sequencing (RNA-Seq) from 3 human samples taken from the cerebral cortex (Fagerberg et al., 2014). Genovese et al. (2016) found that dURVs were enriched in these brain-specific genes but not in genes with expression specific to other cells and there was a larger excess of dURVs in cases enriched in neuron-specific genes. In this study also, no individual gene with a large effect was implicated supporting the idea that SCZ has a large mutational target and is a highly polygenic disease.

The same Swedish Exomes data has been the subject of several further studies (Curtis et al., 2018; Nguyen et al., 2017; Singh et al., 2022). Together with the UK10K cohort, there remains a large percentage of patients for whom the causal gene has not been identified. Using newer methods, larger population databases to further re-classify variants as ultra-rare or common, and more refined gene sets tailored to the pathways involved in SCZ as knowledge of the underlying pathophysiology advances. In 2017, the same group further analysed the Swedish Exome set along with a trio set using an updated Bayesian method for analysing transmission patterns across multiple trio groups to analyse potential gene interactions between different neurodevelopmental disorders (NDD) and SCZ (Nguyen et al., 2017). The authors leveraged ExAC to flag variants as in-ExAC and not-in-ExAC, used 7 *in silico* pathogenicity prediction tools (including SIFT and PolyPhen-2 (HumDiv and HumVar) reviewed here in Chapter 4).

The authors do make an estimate that there are 1,551 risk genes for SCZ with weaker effects compared to NDDs. Using Bayesian analysis with a false discovery rate of 0.3, the authors identified 24 candidate SCZ risk genes, the most significant being SETD1A (also including TAF13, RB1CC1, PRRC2A, VPS13C, MKI67, RARG, ITSN1, KIAA1109, DARC, URB2, HSPA8, KLHL17, ST3GAL6, SHANK1, EPHA5, LPHN2, NIPBL, KDM5B, TNRC18, ARFGEF1, MIF, HIST1H1E, and BLNK) and more were just below the threshold for inclusion (Nguyen et al., 2017).

1.2.4.5 The Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) Consortium

The latest and largest exome sequencing study was a companion study to the PGC3 GWAS study, the Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) Consortium study (Singh et al., 2022) comprised from 24,248 SCZ cases, 97,322 controls (including data from Swedish Exomes and UK10K), and 3,402 trios. The group identified 10 genome-wide significant genes with a burden of dURVs. As well as SETD1A, the genes included subunits from two types of glutamatergic receptor, GRIN2A and GRIA3; a calcium channel subunit, CACNA1G; a Rho guanine nucleotide exchange factor involved in neuronal growth and actin cytoskeleton remodelling, TRIO; genes involved in gene expression (SETD1A, SP4, RB1CC1 and XPO7), and protein ubiquitination and clearance (CUL1, HERC1). The SCHEMA study leveraged approximately one third of the latest population data from 141,456 individuals in the Genome Aggregation Database (gnomAD) v2 (Karczewski et al., 2020) thus enabling the re-classification of more variants identified as ultra-rare in previous studies. The authors also analysed the URVs against 1,732 genes from biological pathways datasets (Gene Ontology (GO), KEGG, and Reactome) showing enrichment in the postsynaptic density (PSD), chromatin modification, regulation of ion transport, axon guidance, voltage-gated cation channel activity, and synaptic transmission. Enriched URVs in genes expressed in the brain were explored using gene sets derived from expression values (Finucane et al., 2018) from V6p of the Gene-Tissue Expression Project (GTEx – reviewed in Chapter 2), finding the highest enrichment in the frontal cortex/cerebral cortex and anterior cingulate cortex.

Despite the large amount of research carried out on the Swedish Exomes dataset in the last decade, the results have been modest. The most recent study (SCHEMA) only found 10 genome-wide significant genes (Singh et al., 2022), although there may be

many ultra-rare variants of uncertain significance remaining to be characterised completely. The studies all have in common the following:

- Assessment of allelic frequency – studies leveraged new population databases as they became available – 1000 Genomes, ExAC, gnomAD v2. The larger the database, the fewer the variants that are misclassified as ultra rare (Whiffin, N. et al., 2019).
- Selection of appropriate gene sets for enrichment analyses. Genovese et al. (2016) state that enrichment scores were potentially weakened because the gene sets included too many non-SCZ-related genes. On the other hand, gene sets that cover a narrow biological domain will show increased enrichment but will not explain all genes with variants.
- Genes with high expression in the brain for enrichment analysis. All studies touched on a set of genes derived from healthy brain donors. Early studies, used by Purcell et al. (2014) and Genovese et al. (2016) used broad sets of genes derived from small samples sizes (N=6 (microarray) and N=3 (RNA-Seq) respectively). The SCHEMA study (Singh et al., 2022) utilised expression values from an older V6p data freeze, analysed in a previous study (Finucane et al., 2018) which was originally analysed against the older human reference genome (GRCh37) and an older gene annotation model (GENCODE v19).
- All studies assessed pathogenicity using a combination of 5 or more *in silico* predictive tools (see Chapter 4).
- All variants are analysed as stand-alone variants.

1.2.4.6 Single-cell and single-nucleus RNA-seq

Over the past decade, single-cell sequencing has revolutionised biological research by enabling the detailed analysis of individual cells, uncovering cellular heterogeneity and complex tissue dynamics (Wen & Tang, 2022). The number of cells that can be analysed per experiment has increased rapidly, expanding from a few cells to hundreds of thousands of cells. This technological leap has been particularly impactful for SCZ research. By analysing gene expression at the single-cell (scRNA-seq) or single-nucleus (snRNA-seq) level, researchers have identified specific and distinct cell subtypes and molecular pathways implicated in SCZ (Batiuk et al., 2022; Huuki-Myers et al., 2024; Jaffe, Andrew E. et al., 2020; Ruzicka et al., 2024; Skene et al., 2018). revealing novel

insights into its pathogenesis. For example, Batiuk et al. (2022) compared snRNA-seq data from microdissected cortical columns between postmortem SCZ cases and healthy controls and identified reduced densities of superficial layer (L2/3) interneurons and increased numbers of excitatory neurons in L2/3 and L4/5. Ruzicka et al. (2024) recently published a single-cell transcriptome including almost half a million nuclei across the frontal cortex from 140 individuals (cases and controls). Their findings suggest that future single-cell based studies could help deconvolute different subtypes of SCZ based on transcriptional landscape differences between patients – for example, those affected by variants in genes involved in synaptic transmission and those affected by variants in genes involved in metabolic reactions. Huuki-Myers et al. (2024) developed a transcriptomic map using over 54,000 nuclei from 10 donors across the DLPFC extending a previous study of 3 donors (Maynard et al., 2021). One of their key findings was the localisation of SCZ-associated ephrin-EPH signalling to deep L5/6 excitatory neuron subtypes. Finally, Jaffe et al. (2020) used single-cell RNA-seq to examine the granule cell layer of the dentate gyrus, which is a part of the hippocampal formation. They found expression signals of GRM3 and CACNA1C and over 1.35 million eQTL that had not been found in bulk RNA-seq of the hippocampus, highlighting the advantages of studying single-cell data over bulk RNA. With a vast amount of data becoming available, researchers are integrating findings from many modalities, such as eQTLs and GWAS , and rare variant studies, and combine them with high-resolution single-cell maps. This integrative approach holds great promise for narrowing in on the complex biology of schizophrenia Some exciting areas will be in mapping robust, known risk genes onto individual cell subtypes across the whole brain, and the investigation of how different gene transcripts are used between different cells subtypes.

Table 1.1 summarises the key findings of studies carried out on patients with SCZ along with the number of patients and controls.:

Table 1.1 Key WES, GWAS studies on patients with SCZ

Study	Cases	Controls	Key Genes Identified
Hypothesis-driven targeted sequencing (Frank et al. 2011)	503 SCZ 268 BPD	538	GRM1 – cluster of 5 variants exclusive to patients
GWAS PGC2 (PGC, 2014)	36,989	113,075	108 genome-wide significant loci Notable gene associations: DRD2, GRM3, GRIN2A, SRR, GRIA1, CACNA1C, CACNB2, CACNA1I
WES Swedish Exomes (Purcell et al., 2014)	2,536	2,543	Burden of variants was highly polygenic Genes enriched for variants: FMR1, CCHCR1, KYNU, UFL1, SYNGAP1, DLG1, DLG4 (PSD-95), ARC, CACNA1C
WES De novo in trios Fromer et al., 2014	623 trios		TAF13, ARC complex, PSD-related genes
WES Case-Control (Singh et al. 2016)	4,264	45,376	SETD1A LoF variants in 10 cases
WES Swedish Exomes (Genovese et al. 2016)	4,877	6,203	High degree of polygenicity of dURVs No individual gene passed exome-wide significance – suggesting no individual gene has rare variants that explain >1% of SCZ cases. Previously identified genes with LoF variants: TAF13, SETD1A, NRXN1.

Study	Cases	Controls	Key Genes Identified
CNV GWAS (Marshall et al., 2017)	21,094	20,227	8 genome-wide significant loci, including: 1q21.1, 22q11.2, 16q11.2, 3q29, 7q11.2, 15q33.3 Rare deletions in ARC, NRXN1, DLG1, DLG2.
WES Swedish Exomes (Nguyen et al., 2017)	6,699	13,028 + 1,077 trios	Estimate 1,551 risk genes. (FDR<0.1) SETD1A, TAF13, PRRC2A, RB1CC1 (FDR<0.3) SETD1A, TAF13, RB1CC1, PRRC2A, VPS13C, MKI67, RARG, ITSN1, KIAA1109, DARC, URB2, HSPA8, KLHL17, ST3GAL6, SHANK1, EPHA5, LPHN2, NIPBL, KDM5B, TNRC18, ARFGEF1, MIF, HIST1H1E, BLNK
GWAS CLOZUK (Pardiñas et al., 2018)	11,260	24,542	140 genome-wide significant risk loci: Candidate causal genes in 33 loci Signal enriched among genes under strong selective pressure Incl. CACNA1D, SLC39A8,
GWAS PGC3 (Trubetskoy et al., 2022)	76,755	243,649	634 significant loci ($p < 1 \times 10^{-6}$) 287 genome-wide significant loci ($p < 5 \times 10^{-8}$) 120 putative genes – including: GRIN2A, SP4, STAG1, FAM120A, BCL11B, CACNA1C, RERE, FOXP1, MYT1L, SLC39A8)
WES SCHEMA (Singh et al., 2022)	24,248	97,322 + 3,402 trios	10 Exome-wide significant genes: SETD1A, GRIN2A, GRIA3, CACNA1G, TRIO, SP4, RB1CC1, XPO7, CUL1, HERC1

1.3 Gap in the Literature

While a lot has been discovered in the last decade, the full range of genes and biological pathways that contribute to the risk of developing SCZ is yet to be completely characterised. One of the challenges in identifying genes involved in SCZ stems from the large locus heterogeneity involved in the pathophysiology of SCZ. Existing studies had demonstrated that enrichment scores can be weakened by the inclusion of too many non-SCZ-related genes or narrowed to a specific domain, potentially overlooking key genes with variants. As a result, the gene sets used for enrichment analyses could be further developed and refined.

Since the start of the GTEx project, data had been released every 6 months for researchers to study. With access to the raw RNA-Sequencing reads in the final release V8, there was the opportunity to study all selected brain regions from the most donors. Furthermore, by aligning the raw RNA-Sequencing reads to the latest human reference genome (GRCh38) and gene annotation models, it would be possible to maximise the information that could be extracted from each sample for a more accurate gene expression profile than provided for by summary statistics that were freely available from the GTEx Portal. Additionally, given the number of brain regions sampled in GTEx, the gene set containing highly expressed gene expression from SCZ-relevant cortical and hippocampal regions could be refined by excluding any genes with greater expression in less SCZ-relevant brain regions. The resulting gene set might therefore be more specific to the processes that may contribute to the SCZ phenotype.

Subsequently, by studying rare damaging variants in these genes in patients with SCZ, a more refined collection of pathways involved in SCZ can be discovered.

The hypothesis of this thesis is that one or more ultra rare variants with a large effect is likely to be causal or will contribute to the development of SCZ. Variants may be inherited, as shown in family studies, or *de novo* (Fromer et al., 2014). *De novo* mutations that occur in the zygote, embryo or foetus will be ultra-rare, private to the individual, and rarely replicate in another patient except their offspring if the variant is carried in germline cells. *De novo* mutations that affected a parent, grandparent, or other ancestor, may be inherited by siblings and cousins in a Mendelian-like manner, and differences in phenotype are due to the specific pathways affected by the rare

variant or combined rare variants (Ryan et al., 2018). These ultra-rare variants are likely to be localised in key genes in key pathways, with a higher expression in the brain, and narrowing this down further, in tissues that are affected by SCZ symptoms.

1.4 Overall Aim and Main objectives

Overall aim

- Identify novel rare genetic variants potentially contributing to SCZ by leveraging a gene set enriched for genes with higher expression in brain regions relevant to the disorder.

Objectives

1. Identify Brain Regions: Utilize existing scientific literature to identify two brain regions known to be most implicated in SCZ and two control brain regions less affected by core SCZ symptoms.
2. Identify a Targeted Gene Set: Analyse RNA-Seq data from healthy control subjects to identify a set of genes with significantly higher expression in the two SCZ-implicated brain regions compared to the two control regions.
3. Identify Rare Variants in SCZ Cases: Analyse whole-exome sequencing data from SCZ patients, focusing on the self-curated gene set to prioritise rare damaging variants exclusive to these patients.
4. Functional Pathway Analysis: Investigate biological pathways enriched by the identified rare variants and analyse their potential role in SCZ pathophysiology.

2 Ranking of Brain Regions in Schizophrenia

2.1 Background

2.1.1 Schizophrenia

Schizophrenia (SCZ) is a complex and heterogeneous neurological disorder that affects the way a person thinks, feels, and behaves. Symptoms of SCZ are categorised into three main categories (Glausier & Lewis, 2018): positive symptoms such as hallucinations and delusions; negative symptoms, such as social withdrawal and apathy; and cognitive symptoms such as deficits in working memory. Whilst the underlying causes of these symptoms are not still fully understood, research in the last 100 years has shown that there are abnormalities in brain structure, connectivity and function that have been associated with symptoms presenting in SCZ (Birnbaum & Weinberger, 2017; Glausier & Lewis, 2018).

2.1.2 Symptoms

The underlying neurobiology of SCZ symptoms is complex and involves several integrated brain regions or neural circuitry (Glausier & Lewis, 2018). Few symptoms in SCZ involve only a single brain region.

- Hallucinations may involve the primary auditory cortex, superior temporal gyrus for processing auditory and visual information, prefrontal and anterior cingulate cortices, the insula, and thalamus.
- Working memory impairments in schizophrenia are thought to involve the abnormal activity and connectivity of the prefrontal cortex, temporal lobes, thalamus, and the hippocampus.
- Executive functions are cognitive functions that include planning, decision-making, working memory, and cognitive flexibility. They are thought to be mediated by the prefrontal cortex,
- Social cognition is a complex cognitive function that involves the integration of several brain regions including the prefrontal and anterior cingulate cortices for regulation of emotions and processing social cues, the amygdala for processing of emotions and the significance of social cues, and the insula for empathy, social pain, and social pleasure.

- Positive and negative movement symptoms, for example, involuntary movements (tardive dyskinesia) and slowing movement (bradykinesia) may involve the motor and premotor cortices, the basal ganglia (including the putamen, caudate nucleus, nucleus accumbens, globus pallidum, substantia nigra and subthalamic nuclei) and the cerebellum.

For the purposes of the current study, it is important to identify which brain regions are likely to be affected in the core symptoms of SCZ to be able to examine gene expression within those regions. This thesis hypothesises that genes highly expressed in those regions likely to cause symptoms in SCZ would not be highly expressed in other regions that that are less often associated with SCZ. Therefore, it is also important to identify regions of the brain that are less often associated with core symptoms in SCZ to exclude from the analysis genes that are highly expressed in those regions.

2.1.3 Brain Study Modalities in Schizophrenia

The brains of patients with SCZ have been studied using a wide range of non-invasive imaging techniques that are able to detect specific aspects of brain biology, for example hydrogen nuclei in water, blood flow, glucose metabolism, or movement of a radioactive tracer. Table 2.1 shows a selection of these study modalities. The modalities can measure anatomical features (MRI, DTI), function (fMRI, PET, SPECT) or biochemical changes (MRS). Magnetic resonance imaging (MRI) uses strong magnetic fields and radio waves to detect small magnetic changes in the nuclei of atoms in the tissues. These properties differ between tissues and can produce high resolution images that can be subdivided into cortical or subcortical regions, and further subdivided into separate subregions or nuclei. This allows for the comparison of cortical thickness or subcortical volumes to be compared between healthy individuals and patients with SCZ. Functional MRI (fMRI) measures changes in blood oxygenation levels (BOLD) and can be used to compare changes in blood utilization between two or more active states (for example, between a resting state and/or while performing specific tasks). Differences in regional activation can be determined between healthy individuals and patients with SCZ. Positron emission tomography (PET) and single positron emissions computed tomography (SPECT) also trace blood flow using a radioactive tracer, showing which brain regions are most active between controls and patients with SCZ. Histological and immunohistochemical studies on brain tissues are

limited to post-mortem (PM) studies. PM studies can identify differences in cell orientation, density, and morphology, as well as altered distribution of receptors and other proteins within cells.

Table 2.1 Brain study modalities

Magnetic Resonance Imaging (MRI)	Structural - high-resolution using the interaction of radio waves and hydrogen nuclei in water and fats.	Cortical thickness Cortical surface area Gyrification Sub-cortical volume
Functional MRI (fMRI)	Functional – low spatial resolution - using changes in blood oxygenation levels between two states	Resting State – Functional connectivity. Working Memory tasks.
Positron Emission Tomography (PET)	Glucose Metabolic Rate – using a radioactive tracer	Glucose uptake / tracer accumulation
Magnetic Resonance Spectroscopy (MRS)	Detects radio frequency of atoms in molecules	Biochemical changes in localised brain regions
Single-photon emission computed tomography (SPECT)	Functional – blood flow via a radioactive tracer	
Diffusion Tensor Imaging (DTI) Diffusion-weighted MRI (DW-MRI)	Structure and connectivity – using the diffusion of water molecules through tissues	White matter pathways
Post-mortem studies	Structural changes in tissues, cell size, shape, orientation and distribution. Distribution and density of specific proteins	Histology Immunohistochemistry Immunohistochemistry

2.1.3.1 Meta- and Mega-analyses of Magnetic Resonance Imaging (MRI) Studies in Schizophrenia

Results from imaging studies in the first few decades of SCZ research were often contradictory due to small sample sizes and differences in equipment and analysis protocols. The last decade has seen efforts towards pooling of data in large-scale, collaborations to increase sample sizes and standardisation of methodologies, using similar equipment, as well as the collection of covariate data (medication use, duration of illness, sex, age, etc) (Thompson et al., 2017). ENIGMA (Thompson et al., 2017) and COCORO (Okada et al., 2016) are two such collaborative consortia. They were set up to combine MRI cohort data from member research centres and analyse it using standard protocols, for example, aligning MRI images to the same reference brain atlas. The Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) consortium was conceived with many smaller working groups each studying differences in brain morphology including structural features (lateralisation, cortical thickness, subcortical volume), genomic effects on structure (Hibar et al., 2015, 2017), disease (including SCZ, Parkinson's disease, ataxias, post-traumatic stress disorder PTSD), as well as variability of healthy brains (plasticity, lateralisation, lifespan working groups) (Thompson et al., 2020). Cognitive Genetics Collaborative Research Organisation (COCORO) was a consortium of Japanese universities and hospitals that was set up to replicate ENIGMA findings using the same protocols and extend them using stronger field MRI equipment, with improved automation and image segmentation (Hashimoto et al., 2017). ENIGMA and COCORO have published several prospective meta- and mega-analyses. These differ from standard retrospective meta-analysis where results (for example, the effect sizes) reported in published studies are analysed, irrespective of their methods, to draw broad conclusions. In prospective studies, data is collected in a collaborative project where researchers from different sites collect MRI data following a standardised protocol. This allows for a more controlled analysis and the ability to investigate specific research questions that would not be possible with individual studies.

2.1.4 Genotype-Tissue Expression (GTEx) Project

While many regions across the brain have been explored with respect to SCZ, there are limited existing resources of gene expression data for the whole brain. The GTEx

project was launched by the National Institutes of Health (NIH) in 2010 to study the effects of genotype on gene expression in 54 tissues, including 11 brain regions, across the body in up to 1000 individuals (GTEx Consortium, 2020). RNA-sequencing data was sampled and analysed for each tissue sample. Whole exome sequencing data was collected from whole blood samples (GTEx Consortium, 2020). The data has been used to study many aspects of the variability of gene expression, for example, the effects on expression due to genetic variation (expression quantitation trait loci (eQTL) and splicing QTL (sQTL) (Saha et al., 2017) and on allele-specific expression (ASE) (Castel et al., 2020). Recently, GTEx brain RNA-Seq samples have been analysed for eQTLs that moderate gene expression differences between the cortex and cerebellum (Sieberts et al., 2020). The brain regions in GTEx were originally selected for study because of their involvement in neurodegenerative diseases – Alzheimer's disease, Parkinson's disease, and Huntington's disease. Figure 2.1 shows the regions of the brain included in the GTEx project for which there is available sequenced gene expression data.

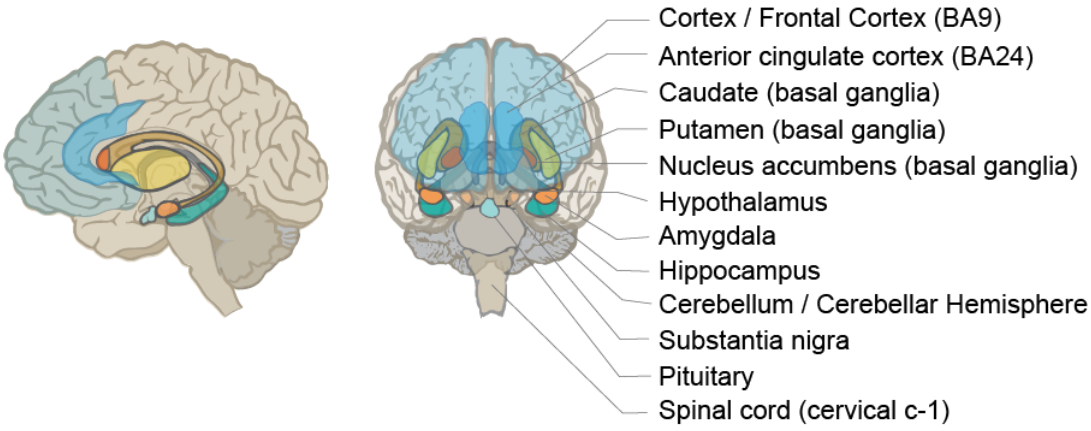

Illustration by Mariya Khan, Pattern Team © 2019 GTEx Portal, Broad Institute

Figure 2.1 Brain regions sampled in the GTEx Project
Permission granted (see image for attribution)

2.1.5 Aim and Objectives

A critical step in this thesis is the selection of two brain regions with the strongest evidence of involvement in SCZ and two control regions with minimal evidence of SCZ involvement. This selection process is crucial because the chosen regions will serve as the foundation for all subsequent analyses. Differences in gene expression between

these gene regions hold the potential to reveal key insights into the molecular mechanisms underlying SCZ.

Selecting the most appropriate regions requires careful consideration. The choice of these four regions will ultimately influence the set of genes identified as potentially relevant to the disease, and consequently, the set of genetic variants investigated in the remaining chapters. While there is some evidence implicating various brain tissues in SCZ, there is no existing definitive ranking system for their level of involvement. Therefore, this chapter focusses on a review of existing literature to identify the two most-affected regions with the most compelling evidence and two control regions with the least association with the disease.

The GTEx project provides the transcriptomes for 11 brain regions, offering a valuable resource for studying gene expression. However, to ensure the chosen regions are relevant to SCZ, an investigation into their anatomical locations within the brain is first conducted. This investigation ensures that the subsequent selection of SCZ-affected regions and control regions aligns with established knowledge about the disease.

Overall aim:

- Identify from existing literature two brain regions with the most evidence of being involved in SCZ and two brain regions for which there is the least supporting evidence for being involved in SCZ.

The objectives are:

1. Identify anatomically locations of the GTEx brain regions.
2. Review these brain regions for their involvement in the core symptoms of SCZ using evidence from a selection of large collaborative MRI imaging studies, post-mortem studies, and other supporting imaging modalities.
3. Select two brain regions with the most evidence for being involved in SCZ.
4. Select two brain regions that are least likely to be involved in core SCZ symptoms.

2.2 Methods

2.2.1 Selection of GTEx Brain Regions

The GTEx project provided a valuable resource for this study. The project collected high-quality RNA-Seq data from post-mortem healthy donors across multiple tissues, including 11 distinct brain regions (Table 2.2). Up to 1000 donors participated, with tissues samples collected from various body sites following a standardised protocol (Carithers & Moore, 2015; Carithers et al., 2015).

For brain tissue collection, the GTEx project adhered to strict ethical protocols. Brains were obtained from deceased individuals who had provided prior consent for brain donation. These brains were removed, frozen and sent to the University of Miami Brain Endowment Bank (MBEB) for standardised processing (Carithers & Moore, 2015; Carithers et al., 2015). The 'Cortex' and 'Cerebellum' regions (Table 2.2) were sampled from the right hemisphere of the brain and preserved in PAXgene at the Biospecimen Source Site (BSS). Upon arrival at MBEB, the brains were sliced into coronal sections and frozen. 'Frontal Cortex (BA9)' and 'Cerebellar Hemisphere' samples were collected from the right hemisphere near the BSS-sampled sites. The remaining 9 brain regions (Table 2.2 and Figure 2.2) were sampled from the left hemisphere. The GTEx project focusses solely on one hemisphere and did not investigate laterality (differences of gene expression between the left and right hemispheres) (GTEx Symposium, 20 May 2015). Finally, tissue samples were sent to the Broad Institute for molecular analysis, with RNA-sequencing performed as described in the GTEx pilot study (GTEx Consortium, 2015).

Table 2.2 GTEx brain regions

11 brain tissue sites – cortex and cerebellum sites were replicated and preserved either frozen or in PAXgene.

GTEx Tissue Site	Hemisphere	Preservation	Known Function
Cortex	Right	PAXgene	Working Memory, Executive Function
Cerebellum	Right	PAXgene	Learned motor sequences, refined motor control
Frontal Cortex (BA9)	Right	Frozen	Working Memory, Executive Function
Anterior cingulate cortex (BA24)	Left	Frozen	Salience, Emotion
Hippocampus	Left	Frozen	Working memory, Memory consolidation, spatial and temporal
Amygdala	Left	Frozen	Emotion, processing of fear
Hypothalamus	Left	Frozen	Brain-body connection, stress, homeostasis
Caudate (anterior, L)	Left	Frozen	Sensory, motor, cognition processing
Putamen (posterior, L)	Left	Frozen	Sensory, motor, cognition processing
Nucleus accumbens (anterior, L)	Left	Frozen	Cognition processing, addiction, reward
Substantia nigra	Left	Frozen	Facilitation of movement, also learning, addiction, and emotion
Cerebellar Hemisphere	Right	Frozen	Learned motor sequences, refined motor control
Spinal cord (cervical c-1)	n/k	Frozen	Motor and sensory to/from body, interneurons

To test the hypothesis that genes differentially expressed between SCZ-affected regions and control regions will be enriched for SCZ risk genes, this study focussed on brain regions, allowing for the identification of these potentially relevant genes. From the 11 brain regions offered by GTEx (as detailed in Table 2.2), specific regions were chosen for further analysis in this chapter. Their anatomical placements within the brain are illustrated in Figure 2.2. Two primary considerations guided this selection process. First, cellular composition was a key factor. Brain regions were chosen based on the assumption that they shared a similar makeup of cell types, rather than non-

neural cell types). This approach aimed to minimize potential confounding variables that might arise from significant differences in cellular makeup. For instance, regions primarily composed of neuronal cells were prioritized for inclusion. The second selection criterion focused on preservation methods. Tissue sites were chosen based on the consistent use of the same preservation technique. This ensured a uniform level of RNA quality across the selected brain regions, minimizing potential biases introduced by variations in preservation protocols.

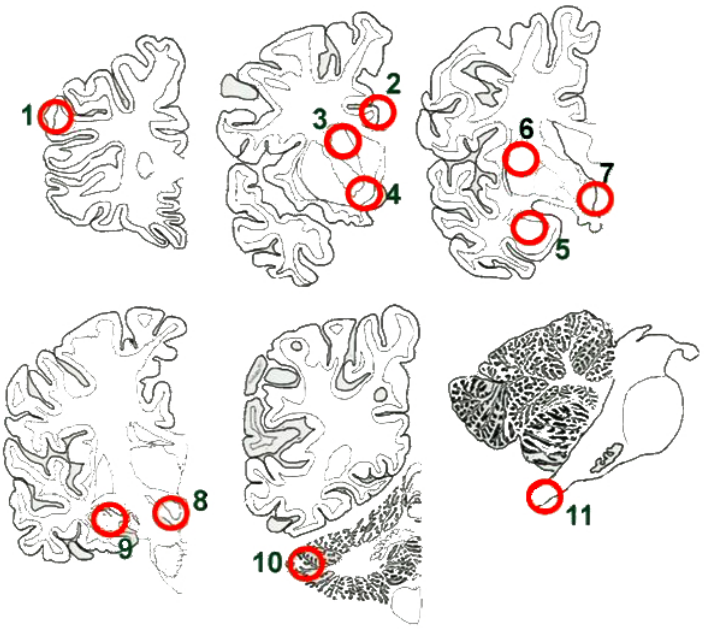

Figure 2.2 GTEx Regions of Interest - Left Hemisphere
At the level of the middle frontal gyrus: (1) Frontal Cortex (BA9). At the level of the anterior striatum: (2) Anterior cingulate cortex (ACC; BA24), (3) Caudate nucleus and (4) Nucleus accumbens (NAcc). At the level of the posterior putamen: (5) Amygdala, (6) Putamen and (7) Hypothalamus. At the level of the midbrain: (8) Substantia Nigra and (9) Hippocampus. (10) Cerebellum. (11) Spinal cord (C1). Image adapted from (Miami Brain Bank, 2015)

2.2.2 Matching GTEx brain regions to brain atlas regions

Where possible, the sampling sites shown circled in red in Figure 2.2 were aligned to more exact subregional names using the Allen Human Brain Atlas (Ding et al., 2016) and Allen Reference Atlas – Human Brain (Allen Institute for Brain Science, 2010).

The names used for brain regions in Table 2.2 represented general anatomical terms used by the GTEx project to describe the sampled areas. However, ensuring a precise correspondence between the brain regions investigated in the reviewed studies and the specific regions sampled by GTEx was important for the literature review undertaken in this chapter.

This matching process is important for several reasons. First, it ensures the accuracy of interpretations drawn when establishing connections between the GTEx data and the findings from the reviewed literature. Any discrepancies in regional definitions could lead to misconstrued relationships between gene expression patterns and specific brain functions relevant to SCZ.

SCZ pathology can manifest differently across distinct subregions within a broader brain area. By achieving a fine-grained correspondence between reviewed studies and GTEx regions, the ranking process could focus on brain regions most relevant to the disease. For example, the GTEx region "Frontal Cortex (BA9)" refers to Brodmann's Area 9 (BA9) of the frontal cortex. The prefrontal cortex was originally defined by Rose & Woolsey (1948, as cited in Carlen, 2017) as the part of the cerebral cortex innervated by the mediodorsal nucleus of the thalamus. Each cortical region has different functionality, for example, emotional processing, memory, or executive function (Carlen, 2017).

2.2.3 Selection of SCZ brain studies from the literature

To explore the degree to which each of the selected GTEx brain regions may be involved in SCZ, data were analysed from a range of studies shown employing various imaging modalities identified in Table 2.1. Primary focus was placed on recent large, collaborative meta- and mega-analyses from the ENIGMA and COCORO consortia. These studies leveraged standardised protocols across multiple sites and encompass numerous brain regions, enabling comparisons for most, but not all, of the GTEx regions. Effect sizes – Cohen's *d* or Hedge's *g* effect sizes – were extracted for the GTEx brain regions. These effect sizes included changes in cortical thinning, surface area, gyrification, subcortical volumes, regional activity, and metabolite concentrations between SCZ cases and healthy controls using magnetic resonance technologies, such as magnetic resonance imaging (MRI), functional MRI (fMRI) and spectroscopy (MRS) respectively. Data were also examined from resting state (functional connectivity)

fMRI (fcMRI) studies for the synchronous correlation of activity between brain regions while the brain was at rest in SCZ cases versus controls. Functional activity and metabolite tracking studies using single-photon emission computerised tomography (SPECT) or functional MRS (fMRS) were reviewed for effects in the 11 brain areas in SCZ. White matter tracts and inter-regional connectivity can be studied by tracing the diffusion of water molecules through axons using diffusion weighted imaging (DWI) or tensor imaging (DTI). Finally, post-mortem studies reveal structural anomalies or asymmetries and allow for the study of the distribution of key receptors in brain tissue using histological techniques.

Results from larger collaborative studies, particularly those focussing on subcortical volumes, were visualised using ENIGMA-TOOLBOX (Lariviere et al., 2021). This visualisation tool displays Cohen's *d* effect sizes on a subcortical structure map with a colour gradient ranging from deep red (increased volume) to white (no change), non-significant difference) to dark blue for a strong reduction in volume in cases vs controls.

2.2.4 GTEx Brain Regions ranking from most affected to least affected

To prioritise brain regions for further investigation, a two-step ranking system was implemented. This system aimed to identify brain regions exhibiting the most prominent and the least prominent alterations in SCZ. First, relevant studies investigating brain structure or function in SCZ were reviewed. Here, effect sizes (Cohen's *d* or Hedge's *g*) were assessed and compared across these studies. These effect sizes quantified the magnitude of differences observed between individuals with SCZ and healthy controls in measures like brain volume or activity. Additionally, other environmental factors that might influence brain function or structure were considered during the review process.

Following this analysis, a scoring system was applied to each selected brain region. Regions exhibiting a significant effect size ($p < 0.05$) exceeding a medium threshold were assigned a higher score, indicating a potentially strong association with SCZ. Conversely, regions with non-significant or weak effects, or those not covered in the reviewed studies, received a lower score.

The total score for each region informed a preliminary ranking, with higher scores suggesting a greater potential for involvement in SCZ. Additionally, a priori knowledge

about the function of each region and its established links to core symptoms of SCZ were considered. Each brain region was plotted against each modality reviewed. The size of the circle reflected the number of large effect sizes. The GTEx regions were ordered by the overall level of evidence for being affected in SCZ based on effect sizes using the sources covered in this thesis. Size of circle shows weight of evidence. Thus, large circles reflect more evidence or larger effect sizes within that modality. Medium circles reflect some evidence but smaller effect sizes or only one or two studies with a large effect. A small circle reflects little evidence or not significant effect sizes. Small blue circles were used where the brain region was not found in any study.

2.3 Results

Brain regions were selected from GTEx RNA-Seq tissue sample sites. Results from recent, large meta-analyses and mega-analyses for structural, functional, and metabolic imaging, along with a review of post-mortem histological studies, were combined to rank these brain regions for their possible early genetic involvement in SCZ. Whilst there is a large *a priori* knowledge of highly affected brain regions (dorsolateral prefrontal cortex, hippocampus, amygdala, and cingulate cortex), ranking the least involved brain regions for the purpose of acting as controls in downstream analyses can be very challenging. In this thesis, a ranking of the GTEx Brain Regions involvement in SCZ is derived from the results of included studies.

2.3.1 Selection of GTEx Brain Regions

The GTEx Project included two brain regions, “Cortex” and “Cerebellum” (Table 2.2) that were sampled from the right hemisphere during the primary donor cadaver processing at the Biospecimen Sampling Site (BSS) and were preserved in PAXgene. In this thesis, these regions were excluded from downstream analyses because they differ in preservation method from all other brain regions. Preservation of tissues in PAXgene versus freezing could potentially introduce a source of technical error into the differential analysis of RNA expression data. In GTEx, the “Frontal Cortex (BA9)” (or “Frontal Cortex”) and “Cerebellar Hemisphere” were sampled close to the sites where “Cortex” and “Cerebellum” respectively were sampled and were preserved frozen along with the remaining brain samples. “Spinal Cord” was excluded from this thesis. The rest of the available frozen brain sites (Table 2.2) were included.

2.3.2 Anatomical location of the GTEx Brain Region sample sites

Table 2.3 shows the approximate locations of each GTEx brain region and any accompanying structures that may have been captured in a 5 mm cube sample. The Allen Human Brain Atlas (Allen Institute for Brain Science, 2010) was used to locate each GTEx sample site to a more precise anatomical structure. This is important because whilst historical MRI and fMRI studies only focussed on macrostructures, more recent studies may include a finer mapping of the heterogeneous landscape of structure and function – especially in the cortical regions.

GTEx provides a histological viewer for viewing all samples of non-brain tissues and the PAXgene-preserved 'Cortex' and 'Cerebellum' samples, but there are no histology slides for the sampled brain regions. In some cases, it was not possible to accurately ascertain which specific nuclei of a subcortical structure or cortical section of a brain region were sampled (including the frontal cortex, cerebellum, amygdala, and hypothalamus).

Table 2.3 Anatomical location of GTEx sample sites in the brain

GTEx Tissue Site	Approximate Brain Site
Frontal Cortex (BA9)	Ventral subdivision of the intermediate portion of the dorsolateral prefrontal cortex at the level of the middle frontal gyrus BA9/46 Rostral-middle-frontal cortex ("R_rostralmiddlefrontal" in FreeSurfer terms)
Anterior cingulate cortex (BA24)	region of the ACC at the level of the head of the caudate and nucleus accumbens
Hippocampus	Hippocampus in the subcortical region of the temporal lobe in line with the subiculum and cingulate tract – GTEx area likely to cover CA1 and parts of CA2/3
Amygdala	Consists of multiple nuclei – no specific nuclei targeted in GTEx
Hypothalamus	No specific nuclei targeted in GTEx
Caudate nucleus	Head of caudate – level with the nucleus accumbens – possibly innervated by the prefrontal cortex or associative cortex
Putamen	Posterior to the NAcc and Caudate nucleus. Possibly innervated by motor neurons.
Nucleus accumbens	Sample may include part of the core and shell of the nucleus accumbens
Substantia nigra	No specific nuclei targeted in GTEx
Cerebellar Hemisphere	lateral pole in the region of Crus I (VIIAf) in the posterior lobe superior to the horizontal fissure

In the following sections, a short summary of the GTEx sample brain regions is provided with their anatomical location and brief description.

2.3.2.1 GTEx: Frontal Cortex (BA9)

The GTEx Frontal Cortex samples comprise a 5 mm cube of cortex tissue taken from Brodmann's area 9 (BA9) at the level of the rostral subdivision of the middle frontal gyrus (MFG) in the right hemisphere (Area 1 in Figure 2.2; Figure 2.3), which is an area that lies in the right dorsolateral prefrontal cortex (DLFPC) (GTEx Symposium, 20 May 2015). The Nissl-staining of nucleic acids (DNA and RNA) in cortex from BA9 (inset in Figure 2.3) shows the cells arranged in vertical columns in several distinct layers. The DLPFC is a region implicated in SCZ and is composed primarily of glutamatergic pyramidal neurons and GABAergic interneurons. This composition is crucial for balanced excitatory and inhibitory neurotransmission, which is a major area of

investigation in SCZ (Glausier & Lewis, 2018). Additionally, the DLPFC receives dopaminergic innervation and expresses dopamine receptors.

2.3.2.2 GTEx: Anterior cingulate cortex (BA24)

The GTEx Anterior cingulate cortex was sampled from a 5 mm cube of cortex tissue taken from caudal region of Brodmann's area 24 (BA24) shown in Figure 2.4 and Figure 2.5. The Nissl-staining of nucleic acids (DNA and RNA) in cortex from BA24 (inset in Figure 2.5) shows the cells arranged in vertical columns in several distinct layers like the frontal cortex. The anterior cingulate cortex (ACC) is a limbic system structure involved in emotion, affect, decision-making, and attention. The ACC has been implicated in schizophrenia, with some studies suggesting dysfunction in this region contributes to difficulties with social interaction (Kerns et al., 2005; Park, I. H. et al., 2008). It is also thought to be involved in autonomic functionality, such as regulating blood pressure and heart rate (Garrett et al., 2022).

Figure 2.3 Ventral Dorsolateral Prefrontal Cortex (BA9/BA46) (red, top right)
Allen Brain Atlas – Human Brain: Modified Brodmann view (Image 13) of the GTEx region "Front Cortex (BA9)". Inset is a 5 mm square section of Nissl-stained cortex tissue from the DLPFC ventral to the intermediate frontal sulcus. Part of the anterior cingulate cortex (A24) is highlighted in red on the medial line of the brain (left) surrounding the corpus callosum ('ccg') (Images credits: Allen Institute. Allen Human Brain Atlas (2010)).

Figure 2.4 highlights these two cortical regions, divided into their caudal and rostral subdivisions, on a cortical structure map (Lariviere et al., 2021). The parahippocampal cortex is marked for reference to indicate where the hippocampus is in relation to these two areas. The following sections describe these two regions.

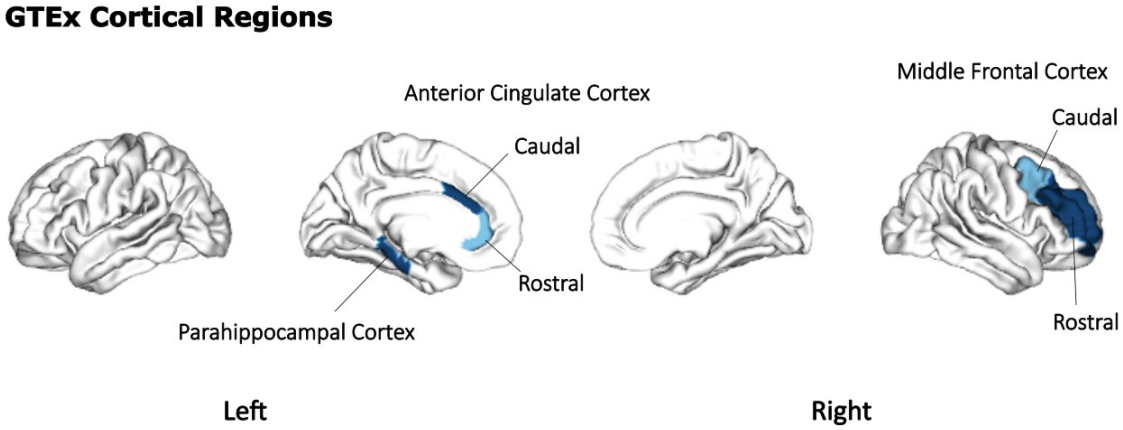

Figure 2.4 GTEx cortical regions
Cerebral cortex regions sampled in GTEx. Views show the left lateral, left medial, right medial, and right lateral perspectives. "Frontal Cortex" corresponds to the rostral subdivision of the middle frontal cortex (dark blue) in the right lateral view. GTEx "Anterior Cingulate Cortex" refers to the posterior part of the caudal ACC (dark blue) shown in the left medial view. Image generated using custom data in ENIGMA-TOOLBOX (Lariviere et al., 2021)

2.3.2.3 GTEx: Caudate Nucleus

The GTEx Caudate Nucleus region (Area 3 in Figure 2.2 and Figure 2.5) comprised a 5mm cube region sampled from the head of the caudate nucleus, which is part of the striatum (Figure 2.5). The striatum is a brain region involved in motor control and reward processing, receiving input from various cortical regions. The caudate itself is primarily composed of medium spiny neurons (MSNs). The progressive loss of MSNs in Huntington's disease leads to a breakdown of specific cortical loops, the order of which underlies the progression of the symptoms. In HD, this begins with executive loops, then motor loops, then finally the motivational loop with accompanying emotional disturbances (Haber, 2016).

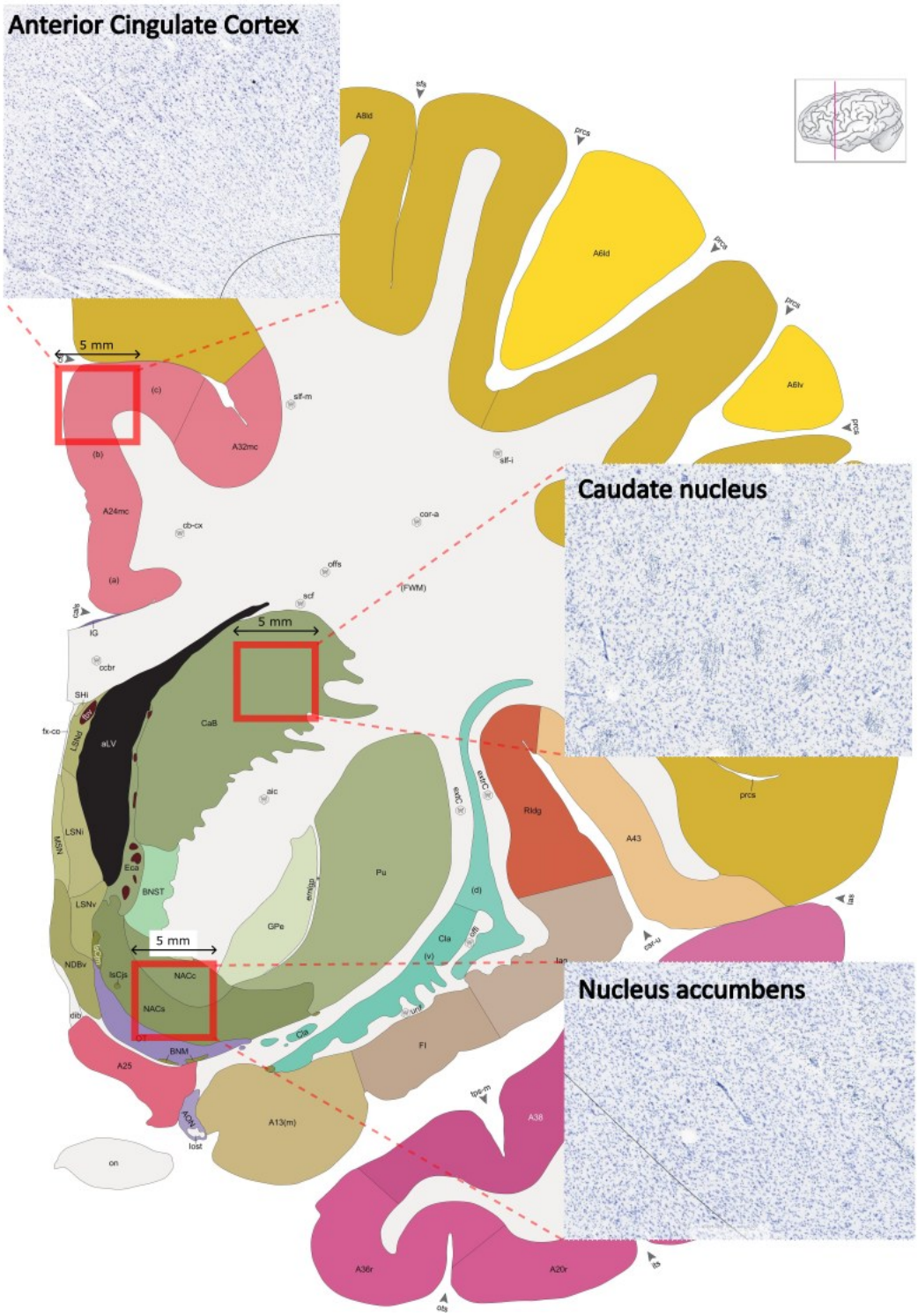

Figure 2.5 GTEx regions: Anterior Cingulate Cortex (BA24), Caudate nucleus, and Nucleus accumbens

Allen Brain Atlas – Human Brain: Modified Brodmann view (Image 20) of the GTEx regions "Anterior Cingulate Cortex (BA24)" (upper left), "Caudate Nucleus" (middle), and "Nucleus accumbens" (lower left). Insets are 5 mm square sections of Nissl-stained cortex tissue from each sampled region. (Images credits: Allen Institute. Allen Human Brain Atlas (2010)).

2.3.2.4 GTEx: Nucleus accumbens

The GTEx 'Nucleus accumbens' region likely originated from a region encompassing both the core and shell subregions of the Nucleus accumbens (NAcc) (Area 4 in Figure 2.2 and Figure 2.5). The NAcc is part of the striatum and is involved in reward processing, motivation, and emotion regulation. It receives inputs from various cortical regions, including the prefrontal cortex, and the limbic regions like the amygdala and substantia nigra. Dysfunction in the NAcc, including its cortical inputs and communication with other limbic regions, has been implicated in SCZ.

2.3.2.5 GTEx: Putamen

The GTEx Putamen samples originated from a posterior region of the striatum compared to the Caudate nucleus and nucleus accumbens (Area 6 in Figure 2.2; and Figure 2.6). The putamen , like other striatal regions, is involved in motor control and reward processing. It receives cortical input, with posterior regions more likely to be innervated by the motor cortex. While this motor input may be relevant to Parkinson's disease, it is less directly implicated in SCZ (Haber, 2016). However, the putamen's role in reward processing and its dopaminergic innervation from the substantia nigra warrant further investigation in the context of SCZ.

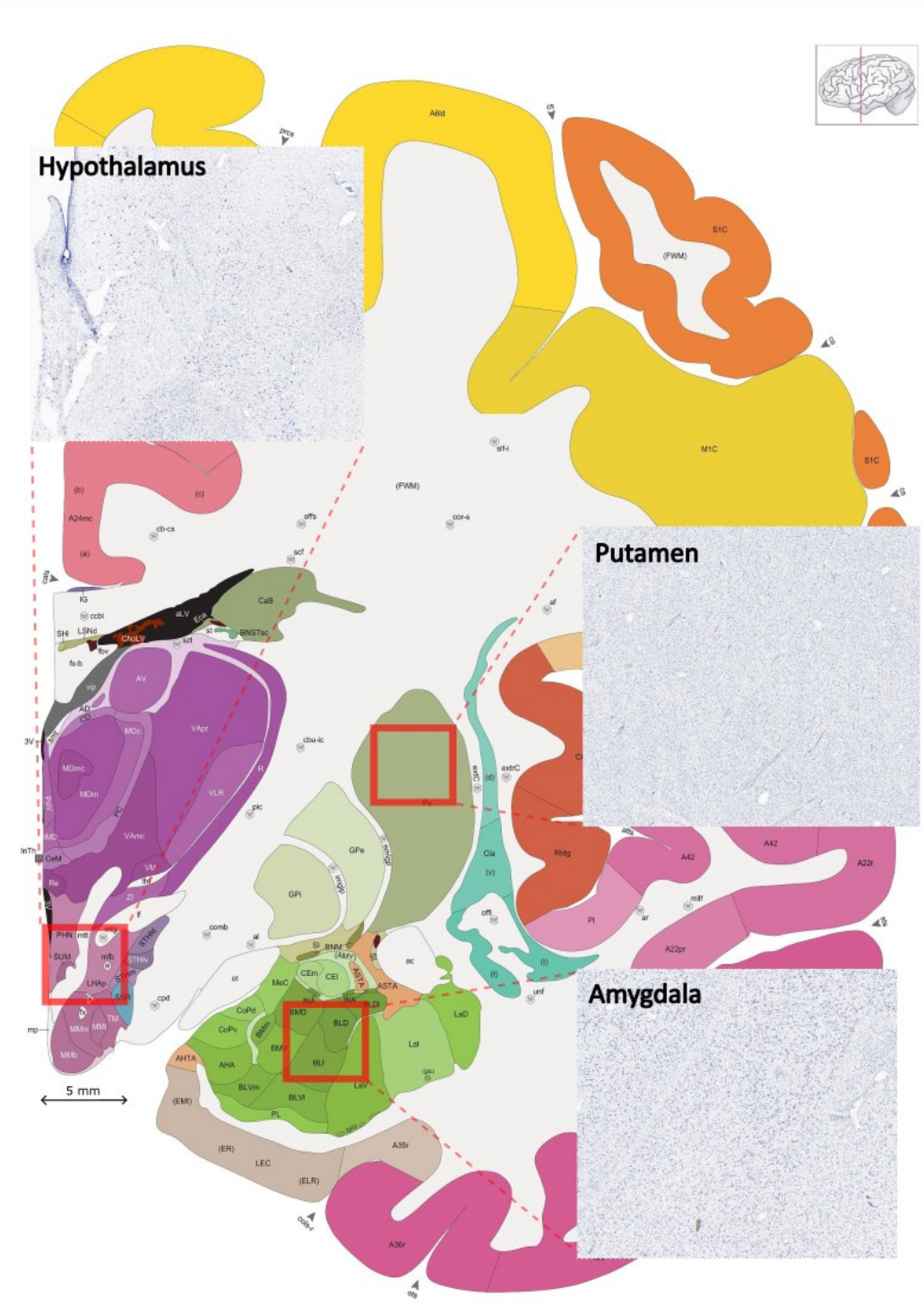

Figure 2.6 GTEx regions: Putamen, Amygdala and Hypothalamus
Allen Brain Atlas – Human Brain: Modified Brodmann view (Image 38) of the GTEx regions “Putamen” (right), “Hypothalamus” (top left), and “Amygdala” (bottom right). Insets are 5 mm square sections of Nissl-stained cortex tissue from each sampled region. (Images credits: Allen Institute. Allen Human Brain Atlas (2010)).

2.3.2.6 GTEx Region: Amygdala

The GTEx "Amygdala" (Area 6 in Figure 2.2; and Figure 2.6) samples originated from a 5mm cube region from the amygdala nuclei. The amygdala is a brain region involved in processing emotions and emotional memories (Meisner et al., 2022), and fear responses (Ressler, 2010). It has connections to various parts of the limbic system and cortices (Janak & Tye, 2015; Sah et al., 2003). Dysfunction in the amygdala has been implicated in schizophrenia, potentially contributing to negative symptoms and emotional processing difficulties.

2.3.2.7 GTEx Region: Hypothalamus

The GTEx "Hypothalamus" (Area 7 in Figure 2.2; and Figure 2.6) samples originated from a 5mm cube region within the hypothalamic nuclei. The hypothalamus is a vital region connecting the brain to the endocrine system and is involved in stress response regulation (Pruessner et al., 2017). Dysfunction in the hypothalamus, particularly within the hypothalamic-pituitary-adrenal (HPA) axis, has been implicated in SCZ (Pruessner et al., 2017).

2.3.2.8 GTEx Region: Hippocampus

The GTEx Hippocampus region (Area 7 in Figure 2.2; and Figure 2.7) samples originated from a 5mm cube region taken from the hippocampal formation in the temporal cortex. The hippocampus is a crucial brain region for learning and memory, particularly involved in consolidating short-term memories into long-term storage in the cortex. Dysfunction in the hippocampus has been implicated in schizophrenia, with studies showing abnormalities in hippocampal structure and function in patients. These abnormalities may contribute to memory difficulties, a common symptom in SCZ (Ragland et al., 2016).

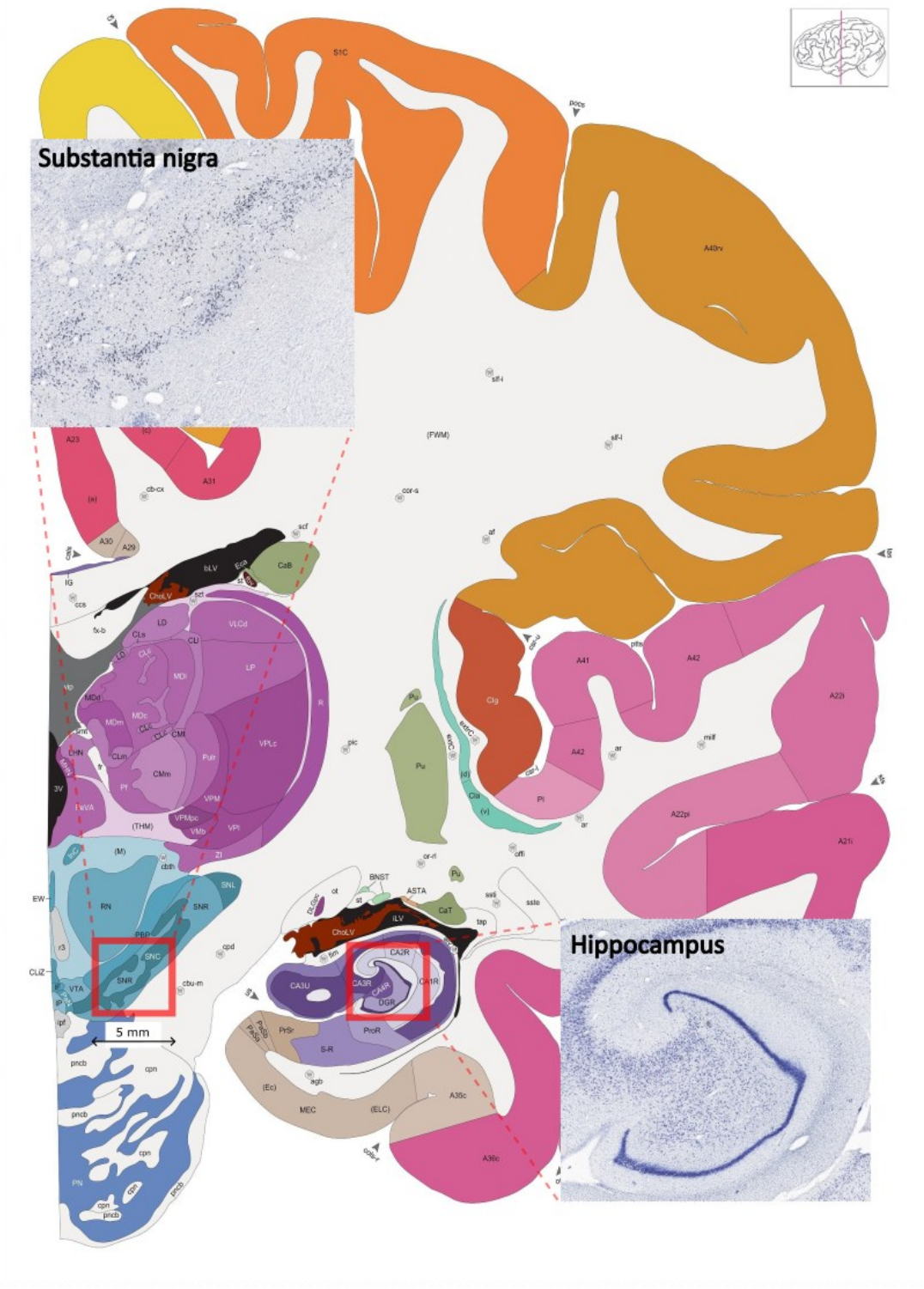

Figure 2.7 GTEx regions: Hippocampus and Substantia nigra
Allen Brain Atlas – Human Brain: Modified Brodmann view (Image 51) of the GTEx regions "Hippocampus" (right) and "Substantia nigra" (left). Insets are 5 mm square sections of Nissl-stained cortex tissue from each sampled region. (Images credits: Allen Institute. Allen Human Brain Atlas (2010)).

2.3.2.9 GTEx Region: Substantia nigra

The GTEx "Substantia nigra" region (Area 8 in Figure 2.2; and Figure 2.7) originated from a 5mm cube region within the midbrain (GTEx Consortium, 2013). The substantia nigra is a midbrain region containing dopaminergic neurons that project to various brain areas, including the cortices and the striatum (caudate nucleus, putamen, nucleus accumbens). Dopamine signalling plays a crucial role in reward processing, movement, and motivation. Dysfunction in the dopamine system, potentially including the substantia nigra, has been implicated in SCZ (reviewed in Schulz et al., 2022).

2.3.2.10 GTEx Region: Cerebellar Hemisphere

The cerebellum is a hindbrain region involved in motor control, potentially including fine motor skills and coordination. The GTEx "Cerebellar Hemisphere" region (Area 10 in Figure 2.2; and Figure 2.8) samples originated from a 5mm cube region of the right cerebellar cortex. The sample includes a dense layer of granular cells and white matter tract. A sample of this volume may include tissue from both the superior and inferior lobes of the cerebellar hemisphere (VIIAf/crus I and VIIAt/crus II – the folium, and tuber semilunar lobules). It was not possible to determine which specific areas were sampled from in the Crus I or Crus II regions. There is ongoing debate into the cerebellum's role in cognitive function and its potential connection to SCZ (Moberget, Torgeir & Ivry, 2019). Some but not all cases of SCZ may have cerebellar involvement, which may be indicated if there are motor-related symptoms.

Figure 2.8 Cerebellar lateral hemisphere cortical structure – Nissl-stained
Allen Brain Atlas – Human Brain: Modified Brodmann view (Image 84) of the GTEx regions “Cerebellum” (right). Inset is an example 5 mm square section of Nissl-stained cerebellar cortex tissue from the sampled region. (Images credits: Allen Institute. Allen Human Brain Atlas (2010)).

2.3.3 Selection of Brain Imaging Literature

For this chapter, the aim was to focus on large, collaborative brain imaging studies that covered as many of the brain regions as possible to enable comparisons between them. Early studies focussed on only one or two regions. Large MRI brain imaging studies now use standardised protocols and have large enough cohorts to enable robust comparisons of data collected in different centres.

Table 2.4 shows the studies that were selected for review in chronological order of the study. Shaded rows show the large meta- and mega-analyses with the greatest number of brain regions in the same study. In most cases the substantia nigra and cerebellum were not included.

Table 2.4 Imaging studies reviewed
List of imaging studies reviewed. Primary focus was on the large collaborative studies that are highlighted. CTL - Healthy control; DF/DN – Drug-free/drug-naïve; FEP – first-episode; APN – anti-psychotic-naïve.

Modality	Description	Study
SPECT	Cerebral blood flow in drug-free and drug-naïve SCZ DF: 8; DN: 9; CTL: 12	(Vita et al., 1995)
MRI	Regional brain volumes – meta-analysis SCZ: 1,588; CTL:	(Wright, I. C. et al., 2000)
MRS	Literature Review of 64 proton MRS for N-acetylaspartate (NAA) SCZ: 1256; CTL: 1209	(Steen et al., 2005)
MRI	Hippocampal volumes in early and late prodromal states (EPS/LPS) EPS/LPS: 36; CTL: 30	(Hurlemann et al., 2008)
MRI	Gray matter volume in ultra-high risk (UHR) and first-episode SCZ patients (FEP) UHR: 30; FEP: 23; CTL: 29	(Witthaus et al., 2009)
MRI	Hippocampus and Amygdala in UHR and FEP SCZ UHR: 29; FEP: 23; CTL: 29	(Witthaus et al., 2010)
MRI	Meta-analysis of brain volumes + anti-psychotics	(Haijma et al., 2013)

Modality	Description	Study
	Medicated SCZ: 8327; APN: 771; CTL: 8292/939	
MRI	Hippocampus and Amygdala SCZ: 29	(Ganzola et al., 2014)
MRI GWAS	Common genetic variants influencing subcortical brain volumes 30,717 individuals	ENIGMA (Hibar et al., 2015)
MRI	Gray matter volume in first-episode SCZ with prominent hallucinatory symptoms FEP: 18; CTL: 18	(Huang, P. et al., 2015)
MRI	Subcortical volume abnormalities in SCZ SCZ: 2028; CTL: 2540	ENIGMA (van Erp et al., 2016)
MRI	Abnormal asymmetries in subcortical volumes in SCZ SCZ: 884; CTL:1680	COCORO (Okada et al., 2016)
fMRI	Effect of SCZ on Spatial Memory and Item Memory Tasks SCZ: 24; CTL: 26	(Ragland et al., 2016)
fcMRI	Functional connectivity/resting state MRI and auditory hallucinations FEP DN: 36; CTL: 18	(Chang, Xiao et al., 2017)
MRI GWAS	Common genetic variants influencing hippocampal volume 33,536 individuals	(Hibar et al., 2017)
MRI	Cortical brain abnormalities meta-analysis in SCZ SCZ: 4474; CTL: 5098	ENIGMA (van Erp et al., 2018)
MRS	Brain metabolism with high-strength MRS in FEP psychosis FEP: 81; CTL: 91	(Wang, A. M. et al., 2019)

2.3.4 Cortex thickness and subcortical volumes are affected by multiple factors

Brain volumes vary between individuals due to multiple factors other than disease. Like the organs of human physiology, brains show a large amount of variability

between individuals. Causes of variability in brain structures may include genetic variation (Hibar et al., 2015, 2017), medication, and urbanicity (Besteher et al., 2017; Haddad et al., 2015). Volumes for 7 subcortical structures, corrected by intracranial volume (ICV), were analysed in 30,717 individuals from 50 cohorts in an imaging GWAS study (Hibar et al., 2015). The study identified 5 novel variants that influenced putamen and caudate nucleus volumes. For example, an intergenic locus upstream from the gene for kinectin (*KTN1*) tagged by the common variant rs945270 showed the strongest effect on putamen volume ($p=1.08 \times 10^{-33}$; 0.52% of variance explained). The remaining variants affecting putamen volume were clustered near developmental genes that regulate apoptosis, axon guidance and vesicle transport. 3 known loci were also strongly associated with hippocampal volume and ICV. Variant effects were specific to the subcortical structure, rather than affecting all structure globally. In a further GWAS on hippocampal volume in 33,536 individuals from 65 centres around the world, the authors with ENIGMA and the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) looked for genetic determinants of hippocampal volume (Hibar et al., 2017). The authors state that heritability of hippocampal volume is around 70% and estimate that 18.76% of the variance seen could be explained by common genetic variants (Hibar et al., 2017).

2.3.5 Ranking of the GTEx Brain Regions

Figure 2.9 illustrates the ranking of the brain regions analysed in this study based on the strength of evidence from large-scale brain imaging studies comparing SCZ cases with controls. Studies were included where they contained a large enough cohort of SCZ and healthy controls, and as many of the brain regions as to allow for robust comparisons (studies highlighted in Table 2.4). In many of these studies the Substantia nigra and the Cerebellum were not included. While all regions showed some evidence for involvement, the frontal cortex (FC), hippocampus (HP) and anterior cingulate cortex (ACC) emerged at the top with the most compelling evidence. The evidence for the remaining brain regions, including the amygdala (AMYG), caudate nucleus (CN), putamen (PUT), nucleus accumbens (NAcc), was stronger than for the hypothalamus (HYP), cerebellum (CEREB), or substantia nigra (SN), but there was not enough evidence to support making a scientific selection of the two control regions based on

these data alone. Therefore, the final selection of control regions was based on additional biological considerations., for example

- CEREB and SN are more commonly associated with motor coordination or motor control, respectively
- Lesions in each region does not tend to cause psychosis-like symptoms; the SN, is more associated with Parkinson's disease or Huntington's disease.

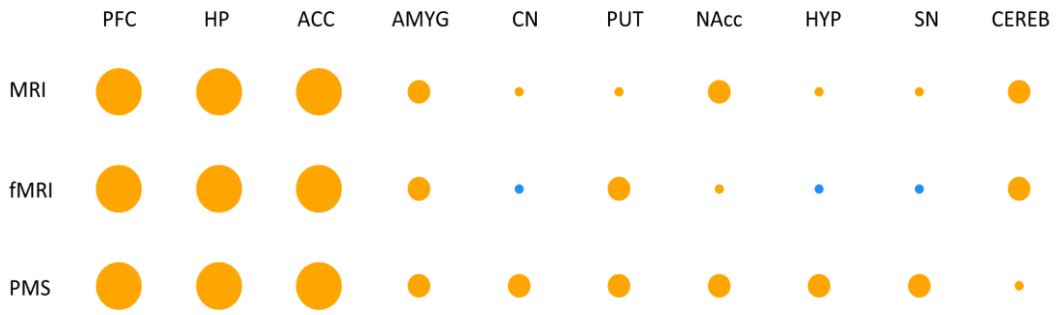

Figure 2.9 Evidence matrix showing modalities reviewed
Brain areas are ordered by the overall level of evidence for being affected in SCZ using the sources covered in this thesis. Size of circle shows weight of evidence – large: lots of evidence or large effect sizes; medium: some evidence but smaller effect sizes; small orange: little evidence or not significant; small blue: not included in the large-scale studies reviewed or no suitable studies identified. PFC: prefrontal cortex; HP: hippocampus; ACC: anterior cingulate cortex; AMYG: amygdala; CN: caudate nucleus; PUT: putamen; NAcc: nucleus accumbens; HYP: hypothalamus; SN: substantia nigra; CEREB: cerebellum; MRI: magnetic resonance imaging; fMRI: functional MRI; PMS: post-mortem studies.

2.3.6 Frontal Cortex and Hippocampus are the two most affected brain regions

Large collaborative subcortical MRI studies provide strong evidence for hippocampal reduction in size in SCZ compared to controls (Haijma et al., 2013; Wright, I. C. et al., 2000). The meta-analysis of structural volumes by Haijma et al. (2013) was one of the first large-scale, cross-sectional, retrospective meta-analyses including over 9000 patients with SCZ and 9000 HCs from 317 studies. Their data included anti-psychotic naïve patients (n=771) and patients with a wide variation in illness duration. The authors also considered cofactors such as duration of illness, use of medication, sex, IQ scores and laterality. The reduction in hippocampal volume in patients with SCZ were largest of all brain regions (Cohen's d L: -0.51, R: -0.44; Figure 2.10). Medium reductions in volume were also seen in the amygdala (Cohen's d L: -0.30, R: -0.27;

Figure 2.10). Large reductions were also seen in the thalamus, which was not sampled in the GTEx project.

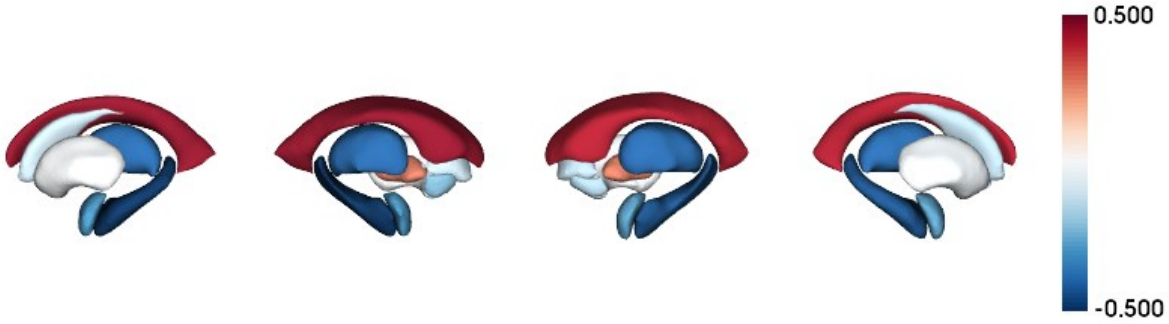

Figure 2.10 Haijma et al. (2013) effect sizes for subcortical structures
Cohen's d effect sizes for subcortical volumes in SCZ subjects compared to HCs from Haijma et al. (2013) plotted on subcortical structure map. Views show left lateral/medial and right medial/lateral. Colours show a gradient of effects sizes ranging from deep red for increased volume to white for no change (or small, non-significant difference) to dark blue for a strong reduction in volume in SCZ. The hippocampus (dark blue) (L d=-0.51; R d=-0.44), thalamus (L&R d=-0.40) and amygdala (L d=-0.30; R d=-0.27) show the strongest reductions in volume. Image created using ENIGMA-TOOLBOX.

The first study from ENIGMA's Schizophrenia Working Group (van Erp et al., 2016) was a cross-sectional, prospective meta-analysis combining MRI data from 2028 subjects with SCZ and 2540 HCs from 15 international centres. Overall, volume reductions, visualised in Figure 2.11, were seen in the hippocampus (Cohen's d=-0.43; 95% CI:), amygdala (d=-0.38), thalamus (d=-0.34) and nucleus accumbens (NAcc; d=-0.20) in patients with SCZ vs healthy volunteers with little signs of lateralisation effects (Figure 2.11). The lateral ventricles showed a strong increased effect on volume (Cohen's d=0.34). The caudate and putamen displayed little or no significant difference. COCORO's independent study collaborating across 11 centres in Japan replicated these findings in 884 subjects with SCZ and 1680 HCs (Okada et al., 2016). The left pallidum appears to show a larger increase in volume in COCORO, but there was greater variation in the results and the volume increase of the pallidum was not significant overall. These results add evidence to the trend seen in Haijma et al. (2013) (Figure 2.10) in which the samples sizes were not large enough to attenuate the variability between individuals caused by other factors. In all three studies, effects were also

seen for medication usage, illness duration and the study centre that did the original imaging.

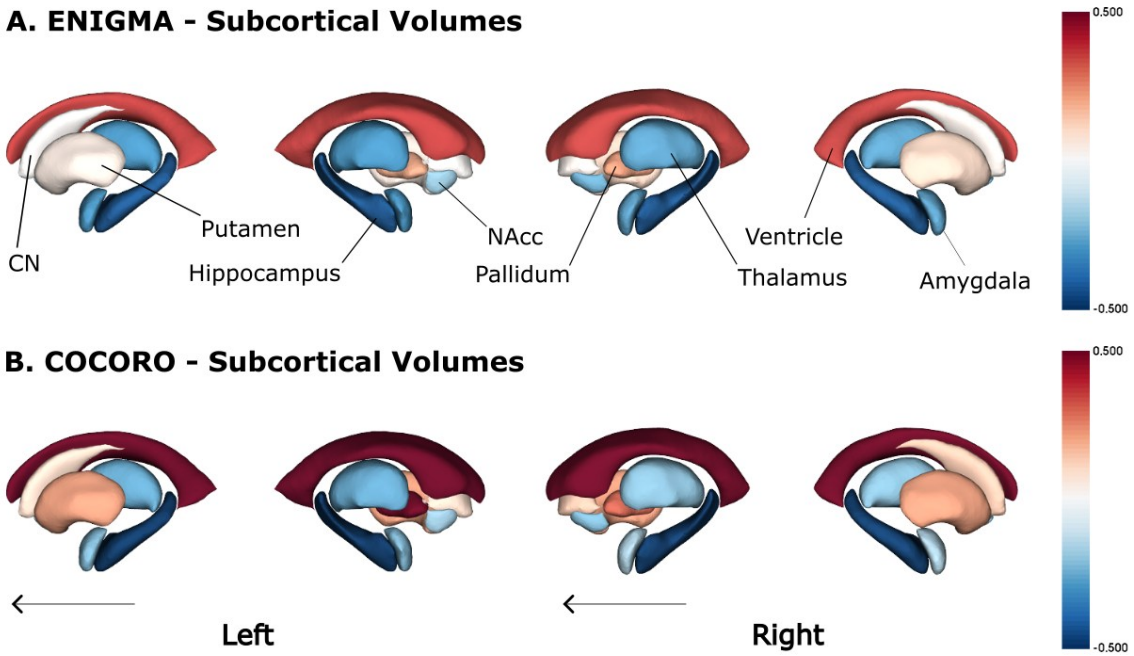

Figure 2.11 ENIGMA and COCORO: Changes in subcortical volumes between patients with SCZ and healthy controls
Cohen's d effect sizes for left and right hemispheres from A) ENIGMA Subcortical Volumes (van Erp et al., 2016) and B) COCORO Subcortical Volumes (Okada et al., 2016) plotted on Freesurfer parcellated and segmented subcortical regions using ENIGMA TOOLBOX (Lariviere et al., 2021). Red regions have an increased volume, and blue regions have reduced volume in patients with SCZ vs healthy volunteers (CN=Caudate nucleus).

Frontal cortex continues to reduce in volume as the disease progresses, but subcortical volumes do not show any significant changes in volume. Figure 2.12 shows changes in subcortical volumes of GTEx regions over the duration of SCZ compared to health controls from Olabi et al. (2011), a review of longitudinal MRI volumetric studies. The meta-analysis covered 27 studies with a total of 928 cases and 867 controls. The 32 areas compared included GTEx regions: cerebellum, hippocampus, amygdala, caudate and a general coverage of frontal cortex as 'frontal lobe tissue'. The left caudate showed a small reduction in volume (standard mean difference (SMD): -0.336; p value: 0.013), but a larger effect was seen in frontal lobe tissue (SMD: -0.478; p value: 0.002)

between patients with SCZ and controls. Global brain volume also continued to decrease, and ventricular volume continued to increase over the progression of SCZ.

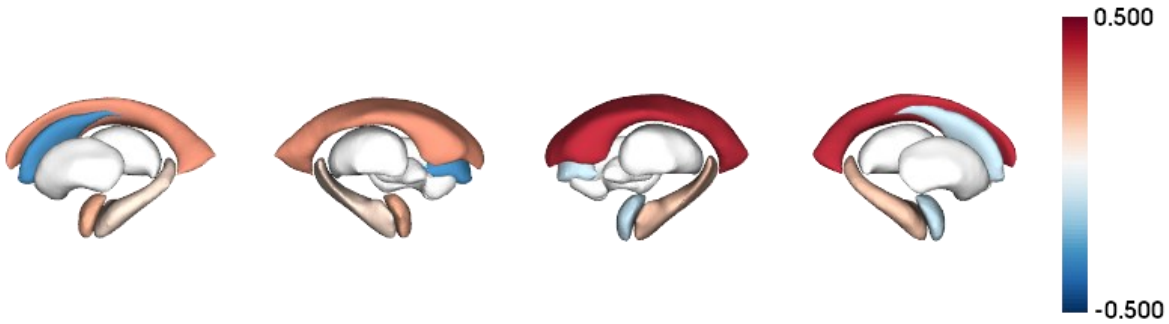

Figure 2.12 Longitudinal changes in subcortical volumes between patients with SCZ and healthy controls (Olabi)
Standard mean differences (SMD) for longitudinal changes in left and right hemispheres from Olabi et al. (2011) plotted on Freesurfer parcellated and segmented subcortical regions using ENIGMA TOOLBOX (Lariviere et al., 2021). Red regions have an increased volume, and blue regions have reduced volume in patients with SCZ vs healthy volunteers. Putamen, pallidum, thalamus, and nucleus accumbens were not included in this meta-analysis. Only 'right lateral ventricle' and 'left caudate' effects were significant ($p<0.05$).

The ENIGMA cross-sectional, prospective meta-analysis of cortical thickness and surface area between SCZ and HCs included imaging data from 4,474 patients with SCZ and 5,098 HCs across 39 international centres (van Erp et al., 2018). All regions of the cerebral cortex showed differing degrees of thinning in patients with SCZ compared to HCs with the smallest reduction seen in the left caudal anterior cingulate cortex (L-cACC; Cohen's d=-0.119; SE=0.036; FDR $p=9.23x10^{-4}$) and the largest in the right fusiform cortex (d=-0.536; SE=0.043; FDR $p=5.22x10^{-34}$). The study covers the two GTEx cortical regions "Anterior Cingulate Cortex", which corresponds to the L-cACC, and "Frontal Cortex" which roughly corresponds to "R_rostralmiddlefrontal" (d=-0.362; SE=0.056; FDR $p=1.07x10^{-10}$).

The amygdala was also significantly reduced in volume (a mean 3.8% reduction) in the ENIGMA meta-analysis of MRI subcortical volumes (van Erp, Theo G. M et al., 2016) and replicated with a smaller effect in the COCORO study (Okada et al., 2016). Meta-analyses that had smaller cohorts found small to large differences in volume but were

not significant after adjusting for multiple testing (Olabi et al., 2011; Wright et al., 2000). Reduced amygdala volumes were more often observed in early-course SCZ compared to chronic SCZ or high-risk relatives and healthy controls (Rich et al., 2016). While there is evidence for the involvement of the amygdala in SCZ, it is not as heavily involved in SCZ as the hippocampus and DLPFC.

2.3.7 Cerebellum and Substantia nigra as the two least affected brain regions

Identifying the least affected brain regions is challenging because some evidence suggests involvement of all GTEx regions in SCZ. Studies like Moberget et al. (2018) observed changes in various brain areas, including the cerebellum. However, Olabi et al. (2011) found no significant volume changes in the cerebellum, suggesting potential indirect effects. Additionally, limitations in studying the substantia nigra due to iron content and differentiation from neighbouring regions make it difficult to draw definitive conclusions (de Hollander et al., 2015; Schulz et al., 2022).

However, the hypothalamus presents another challenge for definitive conclusions. While imaging studies show less evidence for its involvement, the stress diathesis model suggests its role through glucocorticoid (GC) regulation. GCs released from the adrenal glands act on the hypothalamus, which has GC receptors that regulate the stress response (Pruessner et al., 2017). In SCZ, this model proposes that the stress response could lead to further cognitive dysfunction through GC over-activity at the hippocampus (Pruessner et al., 2017).

2.4 Discussion

2.4.1 Most affected: Hippocampus and Frontal Cortex

The analysis of the GTEx brain regions, which employed a ranking system based on the strength of evidence, identified the Frontal Cortex and Hippocampus as the brain regions with the most compelling evidence for being involved in SCZ. This finding was supported by established knowledge in the field, as prior studies had highlighted the critical role of these regions in SCZ (Bahner & Meyer-Lindenberg, 2017; Glausier & Lewis, 2018; Smucny et al., 2022).

Further insights into the potential progression of SCZ come from studies by Ho et al. (2017) and Lieberman et al. (2018). These studies suggest that early SCZ might involve glutamatergic neurotransmission dysregulation in the CA1 sub-region of the

hippocampus. This initial dysfunction could lead to weak psychotic symptoms and eventually progress to a first episode of psychosis encompassing other hippocampal sub-regions and the frontal cortex (Ho et al., 2017; Lieberman et al., 2018). The hippocampus is known to regulate dopamine release in the substantia nigra, and increased glutamatergic signalling in the frontal cortex could potentially lead to dopamine overproduction, further impacting the frontal cortex in SCZ (Heckers & Konradi, 2013).

2.4.2 Least affected: Cerebellar Hemisphere and Substantia nigra

The brain imaging analysis identified the cerebellum and substantia nigra as two of the GTEx brain regions with the weakest evidence for direct involvement in SCZ. This finding aligns with the hypothesis that these regions may be passively modulated by the more affected brain regions, particularly the frontal cortex.

2.4.2.1 Cerebellar Hemisphere

The cerebellum is known for its role in motor control, coordination, and learning. While some studies have observed changes in the cerebellum in SCZ patients, for example, Moberget et al., (2018), these might be indirect consequences of primary dysfunction in other regions, especially the frontal cortex, which is also involved in motor function. Additionally, studies have reported reduced white matter integrity in SCZ (Ohtani et al., 2014), which could disrupt communication between the cortex and cerebellum, further contributing to cerebellar dysfunction without the cerebellum itself being directly involved.

2.4.2.2 Substantia nigra

The Substantia nigra (SN) is a key player in the dopaminergic system, which has been implicated in SCZ. However, the limitations in studying the SN due to iron content and differentiation from neighbouring regions (de Hollander et al., 2015; Schulz et al., 2022) make it difficult to draw definitive conclusions about its direct involvement. The hypothesis suggests that initial changes in the frontal cortex, which is known to regulate dopamine signalling (Heckers & Konradi, 2013), could indirectly impact the SN.

2.4.2.3 Limitations

The selection of negative control regions was complicated by the lack of inclusion of several brain regions in many of the large-scale case-control cohort studies (Figure 2.9). The hypothalamus (in GTEx), cerebellum and substantia nigra were not included in ENIGMA and COCORO studies available at the time of this study. Therefore, it was not possible to robustly compare differences between these brain regions and the changes seen in other brain regions within the same study.

New MRI methods have been developed for imaging neuromelanin and nigrosomes in the substantia nigra (Le Berre et al., 2019). These advancements open new opportunities for studying differences between SCZ and controls.

Historically, differences in hypothalamus volumes between SCZ and healthy controls have contradicted each other (reviewed in Ruggeri et al., 2024). Ruggeri et al. (2024) (published after the analyses in this thesis were completed) analysed differences in the volumes of five hypothalamic subnuclei compared to each other and the total hypothalamic volume. Volumes were extracted from MRIs of 344 patients with SCZ, 340 patients with bipolar disorder, and 684 age- and sex-matched healthy controls. Significant reductions in volume were observed for three subnuclei, and an increase was observed in one subnucleus, whereas no significant changes were found in overall hypothalamic volume.

Another confounding factor in determining whether brain regions can or cannot be used as control regions is that damage in one brain region can exert effects on other brain regions. For example, reductions in volume of cerebellum may be observed in different layers of cerebellar hemispheres, which also show corresponding reductions in correlating regions of the frontal cortex (Olabi et al., 2011). The reductions in cerebellum may be an indirect consequence of loss of projected axons from the frontal cortex.

It is also important to acknowledge that antipsychotic medication use is a potential confounder when it comes to variation in brain volume. Anti-psychotic use could have influenced the observed brain volume changes. Future research that incorporates longitudinal data and tracks medication use in SCZ patients would be valuable to further elucidate the independent effects of antipsychotic medication and SCZ

pathology on brain structure. For instance, a longitudinal study following medication-naive patients from the time of diagnosis could provide more definitive insights.

3 RNA-Seq Differential Expression Analysis of Brain Regions

3.1 Background

3.1.1 Gene Expression

The central tenet of biology is that DNA is transcribed to RNA and RNA is translated to protein. Gene expression is the production of proteins from DNA (Buccitelli & Selbach, 2020). Protein-coding genes only make up about 1-2% of the whole genome; and only a subset of those genes is expressed in any one cell. Furthermore, genes may be spliced to form alternative transcripts of the same gene, which are translated into different isoforms of the final protein. RNA polymerases transcribe the gene from DNA into a precursor messenger RNA (pre-mRNA) template. The pre-mRNA template is immediately read by a splicing complex (spliceosome) that splices out non-coding parts of the gene sequence and retains coding regions required to make the final protein (Wilkinson et al., 2020). Protein coding regions of the gene are called exons which are separated by non-coding regions called introns. These mature mRNA molecules are transported out of the nucleus to be translated into proteins at ribosomes. The genes that are expressed in any cell and the isoforms produced from these genes are determined by the microenvironment around the cell, the needs of the cell, and the functions that the cell performs within the tissue (Wang, E. T. et al., 2008). The complete set of all mRNA transcripts, coding or non-coding, is called the transcriptome. Each cell can be said to have a gene expression profile, which is the set of genes that are actively being expressed at any given moment. This gene expression profile is thought to determine the outward characteristics or traits displayed by the cell, which is referred to as its phenotype. For example, genes active in neurons are likely to be involved in the transmission of signals, the maintenance of synapses and the trafficking of proteins involved in synaptic transmission of neurotransmitters. Genes less likely to be active in neurons are those related to cell growth and division because most neurons are senescent. By studying the gene expression profile of a cell, it is possible to gain insights into the underlying molecular mechanisms that underlie its phenotype. Gene expression profiles can be studied through techniques such as RNA-Sequencing or microarray analysis (Van den Berge et al., 2019).

3.1.2 RNA-Sequencing

The development of massively parallel, high-throughput sequencing in 2008 enabled the study of larger volumes of DNA, RNA as complementary DNA (cDNA), and protein binding sites (CHiP-Seq, histone modifications) (Mortazavi et al., 2008; Wang, Z. et al., 2009). Over the last decade RNA-Sequencing (RNA-Seq) has become a widely used method for the study of transcriptomics and profiling gene expression. In RNA-Seq mRNA molecules are extracted from cells in a tissue sample, fragmented and then converted into cDNA molecules (Wang, Z. et al., 2009). The cDNA molecules are sequenced using high-throughput sequencing, in which millions of DNA (or cDNA) molecules can be sequenced in parallel so that a snapshot of gene expression in cells of a tissue sample can be read in a single experiment (Wang, Z. et al., 2009). The sequencing machine outputs short sequences (reads) that have been read from each fragment of mRNA (as cDNA). In single-ended sequencing only one file of reads is produced. Sequencing read errors tend to accumulate as sequencing progresses beyond 100 bp. In paired-end sequencing, a second file of reads is produced representing reads that start at the opposite end of the mRNA fragment. In this way, a fragment of 500 bases can produce two reads of 100 bases from each end that can be mapped together for increased accuracy.

3.1.3 RNA-Seq Differential Expression Analysis

RNA-Seq is a versatile methodology with many applications, including the quantification of gene expression, discovery of alternative splicing, and identification of non-coding RNAs (Sahraeian et al., 2017). Among these, comparing gene expression profiles to detect genes that are upregulated or downregulated between two or more cells in different conditions can reveal valuable insights into gene regulation and cell responses to various conditions. Specifically, in SCZ research, RNA-Seq has enabled the comparison of genes expression profiles between affected individuals and healthy controls within or across different age groups (Chang, X. et al., 2017; Jaffe, A. E. et al., 2018). The BrainSeq consortium project (BrainSeq: A Human Brain Genomics Consortium, 2015) aimed to collect RNA-Seq samples from brain tissue for a range of neuropsychiatric disorders as well as controls across the whole human lifespan. In Phase I, 746 dorsolateral prefrontal cortex (DLPFC) samples were collected from donors from fetus to adult, including 341 controls and 184 SCZ patients. In Phase II,

with updated RNA-Seq library preparation protocols, this was extended to both DLPFC and hippocampus in 286 SCZ patients and 265 controls (Collado-Torres, L. et al., 2019). In this phase, the authors identified differentially expressed genes (DEGs) between patients and controls that were unique to either the hippocampus (48 DEGs) or the DLPFC (245 DEGs) with little overlap between them. Gene set enrichment analysis implicated some common biological processes between the DEGs from two brain regions, for example in 'myeloid leukocyte activation' and 'regulation of ion transport', despite the lack of overlap between the DEGs themselves. Studying the differences between SCZ patients and controls is important when looking for therapeutic drug targets, or for understanding disease aetiology. However, while this offers a glimpse into how cells cope with SCZ, it does not inherently reveal the underlying genes that are responsible for the onset of SCZ (Porcu et al., 2021). To unravel the genetic basis of SCZ, one targeted strategy would be the exploration of genes that have higher expression within SCZ-affected tissues compared to levels of expression in tissues not affected by core symptoms in SCZ. To support this targeted approach, the GTEx project provides valuable data, acting as a rich resource of RNA-Seq samples from tissues collected from healthy donors.

3.1.4 The Genotype-Tissue Expression (GTEx) project

Gene expression is affected by many factors including cell type, age, gender, and health status. Genome-wide Association Studies (GWAS) conducted in the early 2000s had identified numerous risk loci that are associated with various traits (height, body mass index, bone mineral density) and diseases (cancers, diabetes, colitis) (Visscher et al., 2012). Previous work had established that some genetic loci in the genome influence RNA expression levels (Emilsson et al., 2008). The GTEx project, introduced in Section 2.1.4, was conceived with the aim of enabling the study of eQTLs throughout the genome (GTEx Consortium, 2013), as well as other trait-expression combinations, such as gender and expression (Oliva et al., 2020), and genotype and splicing (GTEx Consortium, 2020). RNA-Seq was performed on bulk tissue samples from each body site. Raw data was released biannually to the National Center for Biotechnology (NCBI) database of Genotypes and Phenotypes (dbGaP). In 2019, GTEx version 8 part 2 was released including 17,382 samples from 948 donors from up to 54 tissues per donor (GTEx Consortium, 2020). GTEx can be considered a large library of multi-tissue RNA-

Seq data from healthy individuals. This made it one of the largest reference transcriptomes available for research into human gene expression, which can be leveraged by researchers to explore normal gene expression in healthy tissues correlating to those in patients affected by diseases, such as schizophrenia.

3.1.5 Software

The success of RNA-Seq has also partly been due to the huge efforts within the research community to develop open-source RNA-Seq analysis software that is freely available to other researchers, and that can run on affordable, high-spec computer hardware. With increased application to RNA-Seq analysis over the last 15 years software has evolved to perform well under diverse experimental conditions. RNA-Seq differential expression analysis fundamentally involves three main steps performed *in silico* (Figure 3.1): alignment, quantification, and differential analysis (Oshlack et al., 2010; Sahraeian et al., 2017; Williams et al., 2017). Different software packages have been developed that address each stage of the analysis, often tackling the steps through different approaches depending on the aims of the study and the nature of the sample data.

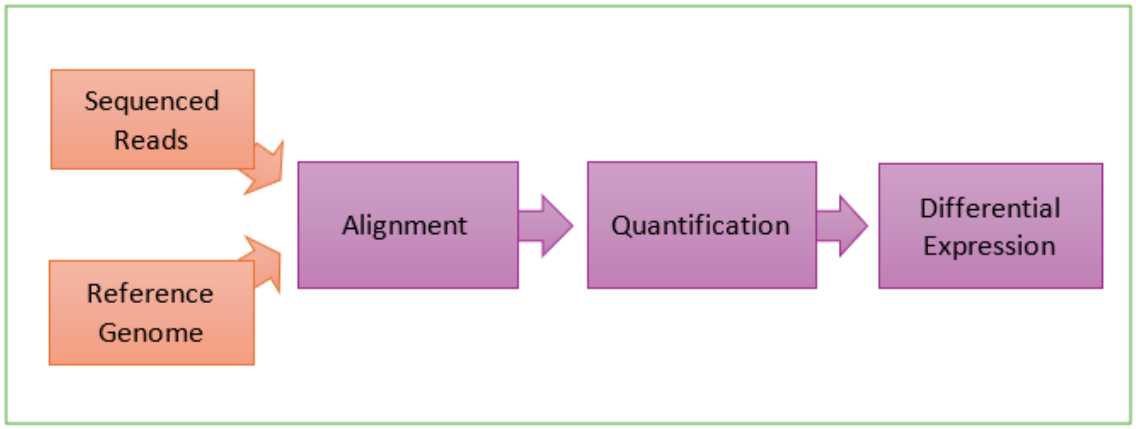

Figure 3.1 Main steps involved in RNA-Seq Differential Expression Analysis

3.1.5.1 Alignment

Alignment is the first stage in analysis of RNA-Seq data. Each paired read in the sample is a fragment of an mRNA that was transcribed from a gene in the genome. Table 3.1 shows a selection of software that take different approaches to this analysis. There

are two main types, those that align reads to the genome (TopHat2, HISAT2, STAR, Rsubread and RUM) and those that align to a constructed transcriptome (STAR, Bowtie2, RUM). Software is said to be 'splice-aware' if it can map reads from exon to exon across introns that might be up to a million bases in length. Before alignment can proceed the Human Reference Genome is first indexed using one of several strategies. TopHat2 and its successor HISAT2 use a Burrows-Wheeler transform (BWT) with a hierarchical full-index in minute space index (FM-index) (Kim et al., 2015). The BWT is a lossless compression algorithm, which compresses the genome sequence and combined with the FM-index enables global alignment for any seed sequence from reads to occur in linear time however large the genome. STAR was developed for the ENCODE Project and uses an index based on suffix arrays and seed searching algorithms (Dobin et al., 2013). For the human genome, a suffix array-based index is very large and requires a lot of memory, but searches are very fast. Aligning to a transcriptome removes the requirement of the aligner to be splice-aware, but it is limited to known transcripts only. It has been suggested that, especially in the brain, there are still many transcripts yet to be discovered (Pertea, 2012; Reble et al., 2018; Yeo et al., 2004). A third type of alignment, pseudoalignment, creates a k-mer hash table from the known transcriptome and has the same limitation. Software that aligns reads to the genome must account for the specific exons in the mRNA transcript (Oshlack et al., 2010; Wang, Z. et al., 2009). Rsubread builds a hash table index of all 16 bp sequences (16 bp k-mers) in the genome and the chromosomal positions where they are located (Liao, Y. et al., 2019). A hash table provides a fast key-value retrieval structure. All 16-mers derived from a sequencing read can be looked up in the hash table and the genomic position with the highest number of matches is taken as the position of the read. The positions with the best matches are then aligned to the reference genome accounting for any variations from the reference genome and exon-exon boundaries.

Table 3.1 RNA-Seq alignment software

Software	Strategy	Reference
TopHat2 HISAT2 STAR	Genome Index-based	(Kim et al., 2013) (Kim et al., 2015, 2019) (Dobin et al., 2013)
Rsubread	Genome K-mer Hash table	(Liao, Y. et al., 2019)
STAR Bowtie2	Transcriptome	(Dobin et al., 2013) (Langmead & Salzberg, 2012)
Sailfish Salmon Kallisto	Pseudoalignment (transcript K-mer Hash table)	(Patro et al., 2014) (Patro et al., 2017) (Pimentel et al., 2017)
RUM	Genome/Transcriptome	(Grant et al., 2011)

Several comprehensive alignment software comparisons and benchmarks have been published (Baruzzo et al., 2017; Sahraeian et al., 2017; Williams et al., 2017) and developers of packages often compare their packages against alternatives. However, the versions compared are rapidly succeeded with updates or new software. For example, Dobin and Gingeras (2015) compared their latest version of STAR (2.1.3) with TopHat2 (2.0.0). However, by 2015 the developers of TopHat2 had released and published HISAT (Kim et al., 2015), their successor to TopHat2. Comparisons between various software tested aligners for their accuracy; how they handled reads that mapped to multiple locations (multimappers); accurate splice junction estimation especially where the end of a read overhangs a junction by a few bases; and their use of a gene annotation file. The two aligners that consistently performed well in comparisons were STAR and HISAT2. In Baruzzo et al. (2017) STAR (2.5.0a/2.5.2a) was top for accuracy, whereas HISAT2 (2.0.0-beta) handled short anchors and canonical and non-canonical splice junctions best. In Sahraeian et al. (2017) HISAT2 (2.0.1-beta) was fastest and most accurate, whilst STAR (2.4.2a) was more sensitive to detecting novel transcripts. In 2019, the developers of Rsubread released a new version that performed favourably in accuracy and speed, but compared it to TopHat2, which had long been superseded by HISAT2, and STAR (Liao, Y. et al., 2019). Using the latest versions of software leverages the newest technologies to gain the maximum

biological insight. With no clear winner, a short exploratory analysis was required to assess the latest versions to maximise the accurate mapping of reads to the reference genome.

RNA-Seq reads may also contain primer and adapter sequences used in the library preparation for sequencing. An additional step to trim off known sequences can be employed using Trimmomatic (Bolger et al., 2014), Cutadapt (Martin, 2011), or built-in trimming if the alignment software has this feature. Trimming can be turned on in STAR using an appropriate switch (Dobin & Gingeras, 2015). To trim or not is a debated topic. Williams, Baccarella, Kim et al. (2016) suggest that people are cautious with trimming. If no minimum length of transcript is specified, then results can be unpredictable. Rsubread's k-mer hash tables and seed-and-vote strategy may mitigate the need to trim datasets (Liao, Yang & Shi, 2020).

All splice-aware aligners can use a gene annotation model as a guide to focus alignment or reads on known gene sequences. Gene annotation models contain gene boundaries and within these, the boundaries of exons, start and stop sites and 5'- and 3'UTR regions. Software that uses the annotation model as a guide, may give preference to aligning reads within genes (rather than in intergenic regions) when multiple genomic sites are returned from the indexed genome. The three most common gene annotation models are the University of California Santa Cruz (UCSC) Known Genes dataset (Hsu et al., 2006), NCBI's RefSeq (O'Leary et al., 2016), and Ensembl/GENCODE (Aken et al., 2017; Frankish et al., 2019). GENCODE was developed for the ENCODE project and has been widely used in large-scale projects (GTEx, the Cancer Genome Atlas (TCGA), Genome Aggregation Database (gnomAD), the 1000 genomes project and the Human Cell Atlas (HCA)). The GTEx v8 pipeline (shared with the TOPMED Project) (Broad Institute, 2020) uses a collapsed GENCODE v26 annotation with the GRCh38 human reference genome (GTEx Consortium, 2020). Ensembl informs and is periodically released in sync with GENCODE releases (Aken et al., 2017). Because they are subject to continuous improvement, the most accurate alignment of RNA-Seq reads and maximum biological insight are achieved by using the most recent annotation model.

3.1.5.2 Read Quantification

Reads that have been aligned to genes are summarised or 'quantified' to produce a metric, for example, "reads per gene" or "reads per exon". Quantification is a pre-processing step to prepare a matrix of counts of mapped reads for all genes in all samples which can then be passed into the differential expression analysis (Van den Berge et al., 2019). Reads per gene is a crude but robust metric that allows for fold changes between samples in two conditions to be accurately calculated. Comparison of gene abundance cannot be made between different genes in a single sample because of the different lengths of genes and their transcripts. featureCounts, part of the Rsubread package (Liao, Y. et al., 2014), can summarise reads to various genomic features: genes, exons, or splice junctions. Both featureCounts and Rsubread run in an R environment and can take advantage of Bioconductor's BiocParallel to leverage multiple processor cores. Other packages for counting reads include htseq-count, from the Python framework, HTSeq (Anders, S. et al., 2015), Rcount (Schmid & Grossniklaus, 2015) and BEDtools (Quinlan & Hall, 2010).

3.1.5.3 Differential expression analysis

Differential Expression analysis is a method that uses a combination of statistical tests to identify genes that are expressed at different levels in two or more groups of samples. The null hypothesis is that the log fold change between the two conditions is exactly zero. There are three main steps to the analysis workflow: normalisation of counts for sample size and gene distribution; the differential expression fitting the data to a data model; and finally, visualisation of the results. Several software packages have been developed to perform these steps with mixed accuracy, handling of false positives, and a variety of different statistical methods (Law et al., 2016; Love et al., 2014; Oshlack et al., 2010; Rapaport et al., 2013; Ritchie et al., 2015). Before read counts can be compared between samples, the counts are normalised for differences in the sample size (size estimation) and how genes are distributed within the samples (dispersion estimation) (Oshlack et al., 2010; Zyprych-Walczak et al., 2015). A sample that contains a total of 10 million reads cannot be compared directly with one that contains 90 million reads. The reads per gene count for each gene in each sample is adjusted (normalised) by a size factor. The reads per gene count is also adjusted by a dispersion estimate. The distribution of counts data generally follows a Poisson

distribution. In actual RNA-Seq data, it has been observed that read counts are over-dispersed; that there is generally more variance between small numbers of samples. As the number of samples increases, the distribution tends towards a Poisson distribution. *edgeR* (Law et al., 2016) and *DESeq2* (Love et al., 2014) both assume that gene counts within a sample follows a negative binomial distribution and estimate a value for dispersion for every gene in every sample. Finally, the adjusted reads per genes counts for samples in each condition are fitted to a design model and tested against the null hypothesis. The log fold change (LFC) between conditions is output along with a standard error (LFCSE), base value, a p value and a q value that has been adjusted for multiple testing. The base value is the normalised reads per gene value in the control condition. Results can be ranked by LFC or by p- or q-value. Positive and negative LFCs indicate higher or lower expression, respectively, in the test condition. LFCs and adjusted p-values can be visualised in a volcano plot to visualise how many genes are differentially expressed.

3.1.6 Aim and Objectives

The aim of this chapter was to identify genes with higher expression levels in either of the two brain regions from GTEx (identified in Chapter 2) - Frontal Cortex and Hippocampus – compared to both control brain regions from GTEx – Cerebellar Hemisphere and Substantia nigra. This analysis was exclusively conducted on tissue samples from healthy donors. Differential gene expression was analysed between brain regions in these individuals correlating to those typically most affected and least affected in patients with SCZ. Unlike conventional methods that focus on gene expression in affected tissues from patients with SCZ or that compare cases to controls, this strategy aimed to uncover potential gene targets by understanding normal gene expression differences between the selected brain regions.

As there was no clear leader in the alignment software packages in benchmarks and comparisons, and because software is continuously being developed, a comparison of alignment software was performed first. STAR 2.5.3a (17/03/2017) with GENCODE v26 had been used in the original GTEx v8.p2 pipeline (Broad Institute, 2020). The updated Rsubread (Liao, Y. et al., 2019) aligner had compared favourably against STAR. The alignment strategies were tested comparing the updated Rsubread (2019) with the latest version of STAR and the latest GENCODE annotations to maximise the alignment

of reads to the reference genome. Since dbGaP access for GTEx had not been granted yet, RNA-Seq datasets from BrainSeq (BrainSeq: A Human Brain Genomics Consortium, 2015) were used in their place for this step only. An assessment was made on the accuracy of alignment to select an alignment strategy for the main experiment. The BrainSeq datasets included DLPFC samples from both patients with SCZ and controls. Their use had no other relevance to the main thesis.

The objectives are:

1. Investigate trimming and alignment strategies using STAR and Rsubread, emphasising maximising accurate mapping of mRNA reads to the Human Reference Genome.
2. Run a differential expression analysis between GTEx Frontal Cortex and GTEx Cerebellar Hemisphere regions only.
3. Run 4 differential expression analyses between the 2 affected and control regions:
    a. Frontal Cortex v Cerebellar Hemisphere
    b. Frontal Cortex v Substantia nigra
    c. Hippocampus v Cerebellar Hemisphere
    d. Hippocampus v Substantia nigra
4. Extract a list of genes with increased expression in either of the 2 affected regions compared to both control regions.
5. Compare the genes against genes associated with schizophrenia from the Open Targets platform.
6. Perform gene enrichment analysis and pathway analysis to identify which molecular functions, pathways and cellular compartments are enriched with the list of genes identified.

3.2 Methods

3.2.1 Study Populations and RNA-Seq Samples

RNA-Seq samples used in this chapter were from two different existing RNA-Seq projects; RNA-Seq samples for (Hippocampus, Frontal Cortex (frozen), Cerebellar Hemisphere (frozen) and Substantia nigra) from healthy donors in the GTEx Project

where all brain regions were available for each donor (GTEx samples) and a cohort of SCZ cases and age-matched healthy donors from the BrainSeq Consortium Phase 1 DLPFC project (BrainSeq samples). All GTEx data was stored and handled in accordance with the NIH Security Best Practices for Controlled-Access Data subject to the NIH Genomic Data Sharing (GDS) Policy and the dbGaP Data User Certification Agreement. The GTEx samples used for the analyses described in this chapter were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v7.p2 or phs000424.v8.p2. BrainSeq samples used for analysis described in this chapter are publicly available and were obtained from Synapse at https://doi.org/10.7303/syn12299750. BrainSeq samples were used to compare alignment methods in place of the GTEx samples while access to dbGaP was granted.

In this chapter, the brain regions refer to the named samples in GTEx rather than the anatomical regions in the brain. "Brain – Cerebellum" and "Brain – Cortex" refer to the PAXgene-preserved samples in GTEx which were sampled and preserved separately to the "Brain - Cerebellar Hemisphere" and "Brain - Frontal Cortex (BA9)" frozen samples. "Brain – " and "(BA9)" are omitted from the GTEx region names for brevity.

3.2.1.1 GTEx Sample – SRR821690 – Hippocampus

The raw RNA-Seq SRA file for a single hippocampus dataset (SRR821690) as an example dataset from GTEx (dbGaP Accession: phs000424.GTEx.v7.p2) was obtained from the Sequence Read Archive (SRA) (Leinonen et al., 2011) using IBM Aspera Fast Adaptive and Secure Protocol (FASP) and sra-toolkit v2.9.6-1 (SRA Toolkit Development Team, n.d.). The SRA file was unpacked and compressed to two paired-end read (FASTQ) files in fastq.qz format using fasterq-dump 2.9.6 from sra-toolkit and parallel gzip (pigz v2.4). This sample dataset provided an early opportunity to assess the properties of GTEx RNA-Seq samples – for example, read length, paired-end characteristics - and enabled the development of a workflow for obtaining further records via dbGaP.

3.2.1.2 BrainSeq – 10 SCZ and 10 controls

The BrainSeq Consortium Phase 1 DLPFC project included DLPFC polyA+ RNA-Seq samples from 738 subjects spanning all ages and 3 diagnostic groups (SCZ, BPD and major depression disorder (MDD)) (BrainSeq: A Human Brain Genomics Consortium, 2015; Jaffe, A. E. et al., 2018). Phenotype data (*phenotype_annotated_szControlEqtl_DLPFC_withDBGAP.csv*) were downloaded from

the Lieber Institute's GitHub repository for Phase 1 (BrainSeq: A Human Brain Genomics Consortium, 2018). BrainSeq only included a single brain region (DLPFC), therefore samples were selected from the SCZ and control groups. Records with a post-mortem interval (PMI) >= 50 hrs were removed. Samples were age- and race-matched, without the use of antipsychotics/antidepressants, where possible. Two samples in each of five age groups (28-33, 36-43, 45-50, 60-67, and over 70) were used. Individuals were assigned a pairing code, for example, F1C = 'female, age group 1, Caucasian' and M3C = 'male, age group 3, Caucasian'. Each pairing code consisted of one SCZ and one control. 10 SCZ and 10 matched control RNA-Seq samples were selected (Chapter3/exp011/data in Slay, 2021) and the FASTQ files were downloaded from Synapse.org (BrainSeq: A Human Brain Genomics Consortium, 2019).

BrainSeq and GTEx library preparations were similar, but the read lengths were longer in BrainSeq samples (100 vs 76 bases).

3.2.1.3 GTEx - 2 Brain Regions - Frontal Cortex and Cerebellar Hemisphere

In a first run of the differential expression analysis only two GTEx brain regions were compared – 'Frontal Cortex' to 'Cerebellar Hemisphere'. 83 Frontal Cortex (frozen) and 83 Cerebellar Hemisphere (frozen) GTEx (dbGaP Accession: phs000424.GTEx.v8.p2) datasets from healthy, matched donors for whom both tissue sites were available, were downloaded using IBM Aspera FASP from the SRA using sra-toolkit v2.9.6-1 (NIH). SRA files were unpacked and compressed to the fastq.qz format using fasterq-dump 2.9.6 and pigz 2.4. All 166 datasets (332 read1 and read2 FASTQ files) were checked for download errors.

3.2.1.4 GTEx - 4 Brain Regions - Frontal Cortex, Cerebellar Hemisphere, Hippocampus and Substantia nigra

In the second and main run of the differential expression analysis, two further regions were added – Hippocampus and Substantia nigra. There were fewer donors from whom all four samples were available. For the main analysis, further filtering of attributes was carried out to select suitable samples. GTEx (dbGaP Accession: phs000424.GTEx.v8.p2) metadata files were downloaded via the GTEx Portal (http://gtexportal.org) on the 23/09/2019. The Sample Attributes metadata dataset contained attributes (atts) for each GTEx sample covering the type of tissue, histological notes, the method of RNA extraction, and whether the dataset was

included in the current data freeze or not: https://storage.googleapis.com/gtex_analysis_v8/annotations/GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt.

The Subject Phenotypes dataset contained phenotype data (pdata) including the Subject ID of each GTEx donor along with their sex, age range, and Hardy Scale (a 4-point death classification scale - see **Appendix 8.1**): https://storage.googleapis.com/gtex_analysis_v8/annotations/GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt.

SRA Run Selector (NCBI, n.d.) is a utility for selecting Sequence Read Run (SRR) read files in the NIH SRA database for downloading. SRA Run Selector for GTEx v8.p2 (https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=phs000424) was filtered to GTEx RNA-Seq paired-end Illumina datasets using the criteria: `DATASTORE.filetype`='sra'; `material_type`='rna:total rna OR <empty>'; `Instrument`='Illumina HiSeq 2000'; and `LibraryLayout`='PAIRED'. The resulting Run List was downloaded as *SraRunTable.txt*.

Further filtering of the runs, attributes and phenotype data was carried out in R. The downloaded full list of RNA-Seq datasets (SraRunTable.txt) was loaded in R. Samples were filtered by body site (`body_site`) to the four GTEx brain regions ("Brain – Cerebellar Hemisphere", "Brain – Substantia nigra", "Brain – Frontal Cortex (BA9)" and "Brain – Hippocampus"). The remaining records were joined by sample id (`Sample.Name=SAMPID`) to the Sample Attributes dataset (atts) and records were filtered to only those included in the GTEx v8.p2 data freeze (`SMAFRZE='RNASEQ'`). The records were joined by Subject ID to the Subject Phenotypes dataset (pdata) to access sex, age, and Hardy Scale data for each subject (`SUBJID = submitted_subject_id`). Finally, the dataset was filtered to only those donors for whom all four brain regions were available.

There were 25 donors for which all four regions were available. 100 SRA files (25 CO; 25 CE; 25 SN; 25 HP listed in **Appendix 8.2**) from 25 donors were downloaded using IBM Aspera FASP from the SRA using sra-toolkit v2.9.6-1 (SRA Toolkit Development Team, n.d.). SRA files were unpacked and compressed to the fastq.qz format using

fasterq-dump 2.9.6 and pigz 2.4. The 200 datasets (100 read1 and 100 read2 FASTQ files) were checked for download and conversion errors.

Note that from this point forward the abbreviations CO, CE, HP, and SN represented the RNA-Seq samples from each of the four GTEx Brain Regions, rather than the anatomical brain region from which they were sampled.

3.2.1.5 Human Reference Genome and Genome Annotation

All datasets were mapped to the Human Reference Genome GRCh38.p12 (hereafter GRCh38) obtained from GENCODE.

During the exploratory stages, gene annotation data from GENCODE v30 (Apr 2019) was used. GENCODE v31 (Jun 2019) was adopted when available to maximise access to the most up-to-date gene model data. The main analysis and all subsequent downstream analyses in this chapter used GENCODE v31 and corresponding data from Ensembl Release 97.

3.2.2 GTEx – Principal Components Analysis for all brain regions

The four GTEx regions selected were first examined in a principal components analysis (PCA) against all 13 available GTEx brain regions and all 2642 samples. GTEx v8 gene read counts for all tissues were downloaded via the Bulk Tissues v8 section of the Open Access Datasets page on GTExPortal.org:

https://storage.googleapis.com/adult-gtex/bulk-gex/v8/rna-seq/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz

A table of v8 Sample Attributes were downloaded from the Metadata section:

https://storage.googleapis.com/adult-gtex/annotations/v8/metadata-files/GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt

The table of Samples was filtered to Brain (`SMTS='Brain'`) and only to those included in v8 (`SMAFRZE='RNASEQ'`). The Sample names remaining were used to filter the samples in the gene read counts matrix. Duplicated gene were removed if they had row counts less than 10. This removed all genes labelled as 'Y_PAR'. Finally the order of samples in the counts matrix was set to the order of sample names in the Samples table and checks were made that all samples matched between the tables.

A DESeqDataSet (`dds`) was created using *DESeqDataSetFromMatrix* in DESeq2 (Love et al., 2014) in R using the gene read counts, sample table, and a design model based on the GTEx Brain Region: "~ *SMTSB*". Genes were filtered all those with more than 10 reads in at least 139 samples (the number of samples in the smallest group). DESeq was run on *dds* followed by the variant stabilising transformation, *vst*, from the DESeq2 package. A PCA graph was generated using *plotPCA* from DESeq2, which analysed PC1 and PC2 from the variance-stabilised read counts for the top 1000 genes by highest row variance across all samples. 95% confidence regions using a normal distribution were added as ellipses for all samples in a collective brain region. Samples were labelled with a combined brain region to simplify the confidence regions where samples clustered together. 'Cerebellum' combined samples from GTEx regions 'Cerebellum' and 'Cerebellar Hemisphere'; 'Cortical' combined all cortex regions (including 'Anterior Cingulate Cortex'); and 'Striatum' combined 'Putamen', 'Caudate nucleus' and 'Nucleus accumbens'. Cortex and Hippocampus samples and confidence regions were given blue hues, and the control brain regions – Cerebellum and Substantia nigra – were given red hues.

3.2.3 Quality Control - Analysis of single dataset, SRR821690 - Hippocampus

The GTEx Hippocampus sample (see 3.2.1.1) was studied for quality control using FastQC v0.11.8 (Andrews, 2015). A reference genome index was built from the reference genome, GRCh38, using the command `genomeGenerate` from STAR 2.7.1a (Dobin et al., 2013) and the following options: sjdbOverhang = 76, runThreadN = 8; annotation = GENCODE v30 patched contigs. A test alignment to this index with STAR 2.7.1a (Dobin et al., 2013) revealed that the reads in the SRR821690 FASTQ files generated from the SRA file were sorted with previously mapped reads at the start of the files and more fragmented and unmapped reads at the end. The reads were analysed for adapter sequences and unmapped sequences to inform the optimal alignment strategy. The FASTQ files were divided into 3 parts (40 million reads, 40 million reads and 14,446,403 reads), which were checked for quality using FastQC separately. The dataset was analysed using custom Perl scripts for mitochondrial sequences and polyA-tail sequences (Chapter3/exp010 in Slay, 2021).

3.2.4 Compare Aligner Software – STAR and Rsubread - Analysis of 20 BrainSeq samples

Software for RNA-Seq read trimming and alignment to the reference genome were assessed using 10 SCZ and 10 age- and sex-matched control BrainSeq samples (see 3.2.1.2). All 40 paired-end FASTQ read files were quality checked with FastQC (v0.11.5). The reports were combined and visualised using MultiQC (v1.7) (Ewels et al., 2016). Based on these reports, trimming in each of the workflows was set to trim TruSeq RNA-Seq adapter sequences without their ID sequences from Read 1 reads and to trim Illumina PCR Primer sequences from Read 2 reads. Trimmed reads were aligned to the indexed GRCh38 reference human genome using STAR v2.7.1a 15/05/2019 (Dobin et al., 2013) or Rsubread v1.34.7 (Liao, Y. et al., 2019) with GENCODE v31 primary assembly gene annotation. Figure 3.2 shows a summary of the strategies assessed. Script paths in the steps refer to source code in PhD-codebase/Chapter3-Differential Expression Analysis/exp011 (Slay, 2021). [Scripts were re-used and may not be sequential within each Alignment.]

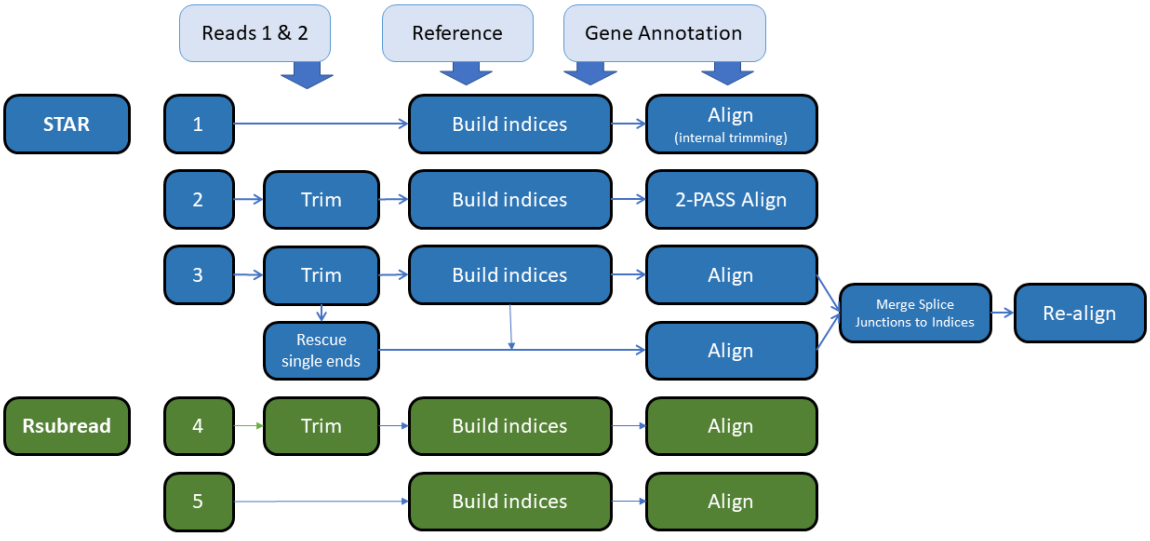

Figure 3.2 Summary of alignment strategies tested on BrainSeq samples

Alignment 1 – STAR + Self-trimming:

1. Build STAR indices (sjdbOverhang = 99, reference = GRCh38, annotation = GENCODE v31) [exp011/pipeline-03a.index.sh].

2. Align paired end reads to the index - Trim ends for adapters and 100 As or 100 Ts (`--clip3pAdapterSeq`) – this occasionally trimmed one whole read from a pair, so a minimum read length was set to 20 bases (`-m 20`). Reads in a pair in which one read fell below this minimum length were discarded, because STAR requires two reads.

Alignment 2 – trimming with CutAdapt and STAR 2-PASS alignment:

1. Trim ends for adapters and 100 As or 100 Ts using CutAdapt [exp011/pipeline-01.trim.sh].
2. Align to the index built in Alignment 1 using STAR 2-PASS alignment [pipeline-03b.align.sh with `--twopassMode=Basic`].

Alignment 3 – trimming with single end rescue and STAR 1-PASS, merge splice junctions and re-align:

1. Step 1 from Alignment 2.
2. Align to the index using STAR alignment [pipeline-03b.align.sh]
3. Recover single reads – where one read in a pair has been discarded due to adapter or polyA removal [exp011/pipeline-02.trim.sh].
4. Align recovered, single-ended reads to the index [exp011/pipeline-04.align.se.sh]
5. Move all *ReadsPerGene.out.tab* and *SJ.out.tab* files and prefix with <sample_id><read> [exp011/pipeline-05-renametables.sh]
6. Filter all *SJ.out.tab* files – exclude any junction that is not supported by at least 5 reads – combine results into a single *combined.SJ.out.tab* file.
7. Generate a new STAR index (sjdbOverhang=99, ref=GRCh38, anno=GENCODE v31) with *–sjdbFileChrStartEnd* set to the combined *combined.SJ.out.tab*. The new index is *gen31sj*. [exp011/pipeline-06.insertSJ.sh]
8. Align trimmed paired end and rescued single end reads to *gen31sj* [exp011/pipeline-03b.align.sh and pipeline-04.align.se.sh with `--genomeDir gen31sj`]
9. Sort and merge all unique alignments from trimmed paired and rescued single ends [exp011/pipeline-10.sortindextrimbam15.sh]

10. Merge all *ReadsPerGene.out.tab* where the reads value for the *stranded* column was greater than 0. [exp011/pipeline-08.nonzero_rpgs.sh]

Alignment 4 – trim reads and align with Rsubread:

1. Trim reads as in step 1 of Alignment 2.
2. Build indices with *buildindex* from Rsubread. (All settings were default except `TH_subread` threshold. The threshold setting for highly repetitive reads was increased from the default 100 to 125.)
3. Align reads to index using subjunc from Rsubread.

Alignment 5 – align with Rsubread (without trimming):

1. Align raw reads to the index from Alignment 4 with Rsubread.

3.2.5 GTEx Frontal Cortex vs Cerebellar Hemisphere – Differential Expression Analysis

The 332 datasets (83 read1 and 83 read2 FASTQ files per brain region) from the GTEx samples (see 3.2.1.3) for the two regions, Frontal Cortex and Cerebellar Hemisphere, were quality checked with FastQC (v0.11.5). The human reference genome, GRCh38, was indexed using Rsubread's *buildindex* with all default settings except the threshold for including highly repetitive subreads that appear in more than one position (`TH_subread=125`) in R. Using *subjunc* from Rsubread v1.34.7 (Liao, Y. et al., 2019) in R v3.6.1 (2019-07-05), all datasets were aligned to the GRCh38 index using GENCODE v31 gene annotation as a guide. Reads per gene were counted using *featureCounts* from Rsubread v1.34.7 and compiled into a table, *ReadsPerGene.out*. PCA analysis was carried out using *ReadsPerGene.out* to examine patterns in the Frontal Cortex and Cerebellar Hemisphere samples.

DESeq2 (Love et al., 2014) performs differential expression on counts of reads aligned to each gene in each sample quantified after alignment (as a *ReadsPerGene* counts matrix). First, DESeq2 normalises the counts for each gene by a scaling factor related to library size to ensure that count sizes can be compared between samples. Next, DESeq2 models the mean and dispersion for each gene independently across all samples based on a negative binomial distribution. Finally, DESeq2 runs a Wald test to compare the distribution of normalised gene counts in each condition according to the

supplied model. This calculates how the fold change of mean counts per gene in each condition differs from the expected variability based on the estimated dispersion for that gene.

Differential expression analysis of read counts per gene was run using DESeq2 to examine which genes are differentially expressed (DE) between the Frontal Cortex and Cerebellar Hemisphere datasets. A paired sample model (*~sex + roi*) was supplied to DESeq2. This tests the brain region of interest (roi) with gender as a covariate. In this experiment, one Frontal Cortex sample and one Cerebellar Hemisphere sample were used from each of the 83 donors. This model aimed to capture the difference in gene expression due to the region of interest (roi), while controlling for gender.

The normalised counts from DESeq2 were transformed to control the variance in the data using the variance stabilising transformation, *vst,* from DESeq2. A PCA was performed on these counts and PC1 and PC2 were plotted using *plotPCA* in R.

3.2.6 Main Differential Expression Analysis - Cortex/Hippocampus vs Cerebellar Hemisphere/Substantia Nigra

In this main analysis, the four brain regions were processed collectively by DESeq2. This enabled the correct estimation of dispersion for genes across all brain regions that were subsequently used in the assessment of log fold changes between the different regions, and their significance levels. The four contrasts were extracted from the final DESeq2 result.

3.2.6.1 Quality Control and Read Alignment

All FASTQ files from the 100 datasets (100 read1 and 100 read2 FASTQ files) were checked for errors with FastQC 0.11.8 (Andrews, 2015). The human reference genome, GRCh38, was indexed using Rsubread's *buildindex* in R v3.6.1 (2019-07-05) with all default settings except the threshold for including highly repetitive subreads (16 kmers) that appear in more than one position (`TH_subread=125`). Using *subjunc* from Rsubread v1.34.7 (Liao, Y. et al., 2019) in R, all 100 datasets were aligned to the GRCh38 index using GENCODE v31 gene annotation (Frankish et al., 2019) as a guide (Chapter3/exp012/01_rubread in Slay, 2021). An option was enabled to create sorted and indexed output BAM files (`sortReadsByCoordinates=TRUE`). The

maximum number of mismatches was set to 4 (`maxMismatches=4`). Reads mapping to multiple locations ('multimappers') were reduced to 2 (`nBestLocation=2`).

The total number of reads covering exons in each gene was counted per gene per sample using the *featureCounts* utility from Rsubread v1.34.7. The ReadsPerGene counts were extracted from the BAM files using exons as the feature type (`GTF.featureType='exon'`) from the GENCODE v31 annotation file, summed over each gene (`GTF.attrType='gene_id'` `GTF.attrType.extra='gene_name'`). Mulitmapped reads were only counted once for their primary alignment (`primaryOnly=TRUE`). This produced a *ReadsPerGene* counts matrix consisting of all genes as rows and all 100 samples as columns for the differential expression analysis.

3.2.6.2 Differential Expression Analysis

The *ReadsPerGene* counts matrix for the aligned samples was analysed for differential expression using a paired sample design. This analysed for fold change between regions of interest (roi) using the subject ID as a covariate. (Sex was not used as a covariate because it is inherent in the subject ID.)

Phenotype data (see **Appendix 8.2**) was extracted from the Sample Attributes table (GTEx V8p2) for the samples included in the study. The DESeq2 (1.24.0) (Love et al., 2014) package was used in R for gene-level differential analysis. The *DESeqDataSetFromMatrix* function was run to generate a DESeqDataset from the *ReadsPerGene* counts matrix with a design model: "~ *subject_id* + *roi*". The *DESeq* wrapper function normalises for sample size and dispersion and runs the negative binomial Wald Test, *nbinomWald*, function to test for differential expression between samples in each brain region. Sets of comparisons between brain regions were extracted from the output DESeqDataset (alpha=0.005; LFC>0.6) for the respective *roi* contrasts (CO v CE, CO v SN, HP v CE and HP vs SN). The results for CO v CE and CO v SN were intersected to find genes highly expressed in the cortex compared to both control regions (CO-genes). The results for HP v CE and HP v SN were intersected to find genes highly expressed in the hippocampus compared to the control regions (HP-genes). Genes in CO-genes and HP-genes were annotated with 1 for cortex, 2 for hippocampus, and 3 where the gene appeared in both lists (CO/HP-genes).

3.2.6.3 biomaRt annotation of genes in three gene sets

The three lists of genes (CO-genes, HP-genes, and CO/HP-genes) were annotated using the Bioconductor package, biomaRt (v2.44.4; Durinck et al., 2009) in R and the July 2019 archive version of Ensembl 97 (Cunningham et al., 2022) for *Homo sapiens* (`host="jul2019.archive.ensembl.org"; dataset="hsapiens_gene_ensembl"`). Ensembl 97 matches with gene IDs in GENCODE v31 (see 3.2.1.5) that were used in the previous read alignment and read counting steps. Genes were annotated with external gene name, description, and biotype. The genes were filtered to protein-coding genes only. The proportion of lncRNA, protein coding and pseudogenes in each contrast/comparison was plotted in a bar chart to show the relative proportions of genes of each of these biotypes.

3.2.6.4 Volcano plots of differentially expressed genes in contrasts

Volcano plots were generated using the package EnhancedVolcano (v1.6.0; Blighe et al., 2019) in R to visualise each of the 4 contrasts in the differential expression analysis [Chapter3/exp012/visualisation/volcano2.R (Slay, 2021)]. For each contrast, the DESeq2 *results* function was used to extract differentially expressed genes greater than an absolute LFC of 0.6. All data points were coloured grey. Data points above an LFC of +0.6 with an adjusted p value < 0.005 were coloured red. This produced a one-sided volcano plot to show genes with increased expression levels in samples from each brain region (CO or HP) compared to the respective control region. Genes in the combined CO/HP-genes list were coloured green in all plots. The top 5 genes in each contrast were labelled with the external gene name retrieved from biomaRt.

3.2.6.5 Venn Diagram of the four contrasts

All gene names from the four contrasts (CO-SN, CO-CE, HP-SN, HP-CE) were combined into a single column of a new table, and duplicates removed. Four columns were added for each of the contrasts. Genes found in each contrast were given a value of 1 in the corresponding contrast column. This data table was then visualised as a 4-set Venn diagram using the VennDiagram (Chen, H. & Boutros, 2011, 2018) package in R. This showed whether highly expressed genes are mostly expressed in the HP or CO or both regions compared to the CE or SN or both control regions. Genes common to all regions (CO/HP-genes) are those highly expressed in both the CO and HP compared to both control regions, CE and SN.

3.2.6.6 Plotting mean expression of CO/HP-genes on GTEx Brain Regions

Mean expression of the differentially expressed genes was calculated and overlaid on the PCA of all GTEx brain regions. This visualised the distribution of the differentially expressed genes across all samples across all of the brain regions pointing to regions where these genes are more actively expressed.

Mean expression was calculated per sample for all protein-coding CO/HP-genes for all GTEx brain samples from 3.2.2. Samples from the same PCA plot were coloured using blue for lower expression and red for higher expression.

3.2.6.7 Expression Weighted Celltype Enrichment in single-cell cortex and hippocampus

The CO-genes and HP-genes were separately analysed for enrichment in single-cell mouse (somatosensory) cortex and hippocampus (Zeisel et al., 2015) from *ewceData* in the package EWCE v 1.10.2 (Skene & Grant, 2016). The function bootstrap_enrichment_test was run with source cell type species 'mouse' which maps the human genes in the CO- and HP-genes lists to their mouse ortholog (`sctSpecies = 'mouse genelistSpecies = 'human'`). For publication, 10,000 repetitions were required (`reps = 10000`). An annotation level of 2 was used to test for enrichment in all cell subtypes in the dataset (Zeisel et al., 2015). Results from each gene list were merged and plotted together using *ewce_plot* to show significantly enriched single-cell subtypes, applying multiple testing correction using the Benjamini-Hochberg method.

3.2.6.8 Comparison of Genes with other SCZ-associated Genes from the Open Targets Platform

The analysis aimed to indicate the number of genes in each comparison that are enriched for genes associated (strongly or weakly) with SCZ in the Open Targets platform (Ochoa et al., 2021). A list of genes associated with SCZ and their 'overall association scores' for SCZ (EFO_0000692) was downloaded from the Open Targets. These SCZ-associated genes were divided into two sets: overall association score (OAS) below 0.75 and overall association score above or equal to 0.75. Two additional columns were added for each contrast in the contrast table from 3.2.6.5. For each gene in the SCZ-associated risk genes, if the gene was found in a contrast, then true

was set in the appropriate column, OAS>0.75 or OAS<0.75. Two further 4-set Venn diagrams were generated for Open Target overall association scores above or below 0.75. This showed how many genes in the Open Targets SCZ-associated genes were found in each of the subset comparisons for the 4 contrasts.

3.2.6.9 GO Pathway Analysis and Functional Annotation Clustering

To verify that the highly expressed protein-coding genes in the two most affected regions are specific to biological processes expected in these brain regions, gene set enrichment analysis was carried out using Ma'ayan Lab's *enrichR* package (Kuleshov et al., 2016; Xie et al., 2021) in R. Gene symbols ('external gene name') for all CO-genes and all HP-genes were run separately against the following databases for biological processes, molecular function, and the cellular component: "GO_Biological_Process_2021", "GO_Molecular_Function_2021", "GO_Cellular_Component_2021". The two gene sets were also run against "KEGG_2021_Human", and "Reactome_2022" in *enrichR* to identify pathways in which the genes are enriched. All significant results that survived multiple testing (adjusted P value < 0.05) were retained. The terms from each brain region were combined and sorted by Enrichr's 'combined score' in descending order. This score combined the Fisher's exact test and odds ratio for a specific term (Kuleshov et al., 2016).

For the Functional Annotation Clustering, the CO-genes and HP-genes were input as gene lists to DAVID Bioinformatics (Huang, D. W. et al., 2009; Sherman et al., 2022) online (https://davidbioinformatics.nih.gov/). A background of Homo sapiens was automatically detected. Functional Annotation Clustering was performed with semantic similarity set to a threshold of 5. Databases selected were GO Biological Process, GO Cellular Component, GO Molecular Function, KEGG, Reactome, and Wikipathways. Each cluster was given an appropriate representative name that encompassed the clustered terms. These representative terms were presented along with the cluster's overall enrichment score – the negative log of the geometric mean of the clustered terms' p values – and the number of genes in the cluster.

3.3 Results

The outcomes of the exploratory analyses and main analyses focused on gene expression in four distinct brain regions are presented in this section. Two key subsections have been organized: "Exploratory Methodological Findings" and "Main Comparative Analysis of Four Brain Regions". Insights into preliminary assessments and methodological choices, including alignment strategy comparisons, are provided in the former. The latter delves into the primary objective of this chapter—comparing gene expression across the identified brain regions.

3.3.1 Exploratory Methodological Findings

This section presents a PCA analysis showing the landscape of gene expression in the selected GTEx brain regions against all other GTEx brain regions, and three key analyses aimed at refining the methodology for gene expression analysis.

3.3.1.1 Principal Component Analysis (PCA) of all GTEx Brain Regions

A PCA was performed (Figure 3.3) on all brain regions to explore the four selected brain regions in the context of the overall gene expression in GTEx brain samples. Each data point represents a sample from different brain regions.

PC1 explained 41% of the variance in the expression counts and PC2 18% of the variance. All cerebellum samples (PAXgene Cerebellum (N=241) and frozen Cerebellar Hemisphere (N=215)) were separated from the rest of the brain samples in PC1. The cerebellum samples were also tightly clustered around the centre of PC2, indicating a high degree of similarity and consistency in their gene expression profiles. This is contributed to by the distinct functions of the cerebellum and relatively homogeneous cellular composition, dominated by granule and Purkinje cells. 'Spinal Cord' (N=159) and 'Substantia nigra' (N=139) showed some degree of separation in PC1 from both the cerebellum samples and the rest of the brain regions. The rest of the brain regions were spread across PC2 highlighting the variability of gene expression reflecting diverse functions and cell types among these regions. There was some overlap between cerebellum samples and some hippocampus, cortex and amygdala samples along PC2, indicating that despite the distinct roles of the cerebellum there was a subset of genes with similar expression levels or variance between these regions.

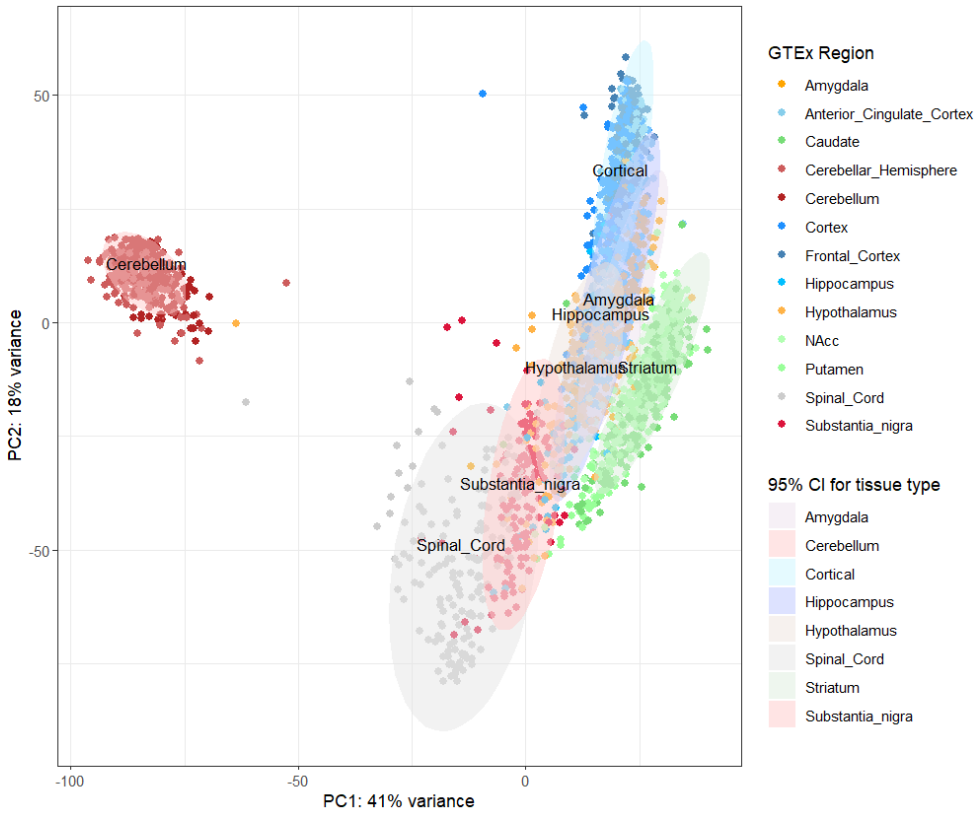

Figure 3.3 Principal Components Analysis of all GTEx Brain Regions
A PCA was performed using plotPCA from DESeq2 using gene read counts from 2642 GTEx brain samples. The PCA was based on the top 1000 genes with the highest row variance across all samples. 95% confidence intervals (CI) as ellipses were added assuming normal distribution (Cerebellum, Cortical and Striatum regions each encapsulate more than one GTEx brain region). Blue hues were used for samples in the target Cortex and Hippocampus regions; red hues were used for control regions, Cerebellum and Substantia nigra.

3.3.1.2 Analysis of Sample GTEx Dataset: Quality Control and Trimming Strategies

Quality control measures using FastQC, including an examination of 'per base sequence quality' for the GTEx dataset after division into two parts each of 40 million reads and a third part of 14.4 million reads (shown in Figure 3.4) revealed degradation of reads in the last 14.4 million reads. The graphs show boxplots of base qualities for each position of the GTEx read. In the first two parts, mean quality across the total 80 million reads remains high with some minimal degradation towards the ends of the paired reads (READ 2). In the final 14.4 million the average per base quality drops in READ_1 and is very poor in READ_2. This degradation necessitated a trimming strategy to remove adapter and primer content to maximise the retrieval of the data for

alignment. Five alignment strategies (three using STAR and two using Rsubread) were devised to address these issues and optimize alignment accuracy.

Figure 3.4 Quality scores across bases in each read pair for SRR821690
Per base sequence quality plots (Sanger/Illumina 1.9 encoding) from FastQC applied to SRR821690 divided into three parts (40 million reads, 40 million reads and 14.4 million reads). Left, READ_1; right, READ_2. Read quality is degraded in the last 14.4 million reads.

3.3.1.3 Rsubread, without trimming, uniquely aligns more reads than STAR

The RNA-Seq aligners, Rsubread and STAR, underwent a detailed comparison on 20 BrainSeq samples, using 5 strategies considering both basic and comprehensive trimming scenarios.

Focussing on the number of uniquely mapped reads, STAR's self-trimming results closely approximated those achieved through pre-trimming. Only a small advantage was gained through a comprehensive trimming process involving STAR (2-PASS) alignment, merging, re-indexing with splice junctions, and a subsequent new alignment. However, a larger improvement was observed in the best Rsubread alignment, which aligned between 2% and 11% more reads per dataset than the best STAR alignment (see Figure 3.5). In concrete terms, this equates to over 10 million additional reads in one dataset – a substantial impact on the "reads per gene" metric and the overall data retrieved from the dataset. Remarkably, this increase was achieved without the need for trimming, rendering the Rsubread approach considerably faster in execution.

As a result of these findings, the decision was made to utilize Rsubread for all further analyses.

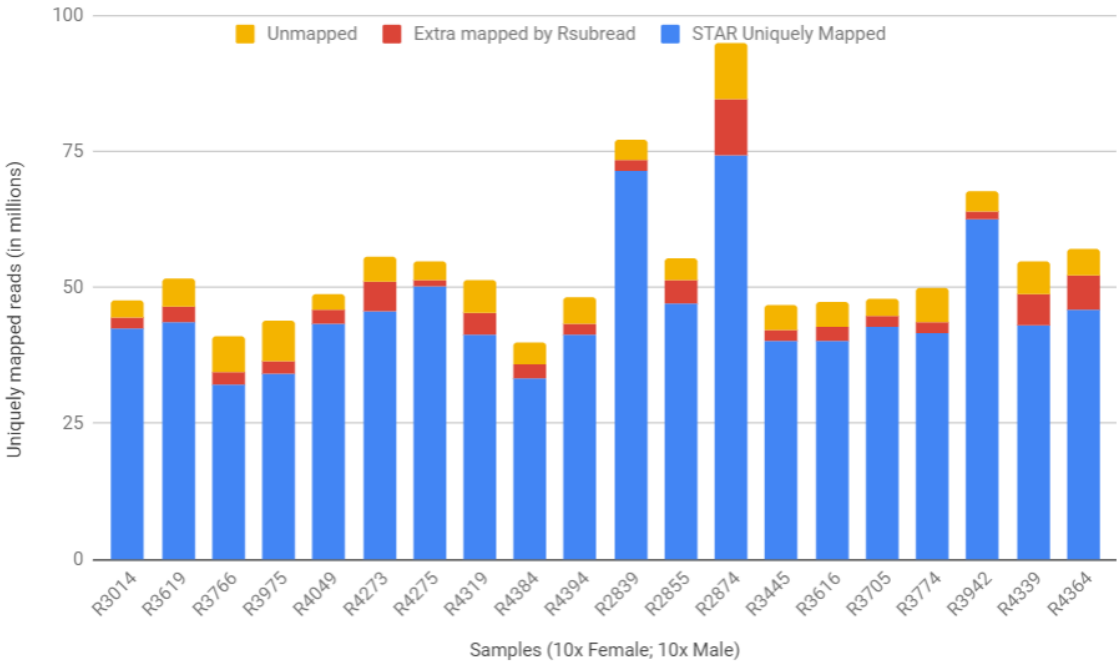

Figure 3.5 Mapping rates for all BrainSeq samples using STAR and Rsubread
Bar chart showing the number of uniquely mapped reads (in millions) for each BrainSeq dataset. The best STAR alignment of the three STAR trimming and alignment strategies is plotted in blue. The best Rsubread alignment mapped 2-11% more reads per dataset (red). Unmapped reads per dataset are shown in yellow.

3.3.1.4 GTEx Frontal Cortex vs Cerebellar Hemisphere - Model Development

83 GTEx Cortex and 83 Cerebellar Hemisphere datasets were aligned using Rsubread and a differential expression analysis was run. 15,690 (coding and non-coding) genes (of a total 60,992) were significantly differentially expressed (either up or down) and surviving adjustment for multiple testing (adj. $P < 0.005$). The first PCA analysis revealed a single donor's Cerebellar Hemisphere sample was plotted directly in the centre of the cluster of Frontal Cortex samples along with its own Cerebellar Hemisphere sample. It was deduced that the donor (GTEX-N7MT) was a mislabelled Cerebellar Hemisphere sample and was removed from the experiment along with the donor's Cerebellar Hemisphere sample. 82 donors (164 Cortex and Cerebellar samples) remained in the analysis. The differential expression analysis and PCA were re-run without the donor. Figure 3.6 shows a second PCA analysis with the datasets removed demonstrating very clear clustering by PC1 (87% variance explained by the brain region of interest) and PC2 (3% variance explained by gender).

The PCA analysis step was important in identifying a mislabelled dataset, but this two-region analysis led to the categorization of genes into two large groups. It was not sufficient for more precise identification of genes highly expressed in the Cortex. The addition of further contrasts between the Frontal Cortex and other control regions would refine this further. Therefore, it was deduced that in the main analysis (3.3.2) encompassing four brain regions with both the Cerebellar Hemisphere and Substantia nigra acting as control regions offered a more refined approach that can contribute to a more accurate identification of region-specific gene expression patterns within the two more affected brain regions, Hippocampus and Frontal Cortex.

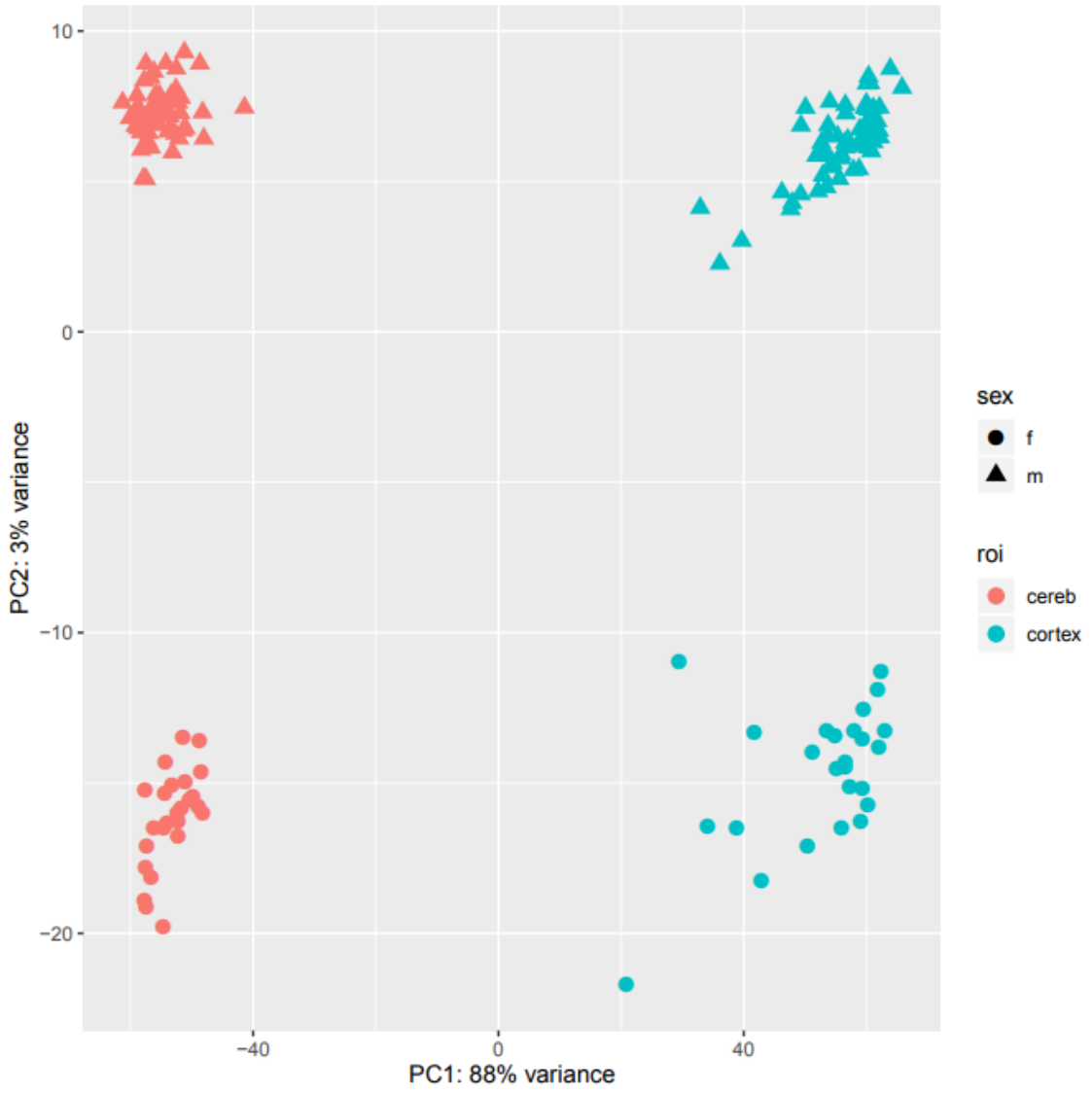

Figure 3.6 PCA for 162 Frontal Cortex and Cerebellar Hemisphere samples
Reads per gene for each of 82 Frontal Cortex and 82 Cerebellar Hemisphere samples were normalised and a principal components analysis was performed between region of interest with sex as a covariate. 88% of the variance in the data can be accounted for by the region of interest, while only 3% of the variance is accounted for by sex.

3.3.2 Main Comparative Analysis of Four Brain Regions – Cortex and Hippocampus vs Cerebellar Hemisphere and Substantia Nigra

In the main analysis the reads for all 100 samples (4 brain regions from 25 donors) were aligned to the GRCh38 reference genome using GENCODE v31 gene annotation. Reads were quantified per gene. A PCA analysis was run to quality check the samples within the four GTEx regions – Frontal Cortex (CO), Hippocampus (HP), Cerebellar Hemisphere (CE) and Substantia nigra (SN). Four differential expression analyses were run as the

four contrasts – HP to CE; HP to SN; CO to CE; and CO to SN. The four contrasts were then combined to identify highly expressed genes in the CO and HP regions.

From 60,992 coding and non-coding genes (in GENCODE v31), this differential analysis of reads per gene found 943 protein-coding genes with significantly increased (>1.5x) expression between CO and both CE/SN and 536 protein-coding genes increased between HP and CE/SN (adj. $P < 0.005$). There were 403 protein-coding genes with increased expression in both CO and HP compared to both CE and SN.

3.3.2.1 PCA Analysis

The exploratory analysis (3.3.1.4) highlighted the importance of the PCA analysis to check that individual samples were correctly labelled. Figure 3.7 shows the PCA analysis of the reads per gene for all four brain regions. PC1 explained 53% of the variance in the expression profiles in the 100 datasets. Colouring each sample by the GTEx brain region identified that this variance explained cerebellar versus non-cerebellar expression. PC2 explained 19% of the variance in the 100 datasets. The Cerebellar Hemisphere samples were tightly clustered between -8 and -16. Substantia nigra samples showed a more spread-out cluster at the top with Hippocampus and then Frontal Cortex samples extending along PC2 to the bottom of the graph. Neurons in the Substantia nigra pars compacta are mostly dopaminergic neurons. Neurons in the Cerebellar Hemispheres are predominantly GABAergic, while neurons in the DLPFC and hippocampus are glutamatergic and GABAergic. PC2 may be explained by combined expression profiles for production of different neurotransmitters in cells from the bulk tissues represented by these RNA-Seq samples. (While techniques to deconvolute bulk RNA-Seq samples are advancing, it was beyond the scope of this thesis.)

The CO sample within the HP cluster and vice versa (samples SRR1487685 and SRR598768) were not considered as outliers and were kept in the study. They were both in the SCZ-relevant brain regions and were at a considerable geometric distance from the control region clusters. Cross-over in expression profiles between some CO and HP samples was to be expected because the hippocampus is considered an older cortical region.

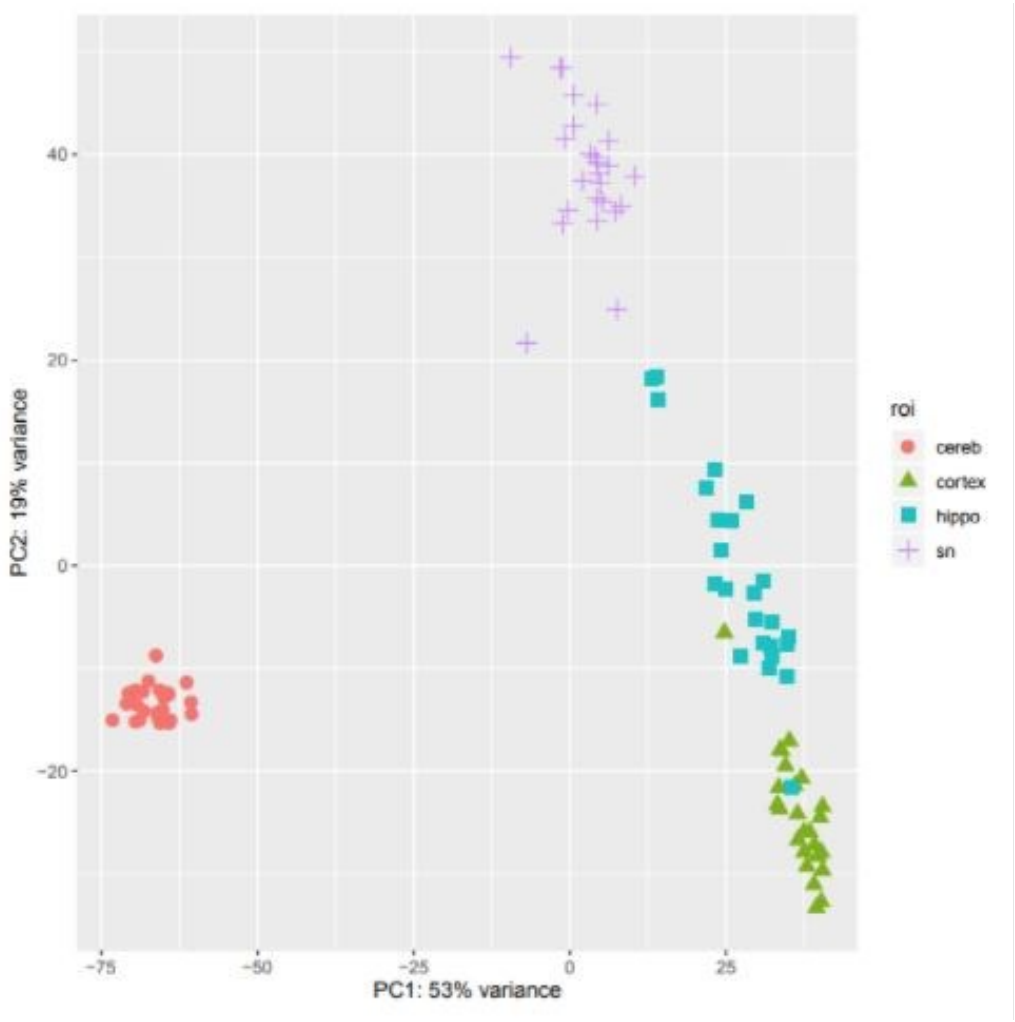

Figure 3.7 PCA plot for all 4 brain regions
PC1 explains 53% of the variance in the samples and separates cerebellar samples from all other brain regions. PC2 explains 19% of the variance in the samples and separates out the non-cerebellar samples into three continuous clusters Substantia nigra samples at the top, then a spread of the two cortical regions (CO and HP).

3.3.3 Genes with higher expression in the Frontal Cortex compared to control region samples

The numbers of genes with higher expression in CO samples compared to control region samples was summarised in Table 3.2 and plotted in the two volcano plots in Figure 3.8. In the comparison of CO to CE, of 56,391 genes with a non-zero total read count, 6,147 genes (11%) had greater than 1.5x higher read counts in the CO samples compared to CE samples within each donor (LFC > 0.6; adjusted p-value < 0.005). 16,334 genes (29%) were removed due to low counts. In the contrast CO to SN, 3,585 genes (6.4%) had a greater than 1.5x higher expression in the CO samples compared to SN samples within each donor (adjusted p-value < 0.005). 16,334 genes (29%) were

excluded due to low counts. (These were the same 16,334 genes excluded in the CO-CE comparison, because all counts were all processed in the same DESeq2 experiment.) Log_2 fold change (LFC) is plotted against -$Log_{10}P$-value in the volcano plots in Figure 3.8 for CO-SN (top) and CO-CE (bottom). Genes that showed higher expression against only one control region are coloured red in each plot.

Table 3.2 Differentially expressed genes for each contrast
For each differential expression test, the number of genes with >1.5x expression is shown. Target region specific genes were the number of genes common to both tests against control regions for each target region, Cortex and Hippocampus. These consisted of both protein-coding and non-coding genes. There were a total of 1076 unique protein-coding genes in both CO- and HP-genes, with 403 in common.

Target Region	Differential Expression Test	>1.5x reads/gene	Target Region specific	Protein coding / non-coding	CO/HP-genes (protein-coding)
CORTEX	v CE	6,147 (11%)	1,730	943	Total: 1076
	v SN	3,585 (6.4%)		787	
HIPPOCAMPUS	v CE	8,045 (14%)	882	536	In common: 403
	v SN	1,542 (2.7%)		346	

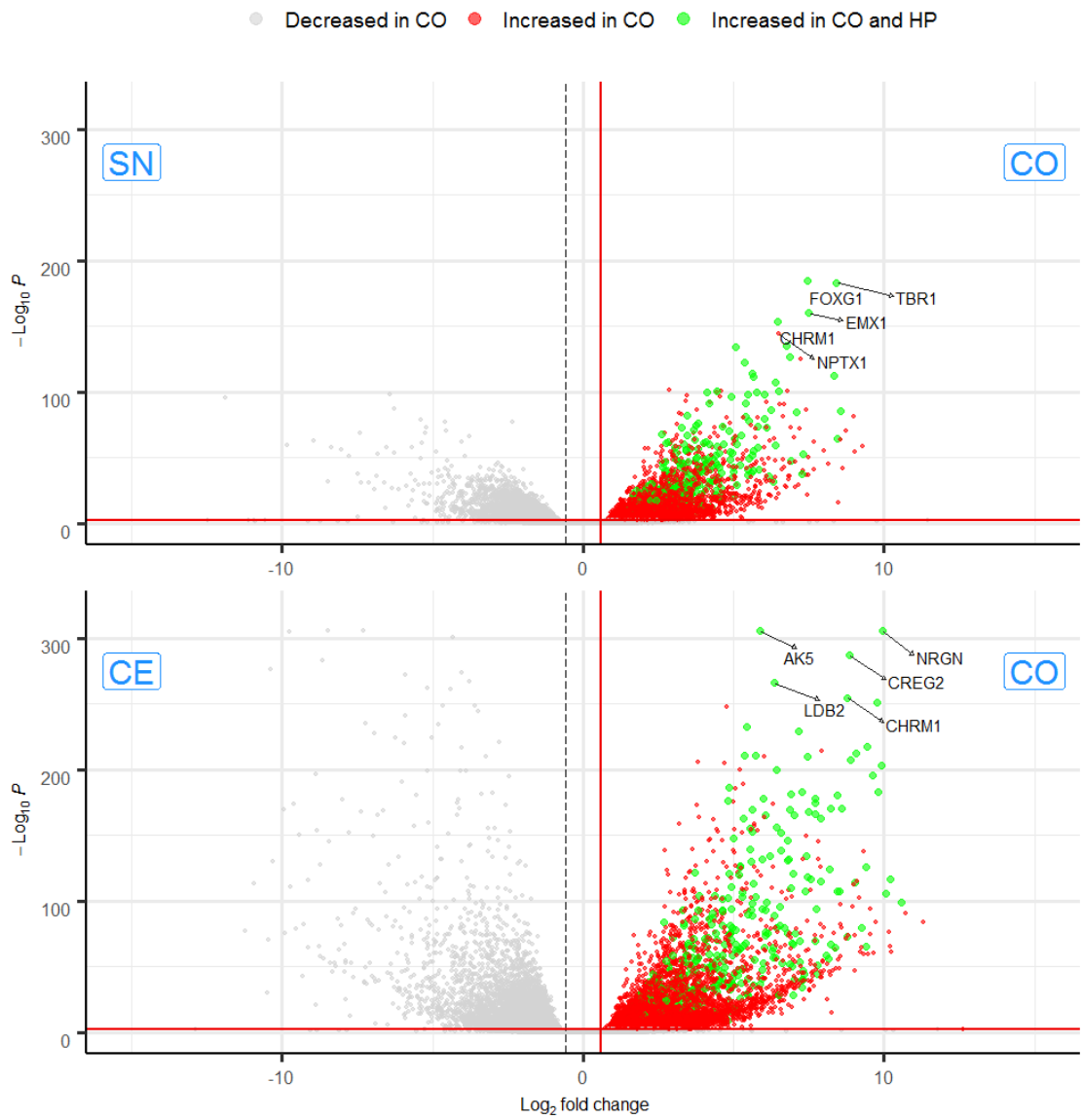

Figure 3.8 Enhanced Volcano plot - expression in Frontal Cortex vs control regions
Volcano plots showing log2 fold change (LFC) between (A) Frontal Cortex and Substantia nigra samples and (B) Frontal Cortex and Cerebellar Hemisphere samples. Genes with >1.5x expression ($p < 0.005$) in the Frontal Cortex samples are shown in red (higher expression against a single control region) or green (protein-coding genes with higher expression in CO and HP than in both control regions). The top 5 significant genes with higher expression in each plot are labelled. Red horizonal lines represent the cut off thresholds of p value < 0.005 (horizontal) and LFC > 0.6 (vertical).

The bar chart in Figure 3.9 shows how many protein-coding and non-protein-coding genes made up each of the contrasts. There were 3,705 and 2,073 protein-coding genes in the 6,147 CO-CE genes and 3,585 CO-SN genes, respectively. 1,730 genes (943 protein-coding and 787 non-coding genes) showed higher expression (LFC > 0.6; alpha = 0.005; p.adj < 0.005) against both control regions (Table 3.2). The genes were more

highly expressed in the CO, and because downstream analyses in this thesis used whole exome sequencing data (WES), only the protein-coding genes were considered (CO-genes). Protein-coding genes that were common to all 4 contrasts (CO|HP-genes) were coloured green in both volcano plots (in both Figure 3.8 and Figure 3.10). Non-significant genes and genes that do not have higher levels of expression than the control regions were coloured grey and were not considered in this thesis. The top 5 significant genes (FOXG1, TBR1, EMX1, CHRM1, NPTX1 for CO-SN; NRGN, AK5, CREG2, LDB2 and CHRM1 for CO-CE) were labelled in each volcano plot.

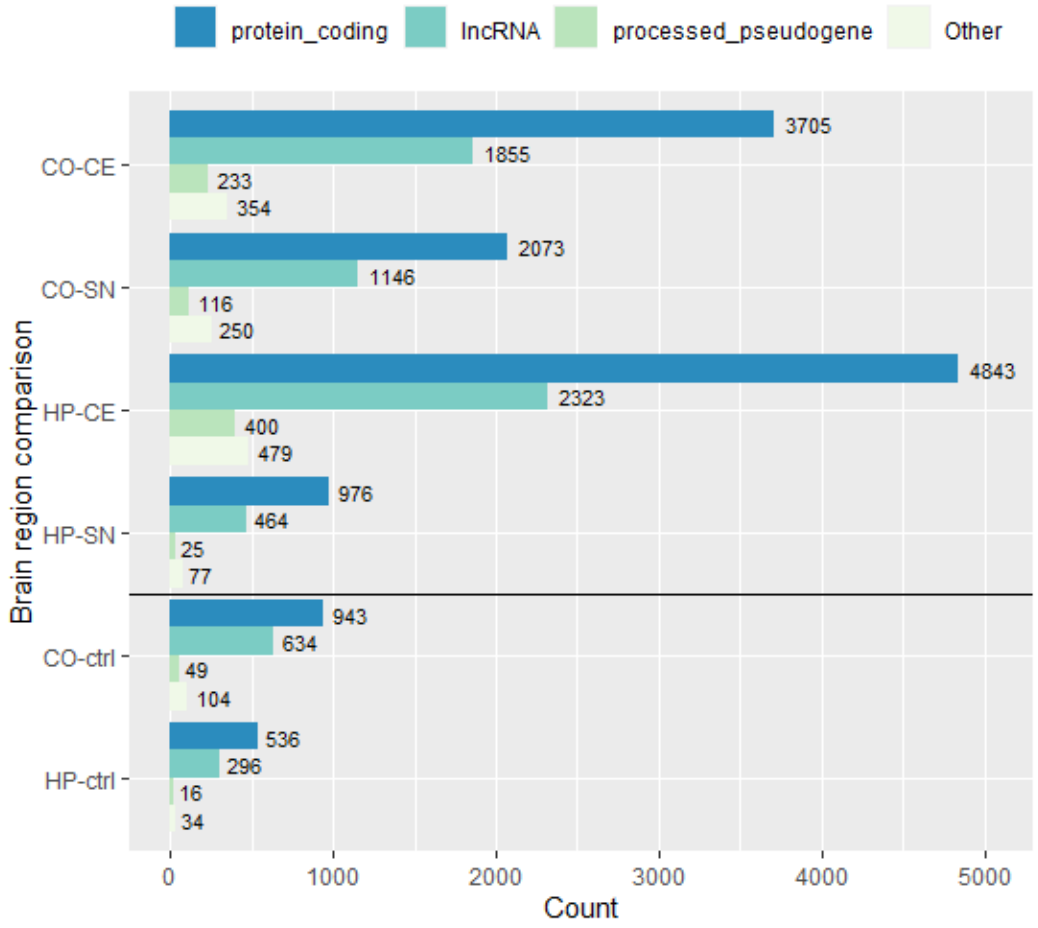

Figure 3.9 Bar chart of the division of biotypes in the brain region comparisons
The three largest components of gene biotype (protein-coding, lncRNA and processed pseudogene) are compared in the four brain region contrasts (CO-CE, CO-SN, HP-CE, HP-SN). CO-ctrl is the intersection between CO-CE and CO-SN. HP-ctrl is the intersection between HP-CE and HP-SN. Overall, 943 protein-coding genes with higher expression were identified from CO and 536 protein-coding genes from HP compared to both control regions CE and SN

3.3.4 Genes with higher expression in the Hippocampus compared to control region samples

The numbers of genes with higher expression in HP samples compared to control region samples were summarised in Table 3.2 above and plotted as volcano plots shown in Figure 3.10. In the contrast of HP to CE, of 56,391 genes with a non-zero total read count, 8,045 genes (14%) had greater than 1.5x higher read counts in the HP samples compared to CE samples within each donor (adj. $p < 0.005$). 17,423 genes (31%) were removed due to low counts. In the contrast HP to SN, 1,542 genes (2.7%) had a greater than 1.5x higher expression in the HP samples compared to SN samples within each donor (adjusted p-value < 0.005). 16,334 genes (29%) were excluded due to low counts. Log_2 fold change (LFC) was plotted against $-Log_{10}P$-value in the volcano plots in Figure 3.10 for HP-SN (top) and HP-CE (bottom). Genes that showed higher expression against only one control region are coloured red in each plot. The bar chart in Figure 3.10 shows how many protein-coding and non-protein-coding genes made up each of the HP contrasts. There were 4,843 and 976 protein-coding genes in the 8,045 HP-CE genes and 1,542 HP-SN genes, respectively. 882 genes – 536 protein-coding and 346 non-coding genes – showed higher expression (LFC > 0.6; alpha = 0.005; p.adj < 0.005) against both control regions (Table 3.2). These genes were considered as highly expressed in the HP for this thesis (HP-genes). Protein-coding genes that were common to all 4 contrasts (CO|HP-genes) are coloured green in both volcano plots (in both Figure 3.8 and Figure 3.10). Non-significant genes and genes that did not have higher levels of expression than the control regions were coloured grey and were not considered in this thesis. The top 5 significant genes (FOXG1, LHX2, SLC17A7, KCNH3 and ICAM5 for HP-SN; AK5, NRGN, FOXG1, CREG2 and DDN for HP-CE) were labelled within each volcano plot.

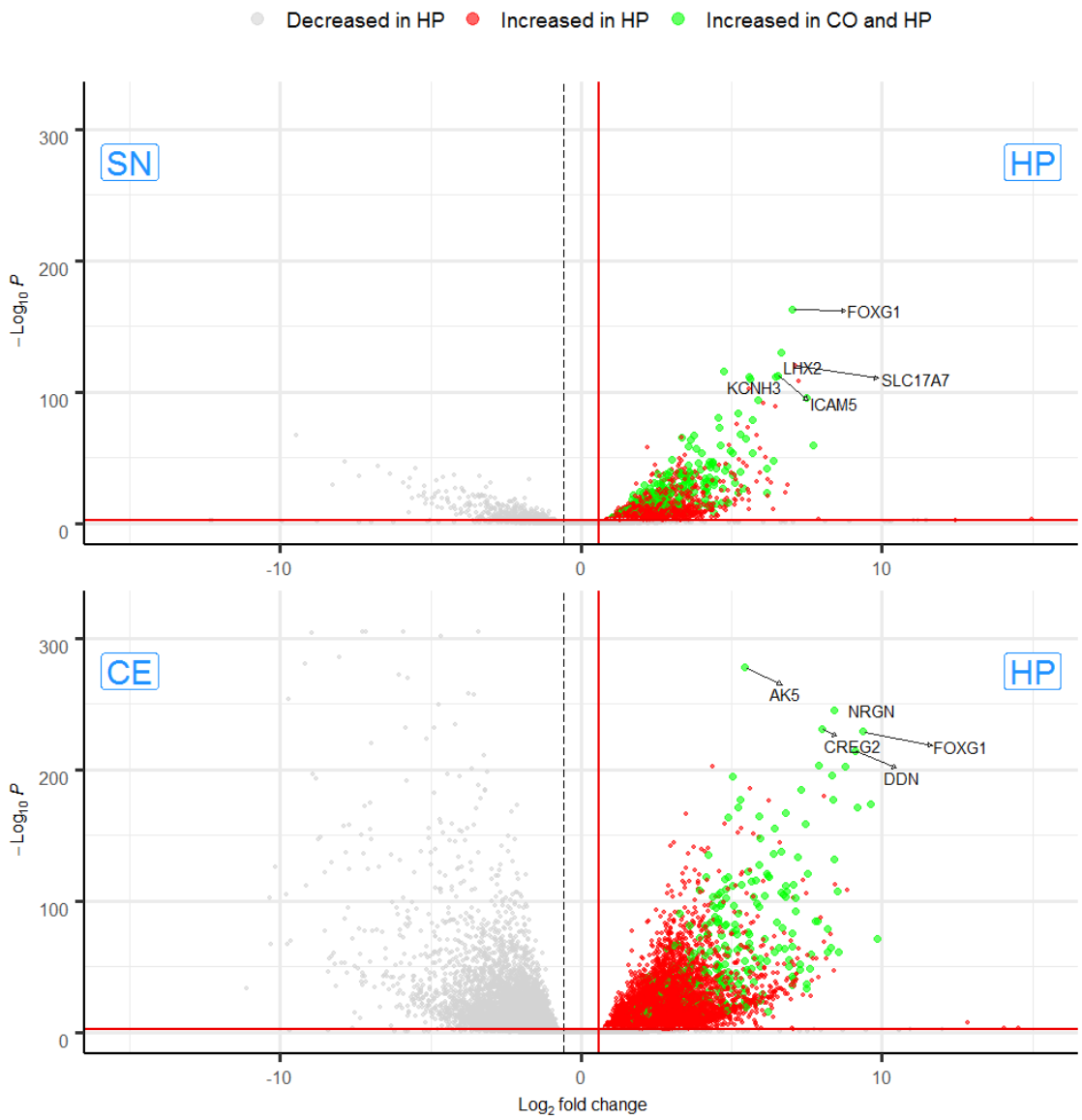

Figure 3.10 Enhanced Volcano plot - expression in Hippocampus vs control regions
Volcano plots showing log2 fold change (LFC) between (A) Hippocampus and Substantia nigra samples and (B) Hippocampus and Cerebellar Hemisphere samples. Genes with >1.5x expression ($p < 0.005$) in the Hippocampus samples are shown in red (higher expression against a single control region) or green (protein-coding genes with higher expression in CO and HP than in both control regions). The top 5 significant genes with higher expression in each plot are labelled. Red horizonal lines represent the cut off thresholds of p value < 0.005 (horizontal) and LFC > 0.6 (vertical).

3.3.5 Protein-coding genes with higher expression in the Frontal Cortex and Hippocampus

Figure 3.11 shows a Venn diagram of the 4 brain region contrasts filtered to protein-coding genes only. There were a total of 1076 protein-coding genes in the combined CO/HP-gene sets, with 403 protein-coding genes common to both sets (Table 3.2).

These represented genes with higher expression in either CO and HP compared to either control region. These genes are referred to in this thesis as CO/HP-genes. In Figure 3.11, the combined 1076 genes consist of all genes in the intersection of CO-CE and CO-SN and the intersection of HP-CE and HP-SN. The 943 protein-coding CO-genes divided across 4 areas (the intersection of CO-CE and CO-SN in Figure 3.11). 203 CO-genes had equal or lower expression in the HP than control regions. 283 CO-genes had higher expression in the HP compared to CE but not HP to SN. 54 CO-genes had higher expression in the HP compared to SN but not HP to CE. The 536 protein-coding HP-genes divided across 4 areas (the intersection of HP-CE and HP-SN). 56 HP-genes displayed equal or lower expression in the CO than control regions. 11 HP-genes displayed higher expression in the CO compared to SN but not CO to CE. 66 HP-genes displayed higher expression in the CO compared to CE but not CO to SN. The two zero values were in subsets where differentially expressed genes would not be expected.

Many protein-coding genes fell outside of the logical subsetscovered by the CO-, HP, and CO/HP-genes sets (Figure 3.11). For example, 2,364 protein-coding genes were common in both CO and HP more than in the CE but had equal or higher expression in the SN; and 315 protein-coding genes were common in both CO and HP more than SN but had equal or higher expression in the CE. These genes were out of the scope of downstream analyses.

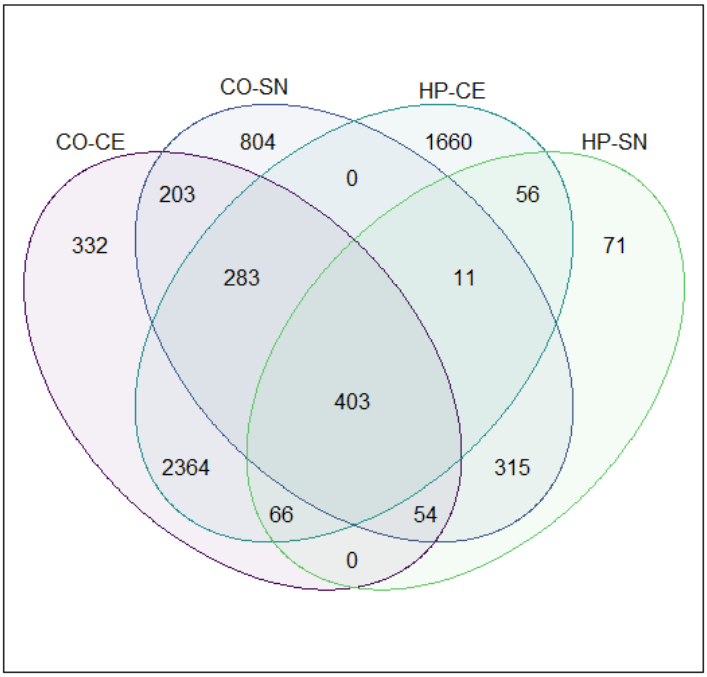

Figure 3.11 Venn diagram of protein-coding genes from the 4 brain region contrasts
The CO-genes fall within the intersection of CO-CE and CO-SN. The HP-genes fall within the intersection of HP-CE and HP-SN. The 1076 CO/HP-genes (Table 3.2) fall within the intersections of CO-CE/CO-SN and HP-CE/HP-SN with 403 genes common to both.

3.3.6 Validation against genes associated with SCZ in Open Targets

This validation step assessed the overlap between the identified region-specific gene sets and SCZ-associated genes from the Open Targets platform (Ochoa et al., 2021). The Open Targets genes were categorised based on a stringent threshold (association score > 0.75 and association score < 0.75).

There were 4,383 genes SCZ-associated in Open Targets. Among these, 309 genes had an association score > 0.75, of which 178 (57.6%) were present in all 6622 protein-coding genes within Figure 3.11 above. 4074 genes had a score less than 0.75, of which 1707 (41.9%) were present in all 6622 genes. Using the Venn diagram from Figure 3.11 as a template, Figure 3.12 shows where the genes in each of these two sets is located among the different intersections of the region-specific CO- and HP-genes.

Of the 309 strongly associated with SCZ in Open Targets, 53 (17.2%) were in the 1076 CO/HP-genes. The overlap for strongly associated genes was statistically significant (Fisher's Exact Test: p < 2.2e-16, OR = 0.25, 95% CI: 0.18–0.34). Of these 53 genes, 27 (8.7%) were in the 403 CO/HP-genes in common. There was a total of 49 (15.9%) genes in the 943 CO-genes and 31 (10.0%) in the 536 HP-genes (Figure 3.12, left). 55 (17.8%)

were found in the 2364 genes of the CO-CE ∩ HP-CE intersection, showing higher expression in the CO and HP against the CE, but not the SN (CO/HP>CE). 20 (6.5%) were found in the 315 genes that showed higher expression in both the CO and HP against the SN, but not the CE (CO/HP>SN). Other intersections contained less than 6.8% of the strongly associated genes.

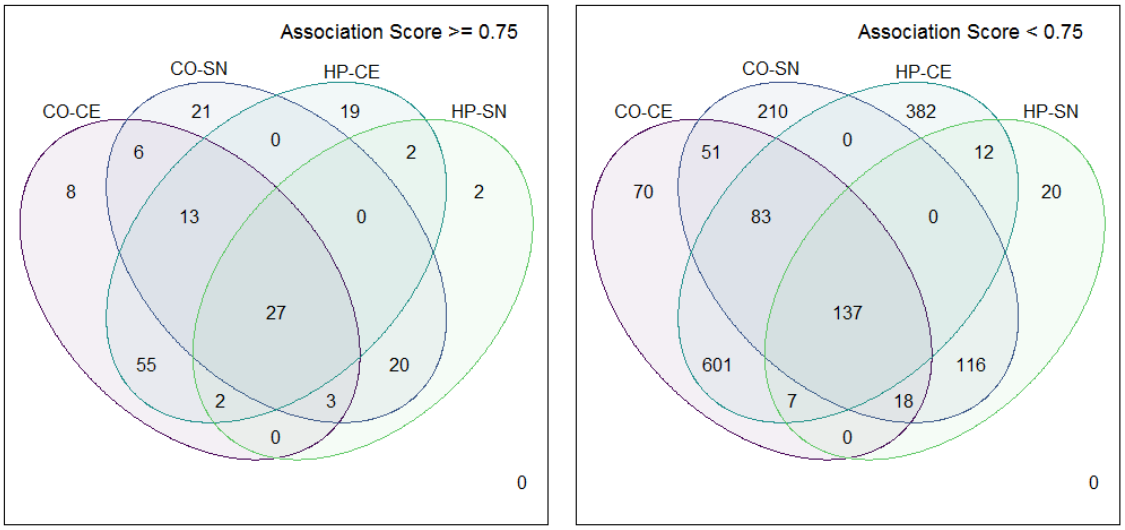

Figure 3.12 Venn diagram showing count of Open Targets platform SCZ-associated genes in the 4 brain region contrasts
The number of SCZ-associated genes from the Open Targets platform that are in each of the four brain region contrasts. Left: all genes with an association score greater than or equal to 0.75 (N=309). 139 genes did not map to any contrast. 53 strongly-associated genes (17.2%) are in the combined CO/HP genes (total of values in the two intersects CO-CE/CO-SN and HP-CE/HP-SN) ($p < 2.2 \times 10^{-16}$). Right: all genes with an association score less than or equal to 0.75 (N=4074) 2367 genes did not map to any contrast. 308 less-strongly associated genes (3.4%) are in the combined 1076 CO/HP genes ($p < 2.2 \times 10^{-16}$).

Of the 4074 Open Targets SCZ genes with an overall association score lower than 0.75 (Figure 3.12, right), 308 (7.6%) were in the 1076 CO/HP-genes (Fisher's Exact Test: $p < 2.2 \times 10^{-16}$, OR = 0.16, 95% CI: 0.14-0.18). Of these 308 genes, 137 (3.4%) were in the 403 CO/HP-genes; 289 (7.1%) were in the 943 CO-genes; and 156 (3.8%) were in the 536 HP-genes. For regions outside of the CO- and HP-genes, 601 (14.8%) were in the 2364 CO/HP > CE genes and 116 (2.8%) in the 315 CO/HP > SN genes. Other intersections contained less than 9.4% of the lower scoring genes from Open Targets.

3.3.7 Mean expression of CO/HP-genes across all GTEx brain regions

The mean expression for all 1076 CO/HP-genes was calculated for each brain sample in GTEx and was overlaid on the PCA plot from section 3.3.1.1 (Figure 3.13) to indicate the distribution of the CO/HP-genes across all brain samples in GTEx v8.

Jointly, CO/HP-genes display highest expression across the Frontal and Anterior Cingulate Cortex regions. Expression is high to medium in hippocampus, amygdala and in some samples among the striatum (putamen, caudate, and nucleus accumbens). Expression is medium in hypothalamus and medium-low in both substantia nigra and spinal cord. Expression of CO/HP-genes is medium-low in cerebellum samples.

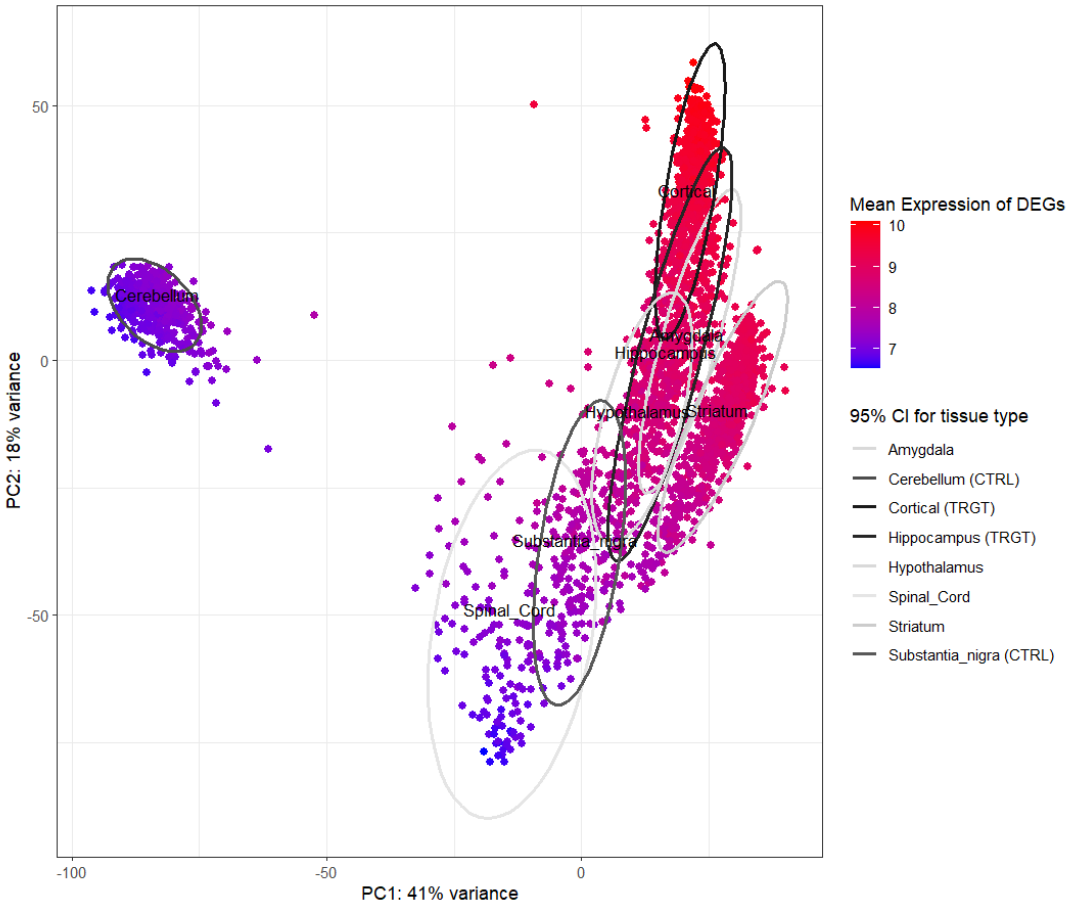

Figure 3.13 PCA plot coloured by mean expression of CO/HP-genes per GTEx sample
The PCA plot of all GTEx brain samples (shown above in 3.3.1.1) was re-plotted with the mean expression of all CO/HP-genes per sample (red = high expression; blue = low expression). 95% CI ellipses were altered to show their outline (darker outlines indicate the four GTEx regions in this analysis). TRGT: target brain regions, CTRL: control/contrast brain regions.

3.3.8 Expression Weighted Celltype Enrichment in single-cell cortex and hippocampus

Genes from CO-genes and HP-genes were tested as separate gene lists for enrichment in single-cell mouse cortex and hippocampus.

Of the CO-genes, 684 of the 943 genes were found in the mouse-human gene name mapping and 27 significant cell type enrichment results were identified ($q < 0.05$). The strongest enrichment signal was in L2/3 pyramidal cells of the somatosensory cortex, and to a lesser degree in pyramidal cells across other cortical layers. Genes were significantly enriched at a lower level in most interneurons (incl. Somatostatin (SST), HTR3a/Reelin, HTR3a/VIP and Parvalbumin (Pvalb) subtypes). There were 403 genes shared between the CO and HP gene lists. Thus, there were also strong signals in hippocampal cell subtypes for CO-genes.

Of the HP-genes, 384 of the 536 genes were found in the mouse-human gene name mapping and 17 significant cell type enrichments we identified ($q < 0.05$). The strongest enrichment signals were in the three glutamatergic pyramidal cell subtypes in CA1 of the hippocampus, one subtype of pyramidal cells in CA2 and another of the subiculum. Significant enrichment signals were also found in pyramidal cells of the S1 cortex, but at a lower level than the CO-genes. Enrichment was only significant in three types of GABAergic interneuron (type 1 and 2 – SST interneurons; and type 9 – HTR3a/Vip/Calb2). Interestingly, a significant signal was also found in ependymal cells – the ciliated cells that line the ventricles. Ependymal cells express many genes related to ciliogenesis, cilia structures, and their regulation (Zeisel et al., 2015)

This single-cell enrichment analysis demonstrated that the differentially expressed genes were significantly enriched in glutamatergic and GABAergic cells in the cortex and hippocampus and the signal from non-neuronal cells, such as oligodendrocytes, astrocytes and microglial cells was absent.

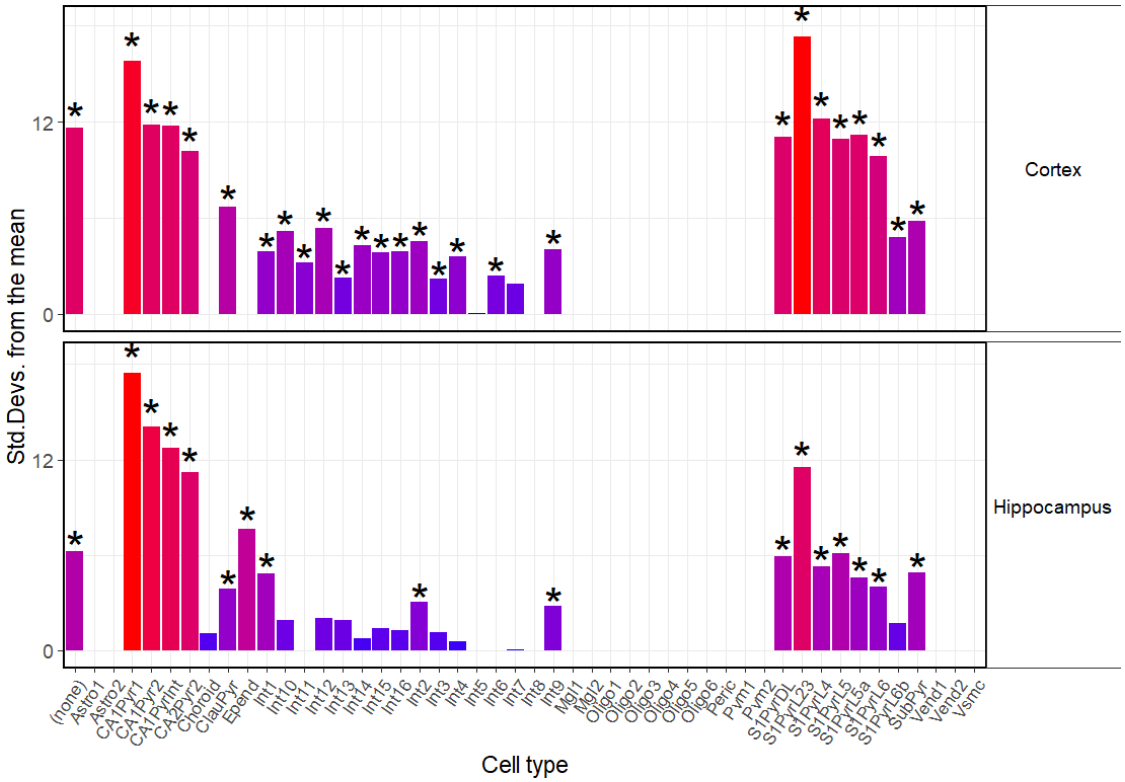

Figure 3.14 Expression Weighted Celltype Enrichment in Cortex and Hippocampus
EWCE plot showing enrichment of CO-and HP-genes in cortical and hippocampal cell subtypes from single-cell mouse brain (Zeisel et al., 2015). Astro: astrocytes; Int: interneurons; Epend: ependymal cells; CA[n]Pyr: pyramidal cells of the hippocampus; Choroid: choroid plexus; ClauPyr: pyramidal cells of the claustrum; Oligo: oligodendrocytes; Mgl: microglia; Peric: pericytes; Pvm: perivascular macrophages; S1Pyr: pyramidal cells of the somatosensory cortex; SubPyr: pyramidal cells of the subiculum; Vend: vascular endothelial; Vsmc: vascular smooth muscle cells. Significant enrichment (adj. p value < 0.05) indicated by asterisks.

3.3.9 Gene set enrichment analysis for CO- and HP-genes

In an effort to extract biologically useful terms all genes and terms were further analysed using Functional Annotation Clustering via the NIH DAVID Bioinformatics server (Huang, D. W. et al., 2009; Sherman et al., 2022). The clustering combined all GO Biological Process, Molecular Function and Cellular Compartment terms, along with KEGG and Reactome pathways terms into one functional chart.

A full enrichment analysis was performed on the CO-genes and HP-genes separately using EnrichR in R and the full lists are available in Appendix 8.3).

3.3.9.1 Functional Annotation Clustering of 943 CO-genes

The CO-genes, genes with higher expression in the frontal cortex compared to both the SN and cerebellum samples, showed the highest enrichment scores in 5 signalling-related clusters including many G protein-coupled receptors (GPCR) and downstream signalling. These also included several different neurotransmitters and internal secondary messenger signalling pathways. Genes were also significantly enriched in terms related to metabolic processes, for example, insulin, and energy metabolism.

Genes were enriched in terms relating to SCZ-associated Ephrin-EPH signalling that was recently implicated in L5/6 excitatory neurons of the frontal cortex (Huuki-Myers et al., 2024).

Although functional annotation clustering is considered significant for enrichment scores over 1.31 (equivalent to $p < 0.05$), if there were damaging rare variants in a gene in this list, then it would affect the pathways in which it resides, irrespective of the enrichment score. Therefore, a few non-significant clusters are included in the clustering tables below to show potential affected processes and pathways.

Table 3.3 Functional Annotation Clustering of 943 CO-genes
Functional annotation cluster performed in DAVID Bioinformatics. Enrichment score is the geometric mean of the gene enrichment scores for all terms in the cluster. Gene enrichment scores greater than 1.3 correspond to p values < 0.05. Several clusters have been retained below significance level, shaded in red.

Functional Cluster	Enrichment Score	Gene Count
GPCR Signalling and Neuroactive Ligand Interaction	14.4	189
Neuronal Communication and Signal Transmission	10.8	64
Monoamine GPCR Signalling Pathways	7.15	32
Potassium Channel Function and Regulation	6	38
Neuropeptide Signalling and Binding	5.37	29
Regulation of Insulin Secretion and Energy Metabolism	3.67	14

Functional Cluster	Enrichment Score	Gene Count
Photoreception and Light Response	3.41	8
Regulation of Synaptic Membrane Potential	3.21	12
Neural Development and Axon Guidance	3.18	82
GABAergic Neurotransmission and Regulation	2.83	57
EPH-Ephrin Signalling Pathways	2.2	11
Cardiac and Muscle Physiology	2.2	16
Phosphorylation and Kinase Signalling	2.15	93
Integrated Cellular Signalling and Metabolic Regulation	2.1	77
Cardiac Function and Calcium Signalling	1.98	41
G-Protein and Second Messenger Signalling	1.87	25
Canonical Cellular Signalling Pathways and Differentiation	1.48	94
Glycoprotein Synthesis and Metabolism	1.34	22
Neuroendocrine and Synaptic Signalling Pathways	1.33	30
Hippo Signalling Pathways and Multicellular Development	1.31	17
Cadherin-Catenin Complex and Cell Junction Dynamics	1.21	35
GPCR-Driven Signalling Networks and Cellular Functions	1.17	26
Platelet Function and Signalling Pathways	1.17	20
Signalling Pathways in Inflammation	1.13	23
Neurotrophic and Signalling Pathways	1	11
MET Signalling and Cell Motility	0.99	6
Cell Migration, Survival, and Growth	0.88	25
Calcium-Regulated Ras and MAPK Signalling	0.85	48
Protein Glycosylation	0.81	14

Functional Cluster	Enrichment Score	Gene Count
Protein Tyrosine Phosphatases in Receptor-Mediated Signalling	0.79	11
DNA Binding and Transcription Regulation	0.78	104
Extracellular Matrix and Collagen Organization	0.69	25
Heme and Iron Binding in Cytochrome P450 Oxidation	0.65	14
GTPase-Mediated Signalling and Apoptosis Pathways	0.56	23

3.3.9.2 Functional Annotation Clustering of the 536 HP-genes

HP-genes, genes with higher expression in the hippocampus compared to both the SN and cerebellum samples, similarly showed the highest enrichment scores in signalling-related clusters, including GPCR-signalling, excitatory neurotransmitter receptors, synaptic transmission, neuronal development, axon guidance.

Table 3.4 Functional Annotation Cluster of 536 HP-genes
Functional annotation cluster performed in DAVID Bioinformatics. Enrichment score is the -log of the geometric mean of the gene enrichment p values for all cluster terms. Gene enrichment scores greater than 1.3 correspond to p values < 0.05. Several clusters have been retained below significance level, shaded in red.

Functional Cluster	Enrichment Score	Gene Count
GPCR-Mediated Neuroactive Signalling	10.62	107
Potassium Channel Signalling and Transport	6.95	31
Rhodopsin-Like GPCRs and Monoamine Signalling	6.63	46
Neurotransmitter Receptors and Synaptic Transmission	3.1	30
AMPA Receptor Trafficking and Synaptic Plasticity	3.01	12
Nervous System Development and Axon Guidance	2.54	46

Functional Cluster	Enrichment Score	Gene Count
Glutamatergic and Serotonergic Synapses and Ras Signalling	1.75	17
Regulation of Insulin Secretion and Energy Metabolism	1.74	7
MET Signalling and Extracellular Matrix Components	1.7	5
Collagen Biosynthesis and Structural Integrity	1.7	28
Wnt Signalling and Neuron Differentiation	1.36	64
Protein Kinase Activity and Phosphorylation	1.33	44
Oxytocin and MAPK Signalling Pathways	1.32	16
DNA Binding and Transcription Regulation	0.96	55
Heparan Sulphate and Glycosaminoglycan Metabolism	0.95	6

In summary, further to the enrichment found in mouse cell subtypes of the hippocampus and somatosensory cortex in 3.3.8, the functional annotation clustering supported the neuronal specificity of the CO- and HP-genes, revealing a substantial number of functions, compartments, and pathways with distinct regional relevance to the Frontal Cortex and/or Hippocampus. These findings highlight the complex molecular landscape associated with these specific brain regions. Although the identified pathways are predominantly neuronal, the presence of additional non-neuronal pathways hints at potential pleiotropic effects of certain genes within the CO- and HP-genes.

3.4 Discussion

The analyses in this chapter aimed to identify a set of genes with higher expression in two SCZ-affected brain regions compared to 2 control regions with lower evidence for being affected by the core symptoms of SCZ. For this study, the analysis involved exploring normal gene expression differentials in healthy brain regions that correlate with those regions thought to be affected in SCZ patients. In the first instance, several methodological analyses were conducted to establish an alignment strategy and model for the main analysis. In the main differential expression analysis, four brain regions

were used with two target regions (Frontal Cortex and Hippocampus) and two control regions (Cerebellar Hemisphere and Substantia nigra).

3.4.1 Exploratory Methodological Findings

The investigation into optimal alignment strategies and models was undertaken to enhance the precision of subsequent differential expression analyses. This was important for identifying genes that exhibit higher region-specific expression in the target brain regions.

3.4.1.1 PCA analysis showed clustering of target and control regions

A PCA analysis was performed on gene read counts obtained from GTExPortal.org to explore the distribution of the four selected GTEx brain regions among 2642 samples across all 13 brain regions in GTEx. In the PCA plot (Figure 3.3), PC1 explained 41% of the variance in the top 1000 genes with the highest variance across all samples, and this separated out cerebellum, spinal cord, and partially the substantia nigra from the rest of the brain regions. PC2 explained 18% of the variance in the data, but the cause for the variance was less obvious. Cerebellum samples were tightly clustered in both PC1 and PC2 suggesting uniformity of gene expression between samples. Cerebellum samples were tightly clustered in PC2, but overlapped a subset of samples in hippocampus, cortex, and amygdala samples. This suggested a set of genes of similar function may be common to specific cell types in these regions. The PCA indicates how other GTEx brain region compare to the selected Cerebellum and SN as control regions. If other regions had been selected as controls, there would have been a much larger overlap in both PC1 and PC2 in the expression with the two target regions, potentially eliminating many key genes. The two brain regions selected, Cerebellum and Substantia nigra appear the best choices to use as controls from among the GTEx brain samples.

3.4.1.2 Rsubread mapped more reads uniquely than STAR

A comparison of STAR v2.7.1a 15/05/2019 (Dobin et al., 2013) and Rsubread v1.34.7 (Liao, Y. et al., 2019), employing various alignment protocols and trimming strategies (3.2.4), revealed a considerable performance gain for Rsubread. Aligning 20 BrainSeq RNA-Seq datasets (described in 3.2.1.2) to the human reference genome (GRCh38) with GENCODE v31, Rsubread consistently mapped between 2% and 11% more reads per dataset compared to STAR (Figure 3.5).

The decision to compare Rsubread was supported by its performance in a comparative analysis by the developers in Liao et al. (2019). The study demonstrated slightly greater correlation of Rsubread's alignments to known log2-expression values for 958 genes in two Sequencing Quality Control Consortium (SEQC) Project (SEQC/MAQC-III Consortium, 2014) RNA-Seq datasets and greater correlation of 28,395 reads in a simulated dataset compared to STAR. This held true despite STAR's widespread use in projects such as GTEx v7 (Mele et al., 2015), GTEx v8 (Broad Institute, 2020), and the ENCODE project (ENCODE Project Consortium, 2012).

Rsubread uses a mapping paradigm known as 'seed-and-vote' utilising a hash table to store the positions of all unique 16-mer subreads in the human reference genome. This approach enables efficient and fast read mapping without the need for trimming (Liao, Yang & Shi, 2020). Rsubread can run with parallelization in R (covered in Liao, Y. et al., 2019). Integrated with quantification using featureCounts and differential expression analysis using DESeq2, the entire workflow can be streamlined.

It is important to acknowledge that this study relied on uniquely mapped reads as the main metric for mapping accuracy. While this metric is highly informative, capturing the majority of reads, the complexity of the transcriptomic landscape introduces subtle variations and potential sources of error, particularly in the presence of shorter fragmented reads. As highlighted by Saha and Battle (2018), the choice of alignment and quantification methods significantly influences the occurrence of false positives in gene expression analyses. Their study emphasizes the importance of employing improved methods to generate reliable gene expression data, aligning with the approach in this chapter, where the GTEx raw data was independently mapped. The precision of the alignment strategy and maximizing alignments of noisy datasets set the foundation for our subsequent differential expression analysis, contrasting two SCZ-relevant brain regions against two control regions.

3.4.1.3 Model development - using 2 brain regions

In the development of the differential expression methodology for the main analysis, GTEx RNA-Seq datasets of the Frontal Cortex were compared with Cerebellar Hemisphere. Aligned with Rsubread to GRCh38 and GENCODE v31 gene annotation, 83 healthy donor datasets for each region were analysed for differential expression using DESeq2, with sex selected as a covariate.

During this analysis, a potentially mislabelled Frontal Cortex dataset among the Cerebellar Hemisphere datasets was identified through Principal Component Analysis (PCA). This anomaly was corroborated using the GTExPortal's Expression PCA tool (https://www.gtexportal.org/home/expressionPca), leading to its removal. Figure 3.15 shows the output when "Brain - Cerebellar Hemisphere" is selected. Notably, the dataset remains in the GTEx v8 data freeze, emphasising the importance of independent raw data analysis.

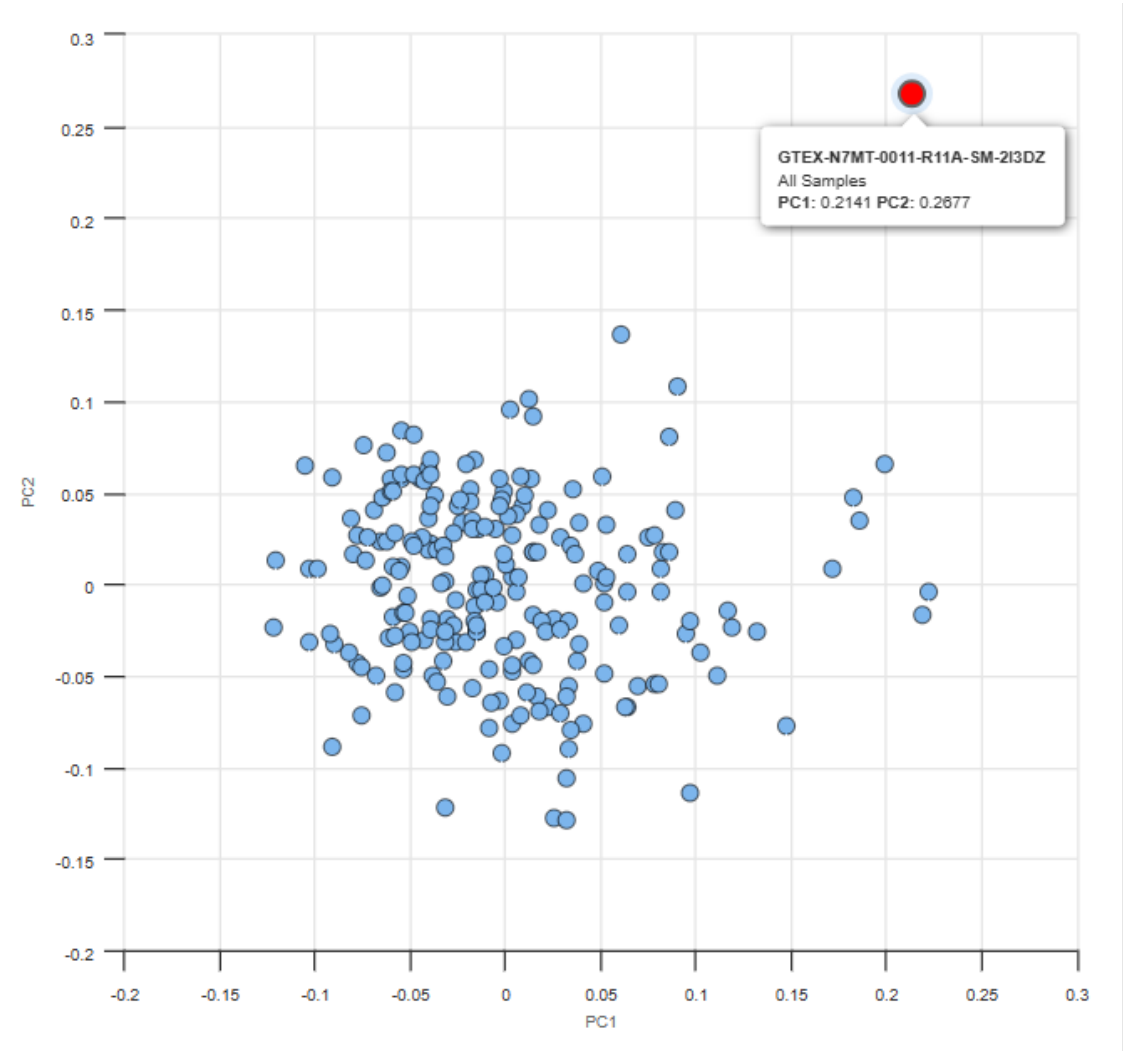

Figure 3.15 GTExPortal: Expression PCA for all GTEx Cerebellar Hemisphere samples
Expression PCA analysis was run for "Brain – Cerebellar Hemisphere". The outlier dataset GTEX-N7MT-0011-R11A-SM-213DZ in this chapter (3.3.1.4) showed an expression signal close to all female Frontal Cortex datasets. (source: GTExPortal.org)

The subsequent differential expression analysis revealed significant differential expression in 15,690 genes surviving adjustment for multiple testing (adj. P < 0.005). A revised PCA reaffirmed distinct clustering by PC1 (87% variance explained by brain region) and PC2 (3% variance explained by sex), highlighting the robustness of the remaining datasets.

In the absence of existing literature comparing Frontal Cortex and Cerebellar Hemisphere gene expression, these results suggest a substantial divergence in their transcriptomic profiles. Acknowledging the broad insights from the two-region analysis, the main analysis (3.4.2) encompassed four brain regions. This provided a more nuanced identification of region-specific gene expression patterns within the Hippocampus and Frontal Cortex – identified in Section 2.4 as more relevant in SCZ.

This distinction between Frontal Cortex and Cerebellar Hemisphere was further validated with a brief analysis using BioJupies (Torre et al., 2018), which is a customised RNA-Seq workflow in an online Jupiter notebook with access to pre-analysed GTEx datasets. This confirmed the broad division of gene expression between Frontal Cortex and Cerebellar Hemisphere gene expression landscapes using sample sizes over 5 times larger than in this study. BioJupies was not used in this analysis; and it cannot perform a 4-way analysis of datasets. Future work could explore the similarities and differences between the four contrasts in this study and individual contrasts analysed via BioJupies.

3.4.2 Main Differential Expression Analysis in Four Brain Regions

RNA-Seq datasets from GTEx, including Frontal Cortex (CO), Hippocampus (HP), Cerebellar Hemisphere (CE) and Substantia nigra (SN) from 25 healthy donors, underwent alignment to the GRCh38 reference genome and GENCODE v31 gene annotation using Rsubread. DESeq2 enabled the collective analysis of counts per gene per sample. Four pairwise contrasts were derived – CO-CE, CO-SN, HP-CE, and HP-SN. These contrasts were combined, and genes exhibiting increased regional expression (> 1.5-fold change; p. adj. < 0.005) compared to the control brain regions were categorised as CO-, HP- or CO/HP-genes (summarised from Table 3.2).

- Out of 60,992 coding and non-coding genes (GENCODE v31), the RNA-Seq analysis revealed:

- 1,730 genes (942 protein-coding) CO-genes (Section 3.3.3)
- 882 genes (536 protein-coding) HP-genes (Section 3.3.4)
- 1076 protein-coding CO/HP-genes, of which 403 genes (Section 3.3.5) were common to both..

Despite the absence of a single comprehensive study validating these results, two studies were identified that focussed on the regional expression of gene families (Sequeira et al., 2019; Shen et al., 2020) important in neuronal signalling and SCZ (Steiner et al., 2016). Specifically, 6 $GABA_A$ receptor subunits (GABAAR) in the CO- and HP-genes aligned with the results in Sequeira et al. (2019), which studied the regional expression across 111 brain regions of 14 GABAAR genes. The same group extended their study to ionotropic glutamate receptors (AMPAR), GRIA1, GRIA2, GRIA3 and GRIA4 (Shen et al., 2020), of which GRIA3 was exclusive to CO-genes. The following summarises the validations:

- **GABAAR Validation**
  - GABRA2, GABRA3, GABRA4, GABRA5, GABRG3, and GABRQ showed consistent expression patterns with Sequeira et al. (2019).
    - GABRA4 was exclusive to CO, not HP.
    - GABRQ was exclusive to the HP, not CO.
  - GABRA4 exhibited higher expression in SNc compared to SNr, suggesting potential GTEx sample origin in the pars reticulata rather than pars compacta.
  - GABAAR genes not in CO- or HP-genes showed regional patterns that agree with their non-inclusion.
  - GABAAR not in our results were consistently upregulated in the control regions or in low amounts in HP and CO – there were no false positives from this family. For example, GABRA1, GABRB2, GABRD, GABRG2, GABRA6 were highly enriched in the Cerebellar Hemisphere.
- **AMPAR Validation:**
  - GRIA3 showed consistent expression patterns in Shen et al. (2020) – high expression in CA1 in the HiF (hippocampal formation) and the dentate gyrus, and lower enrichment in other regions studied here.

- GRIA1 and GRIA2 were not consistent with CO- and HP-genes. GRIA1 had 10x enrichment in CA4 of the HiF.
- GRIA4 was enriched in the cerebellum consistent with the findings in this chapter.

These two studies only validate 10 of the 1076 CO- and HP-genes. In the absence of studies in regional expression signatures in healthy brains for many of these genes, two sets of analyses were performed:

1. Overlap analysis – This analysis aimed to ascertain whether CO- and HP-genes were enriched in genes associated with SCZ using data from the Open Targets platform.
2. Enrichment analyses – These analyses aimed to assess whether CO- and HP-genes were involved in pathways associated with the Frontal Cortex and Hippocampus. This involved examining enrichment in Gene Ontology (GO) and pathways from KEGG and Reactome.

3.4.3 Validation against genes associated with SCZ in Open Targets

A set of genes associated with SCZ were downloaded from the Open Targets platform and divided into two sets based on an overall association score (OAS) greater or lower than 0.75. These two sets were assessed against CO-, HP- and CO/HP-genes.

- Between 8.7% and 15.9% of 309 SCZ-associated genes with a more stringent OAS (>0.75) overlapped with protein-coding genes:
  - 49 of the 309 genes (15.9%) overlapped with the 943 CO-genes (5.2%)
  - 31 (10.0%) overlapped with the 536 HP-genes (5.8%)
  - 53 (17.2%) overlapped with the 1076 combined CO/HP-genes ($p<2.2 \times 10^{-16}$) and of these 27 (8.7%) overlapped with 403 genes (6.7%) in common to both CO and HP
- Between 3.4% and 7.1% of the remaining 4,074 SCZ-associated genes (OAS < 0.75) overlapped with protein-coding genes:
  - 289 of 4074 genes (7.1%) overlapped with the 943 CO-genes (30.6%)
  - 156 (3.8%) overlapped with the 536 HP-genes (29.1%)
  - 137 (3.4%) overlapped with 403 genes (34.0%) common to both CO and HP

The CO-genes represent 943 genes with regional specificity to the Frontal Cortex compared to the Cerebellar Hemisphere and Substantia nigra. The genes are involved in multiple pathways (3.3.7 and 3.4.4) that are specific to the Frontal Cortex. Many of these may have no relation to SCZ. There is a modest overlap of 15.9% of the genes with a stringent OAS above 0.75.

More SCZ-associated genes (57 of the 309 with OAS > 0.75; 216 of the 4474 with OAS < 0.75) had increased expression also in the substantia nigra. This could indicate that the substantia nigra is problematic as a contrast brain region. Including the substantia nigra for this thesis could hide a small number of genes relevant to SCZ. Some genes are differentially expressed between the SN pars compacta and SN pars reticulata (GABRA4 – see 3.4.2) and GTEx tissue sampling procedures may not have discerned between the two nuclei (see Figure 2.7).

Open Targets employs a set of heuristics to link genes with diseases, distinguishing between direct and indirect associations based on the original source. While the overall association score serves as a valuable metric within this data source, it may not inherently indicate a direct association with SCZ. Curated databases, such as DisGeNet (Pinero et al., 2020) may offer more nuanced gene-disease relationships. A DisGeNet enrichment analysis could also reveal a wider spectrum of diseases enriched with the CO- and HP-genes (proposed in Section 3.5).

3.4.4 Differentially expressed genes showed greater expression across target regions

The combined expression level of CO/HP-genes derived from all brain samples in GTEx v8 (see 3.3.7) indicated that the analysed gene lists are expressed at higher levels in the frontal cortex and hippocampal regions, as well as in the anterior cingulate cortex. This plot also suggested that the genes had higher expression in the amygdala. The high-to-medium level of expression seen in a portion of the striatal samples could be accounted for by GABAergic transmission-related genes active in the cortex and throughout the striatum. The signal was similar in the hypothalamus, but this area overlapped the hippocampus and amygdala, more compared to the striatal regions. If rare variants are to be found in these genes, there is a strong likelihood that their effects will be seen in any of those regions of the brain, even to a much lesser extent in the cerebellum, substantia nigra or spinal cord. However, if SCZ is oligogenic, any

individual would only require one or more damaging variants in any of these 1076 genes.

The distribution of mean expression across the different samples also suggested that the spinal cord region may make a third control region or replace the cerebellum, as it separates out from the rest of the brain samples in both PC1 and PC2.

The overlap of some cerebellum samples in PC2 with the cortex and hippocampus could have removed some SCZ-relevant genes, except that for genes to be included in the CO-genes or HP-genes, they must have higher expression in either of those regions compared to both the cerebellum and substantia nigra. This would explain why the mean expression in hippocampal samples did not show a drop in samples that overlapped with the cerebellum samples along PC2.

3.4.5 Genes showed enrichment in key cell types from cortex and hippocampus

The enrichment of expression signals for the CO/HP-genes found in mouse hippocampus and cortex indicated that the genes differentially expressed against the control regions were expressed in some established subtypes of excitatory and inhibitory neurons found in the hippocampus and cortex. It was interesting to note that no enrichment was found in expression data from oligodendrocytes, astrocytes or microglia. Another interesting finding was the enrichment in ependymal cells. These multiciliated cells in the ventricles, may have been part sampled along with the hippocampus. Genes relating to the cilia complexes are the highest signal in ependymal cells and could explain the cilia genes identified in the HP-genes list (Zeisel et al. 2015).

This analysis was performed using the default CellTypeDataset in the EWCE package (Skene & Grant, 2016), however it is possible to create tailored datasets for further study. A more comprehensive analysis could be designed in which cell type data from multiple brain regions is collated from resources such as the Human Cell Atlas (Amit et al., 2024) or scBrainMap (Chi et al., 2023) and compiled into a EWCE dataset. The computing resources that would be required to perform this would have to be factored into the project.

3.4.6 CO/HP-genes were enriched in neuronal and metabolic pathways

Gene set enrichment analyses and functional annotation clustering were performed for CO- and HP-genes in Gene Ontologies (GO), KEGG, Reactome, and Wikipathways pathways.

The main findings from these analyses were that CO and HP genes were enriched in:

- Biological processes including neurotransmitter signalling, cellular transport and cell orientation, cell metabolism, cellular structure.
- Cellular components including dendritic and axonal components, synaptic cleft, postsynaptic density, ion channels, plasma membrane components.
- Molecular functions including enzyme activity (differentially affecting CO and HP), receptor activity, ion channel activity, hormone activity.
- KEGG pathways including signalling pathways, vascular smooth muscle contraction (HP-genes)
- Reactome pathways - neurotransmitter receptor, ion channel, neuronal system and GPCR signalling, synaptic transmission and plasticity, insulin regulation

The pathways discovered agree with, but also add to those pathways identified in other studies including in the latest SCHEMA (Singh et al., 2022) and PGC3 GWAS (Trubetskoy et al., 2022).

In the next chapter, the CO- and HP-genes were used to filter rare variants in whole exome sequencing data from a group of 100 patients with SCZ.

3.5 Limitations and Further Work

The experiments in this chapter have several limitations and as newer resources become available, they present avenues for further analysis.

1. Threshold for 'Higher Expression' was set to 1.5 times. The selection of this value was arbitrary. It produced a reasonably sized set of genes. It may be more appropriate to explore different values or explore using a scaled value. For example, cytoskeleton genes are involved in the structure of synapses and maintenance of dendrites and axons, but because the level of mRNA in a cell

for these genes may depend on cell size rather than location, important regional expression differences could be difficult to analyse through mRNA levels. Differences are more likely to be revealed in single-cell mRNA data.

2. Transcription Level Differential Expression Analysis. This chapter explored gene level expression analysis. Differential gene expression can hide certain changes in transcripts between brain regions that do not affect the 'reads per gene' metric. If the difference between two brain regions for a single gene involves alternate splicing of two exons of equal length, this study will not detect it. Further work could explore exon-level or base-level differential expression analysis. This would detect regional differences expression with more resolution to identify potential loci that contain variants in SCZ.
3. Bulk RNA-Seq data. GTEx sampled 5mm cubes of brain tissue. The gene expression detected in these samples lacks cell-specificity due to noise from the wide variety of cell types. The highest expressed genes detected could be the top genes expressed in each cell type, rather than a signature of any one cell type. Single-cell RNA-Seq could counter this allowing for selection of SCZ-relevant cell types in SCZ-relevant brain regions. One recent, on-going project from the BRAIN Initiative's Cell Census Network sequenced over 3 million cells from multiple brain regions (Hawrylycz, Michael et al., 2023).
4. Validation Resources. Human Brain Map (Hawrylycz, Michael J. et al., 2012) from Allen Institute. The Human Brain Map online, exploratory utility can be used to explore regional gene expression from 6 healthy donors. Target brain structures and contrast structures can be selected to produce a list of microarray gene probes that are differentially enriched in the target structures against the contrast structures. A comparison of the CO- and HP-genes from this chapter could be validated using this data, considering the experimental differences highlighted in Section 3.4.2.
5. Gene Set Enrichment Analysis. Toppgenes is a comprehensive resource for exploring which sets of genes are enriched in genes determined from this study. One important step would be to enrichment in cell type gene sets to try to narrow down the most likely individual cells with expression profiles that align with the highly expressed genes. This could inform further research using single-cell RNA-Seq data.

6. Disease relevancy of the Identified Genes. DisGeNet (Pinero et al., 2020) is a large public collection of genes and variants associate with human diseases. Utilities are available to perform enrichment analyses to identify which diseases are more likely to be associated with a list of genes. In a brief test, CO- and HP-genes in this chapter appeared to be strongly associated with the curated database in DisGeNet. Further work would be needed to explore this resource in more detail.
7. Open Targets Overlap. During the overlap analysis (3.3.6), the overlap of genes associated with other diseases were briefly looked at. The common CO/HP-genes overlapped with 27 genes in the stringent SCZ set (OAS>0.75), but also in the following sets (OAS>0.75): 20 genes in coronary heart disease, 19 in epilepsy, and 17 in migraine disorder. Future work could explorer the use of the CO- and HP-genes in multiple diseases beyond SCZ.
8. Four Brain Regions. The strategy of using more than one control region aimed to eliminate genes that would be found in brain regions not involved in SCZ. Pathogenic rare variants affecting genes in non-SCZ-relevant brain regions would be expected to produce symptoms in those brain regions as strongly as in SCZ-relevant regions. Future work using single-cell or single-nucleus RNA-Seq could expand this to produce a wider spectrum of potential targets and contrast regions. This could narrow down further the number of genes.
9. Comparison with GWAS genes. This chapter looked at the overlap of genes with SCZ-associated genes in the Open Targets platform. This could be extended to look at the overlap of genes with risk loci found to associated with SCZ in the recent GWAS (Trubetskoy et al., 2022) from the Schizophrenia Work Group of the Psychiatric Genomics Consortium (Trubetskoy et al., 2022)

4 Variant Filtering of SCZ Exome data

4.1 Overview

In the previous chapter, 1,076 genes were identified as demonstrating higher expression in either or both the Cortex or Hippocampus brain regions compared to Cerebellum and Substantia Nigra, selected as control regions. Based on the premise of the current study, these genes are hypothesised to include genes that are relevant to SCZ in the Cortex and/or Hippocampus and the list will act as a filter to focus the prioritisation of rare variants identified from a cohort of patient exome data. Variants will be aggregated into the biological pathways in which the genes play a part.

This chapter describes a set of methods to filter variants from whole exome studies in a small group of SCZ patients and aggregate rare variants within the CO- and HP-genes into biological pathways. The methods were run across multiple groups of cases and groups of controls in Chapter 5, where the focus was on the differences in the outputs between patients with SCZ and controls.

4.2 Background

4.2.1 Genomic Variants

Genomic variants are defined as differences between an individual's genome and the Human Reference Genome. Characterised by frequency (rare or common), type (SNV, indel, structural change), consequence (silent, missense, nonsense, or splice site), and pathogenicity (benign, damaging), they play a crucial role in shaping individual differences and disease susceptibility.

Rare variants (RV) occur with an allele frequency of less than 1% in a population, while ultra-rare variants fall below 0.1%. Singletons and doubletons are variants that are only found in one or two individuals, respectively.

Single nucleotide variants (SNVs) represent the most common variations, with millions present in a typical genome (1000 Genomes Project Consortium et al., 2015; Karczewski et al., 2020; Lek et al., 2016). Multinucleotide variants (MNVs) combine neighbouring SNVs. Small insertions and deletions (indels) affect stretches of nucleotides. Copy number variants (CNVs) encompass larger deletions, insertions, or repeats. Other structural changes include inversions and translocations.

Variant consequence describes the effect of the DNA change on gene transcription and translation. Variants can influence pre-mRNA transcription by altering regulatory regions (5' UTR variant, 3' UTR variants) and start and stop sites (start gain, start lost, stop lost, stop gained), pre-mRNA splicing (splice region variant, splice donor/acceptor variants), protein coding sequence (missense, frameshift), and these may further influence protein folding, protein complex formation, and protein interactions, including ligand-, metal- and co-enzyme-binding.

4.2.2 Genome-Wide Association Studies in Schizophrenia

Genome-wide association studies (GWAS) have identified increasing hundreds of common SNPs associated with SCZ, primarily with weak effects. While valuable for understanding the polygenic nature of SCZ (Neale & Sklar, 2015), these common variants explain only a fraction of the heritability (Trubetskoy et al., 2022). Rare and ultra-rare variants (RV/URV) are suspected to play a significant role in the "missing heritability," but their low frequencies render them undetectable by conventional GWAS (Cirulli & Goldstein, 2010; Dickson et al., 2010).

Targeted sequencing approaches, as well as whole-exome sequencing (WES) or whole-genome sequencing (WGS), offer a promising avenue for uncovering the contribution of RVs/URVs to SCZ susceptibility at these loci.

4.2.3 Whole Exome Sequencing of SCZ cases

While large-scale WES projects like the Swedish Exomes (Genovese et al., 2016; Purcell et al., 2014) and UK10K (UK10K Consortium et al., 2015) identified thousands of rare variants in SCZ patients, they fell short of uncovering any definitive genome-wide significant loci associated with the disease. This lack of robust findings despite impressive cohort sizes (4,877 cases (with 6,242 controls) in Swedish Exomes and ~1,745 in UK10K) highlighted the limitations of using existing reference databases like 1000 Genomes (1000 Genomes Project Consortium et al., 2015) and the Exome Aggregation Consortium (ExAC) (Lek et al., 2016) databases to define "rare." Variants considered rare within these studies might still occur in many individuals within the wider population, potentially masking their lower specificity to SCZ.

To overcome this limitation, efforts were made to combine data from these studies to increase statistical power and identify potentially missed rare variants. A follow-up

study (Singh et al., 2016) analysed a combined dataset of 4,264 SCZ cases and 9,343 controls from both UK10K and phase 1 of the Swedish Exomes (2,519 cases). While this larger analysis remained inconclusive for per-gene burden of disruptive variants, it did yield interesting findings when combined with *de novo* mutations from 1,077 probands from several other studies and focussing on loss-of-function (LoF) mutations with minor allele frequency (MAF) < 0.1%. Interestingly, rare mutations in the protein-coding gene, SET Domain Containing 1A, Histone Lysine Methyltransferase (SETD1A), recently linked to dendritic complexity in human glutamatergic neuronal iPSCs (Wang, S. et al., 2022), were found in 10 cases (3 de novo and 7 in the combined exomes) and none in controls. This finding, despite it not reaching genome-wide significance, suggested the potential role of rare *de novo* LoF mutations in SCZ susceptibility, even if their frequency is extremely low.

These efforts, despite not delivering immediate breakthroughs, significantly advanced our understanding of the potential contribution of rare variants in SCZ. They revealed the limitations of existing reference databases in defining "rare" and highlighted the need for larger, more diverse datasets, like gnomAD, which offered a larger dataset to identify rare variants more accurately.

4.2.4 Reference Population Databases

As reviewed in Early WES projects aimed at identifying rare variants associated with schizophrenia (SCZ) faced a constant challenge: the evolving definition of "rare". What appeared rare within a smaller study often turned out to be more prevalent as reference databases expanded. This limitation was evident in projects like Swedish Exomes, where allele frequencies in the first phase were calculated using 2,504 individuals from 1000 Genomes (Purcell et al., 2014) and later reassessed against 45,376 individuals from ExAC (Genovese et al., 2016).

The arrival of gnomAD marked a significant leap forward. With its curated data from 125,748 exomes and 15,708 genomes in its 2.1.1 release (Karczewski et al., 2020), gnomAD offered a vastly expanded pool of genetic information. This larger scale significantly altered the landscape of rare variant research. Variants initially considered rare within smaller studies and potentially linked to SCZ could be reclassified against gnomAD's latest datasets. This reclassification process encourages the re-evaluation of past findings. This ongoing process of refining our understanding

of “rare” through expanding reference populations is crucial for understanding the complex genetic architecture of diseases like SCZ. The gnomAD Exomes covers 17.2 million variants, which the authors estimated represents only 11.5% and 3.5% of all the synonymous and stop-gained mutations that are possible in the exome (Karczewski et al., 2020). Further, releases of gnomAD since have increased the number of individuals fivefold. Although there is low representation of many ancestries in the current reference populations, it is becoming possible to prioritise rare variants from smaller numbers of sequenced exomes in a diagnostic setting. Furthermore, the creation of local biobanking initiatives, for instance the UK Biobank (Bycroft et al., 2018) with over 400,000 deeply genotyped and phenotyped individuals is enabling reassessments of previous GWAS and WES results across many phenotypes (Sun et al., 2022). The most recent gnomAD release (v4; preprint) combines data from the UK BioBank and versions 2 and 3 gnomAD into a single cohort of approximately 800,000 individuals using GRCh38 coordinates (Koenig et al., 2024).

4.2.5 Pathogenicity / Variant Effects

After filtering common variants , the next important step in variant analysis involves predicting the potential coding consequences and the effect of any coding changes on protein function and structure. This intricate task relies on a diverse range of databases and software tools, each equipped to assess how individual variants affect different known transcripts for each gene (Cingolani et al., 2012; Danecek & McCarthy, 2017; McLaren et al., 2016; Wang, K. et al., 2010).

These tools employ a variety of algorithms and resources to evaluate the biological effects of discovered variants and annotate them with informative labels. For example, ANNOVAR (‘ANNOtation of VARiants’) (Wang, K. et al., 2010) and Ensembl’s Variant Effect Predictor (VEP) (McLaren et al., 2016) serve as comprehensive toolboxes, drawing data from multiple databases to paint a detailed picture of each variant. They utilise predictions from coding sequence effects, conserved regions of the genome, GWAS hits, inclusion in functional predictions and annotations in databases, for example the database for non-synonymous function prediction (dbNSFP) (Liu et al., 2016), ExAC, 1000 Genomes, and use of candidate gene lists for Mendelian diseases to enrich their annotations (Wang, K. et al., 2010). These also include tools like SnpEff (Cingolani et al., 2012), SIFT (Sorting Intolerant from Tolerant)

(Ng, P. C. & Henikoff, 2001) and PolyPhen-2 (Polymorphism Phenotyping v2) (Adzhubei, I. A. et al., 2010), which focus on predicting the impact of amino acid changes on protein function by analysing the physicochemical properties of the modified residues.

A common limitation of many tools lies in their tendency to assess variants in isolation, overlooking the potential interplay between neighbouring variants that might collectively affect the same codon (Wakeling et al., 2019). These multinucleotide variants (MNV) necessitate a different approach, one that recognises their combined effect. The BCFtools 'csq' function (Danecek & McCarthy, 2017) constructs transcripts from a gene annotation model and the human reference genome to identify the potential coding consequence of such MNVs – even across exon-exon boundaries. This approach opens the door to uncovering the combined effects of rare variants and/or common variants, potentially revealing hidden intricacies in the complex architecture of SCZ.

The above steps align with the updated ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2024 (Durkie et al., 2024). These guidelines provide a framework for classifying and reporting the pathogenicity of rare variants along with the strength of the evidence. Variants are categorized into five groups: 'pathogenic', 'likely pathogenic', 'variants of uncertain significance (VUS)', 'likely benign', and 'benign' using terms to weight the evidence, such as 'very strong', 'strong', 'moderate', and 'supporting'. Accurate classification of variants is critical in clinical settings and must consider various types of evidence, including population frequency data from resources like gnomAD, computational predictions, functional studies, and familial segregation data where variants are observed in affected family members. The ACGS Best Practices 2024 also updates how combinations of evidence can and cannot be used in combination with different variant types and details whether specific variant classifications should be reported or not.

In a purely bioinformatics pipeline using a cohort of case and control 30X WES data, without access to detailed patient histories, family segregation data, the ACGS Best Practices cannot be adhered to completely for accurate classification of the pathogenicity of rare variants, although it is recommended to use up-to-date annotation and stringent filtering for variants not related to the phenotype (Ellard et

al., 2024). Filtering common variants found in gnomAD Exomes roughly aligns to guideline PM2 ('Pathogenic Moderate 2') 'Rare in population studies' (the most up to date version at the time of writing being gnomAD 4.0). Determining coding consequences and filtering to truncated or stop loss variants allows a downstream assessment of PVS1 (pathogenic, very strong) 'Null variant'. The use of *in silico* prediction tools only provides 'supporting' evidence according to the guidelines. While adhering to these guidelines is essential in a clinical setting where a correct diagnosis is critical for a patient, in a bioinformatics pipeline the ACGS guidelines serve as a valuable framework for standardising variant analysis and ensuring consistency. However, the focus shifts towards identifying common biological pathways and mechanisms, rather than providing direct clinical interpretations or patient-specific diagnoses.

4.2.6 Variant Aggregation in Pathways – systems biology

While deciphering the individual effects of rare variants in polygenic diseases like SCZ holds vast potential, it also encounters a functional limitation. Many of these variants are extremely rare, potentially impacting only a single patient or handful of individuals. Within the framework of the "common disease, rare variant" model, these variants are rarely found across unrelated patients, often remaining confined to isolated cases (McClellan et al., 2007). To navigate this challenge, researchers have turned to a powerful strategy: aggregating rare variants based on shared characteristics.

This aggregation approach transcends studying individual variants and instead focuses on the emergent picture formed by clustering them. By grouping them according to shared properties like gene affiliation, predicted pathogenicity, or involvement in specific biological pathways, researchers can begin to discern larger patterns within the complex architecture of SCZ. Such insights may be invaluable in illuminating specific characteristics of the underlying biology of SCZ as well as narrowing down which variants may become candidates for further investigation.

As well as determining the individual variant effects, an important step lies in mapping these effects onto the functional pathways in which the genes operate. Over the past two decades, a rich collection of curated resources has emerged to facilitate this. Manually curated pathway databases like GO, KEGG, and Reactome, alongside enrichment tools such as PANTHER, DAVID, and R packages like Clusterprofiler, GOseq,

and EnrichR empower researchers to link genes to relevant biological pathways (Pavlopoulos et al., 2015). PANTHER (Protein Analysis Through Evolutionary Relationships), for instance, offers rapid gene annotation via gene ontology (GO) terms (Gene Ontology Consortium, 2021), gene families or protein classes, and biological pathways curated by PANTHER or Reactome (Gillespie et al., 2022; Mi et al., 2021). Furthermore, novel data mining techniques are being leveraged to connect genes to scientific literature, as exemplified by the Open Targets Platform (Ochoa et al., 2021) and STRING (Szklarczyk et al., 2021). Notably, while manually curated resources like PANTHER and Reactome offer high reliability, they may not be as comprehensive as computationally derived datasets like GO Complete.

4.2.7 Aims and Objectives

Building on the identification of CO/HP-genes in Chapter 3, this chapter describes a variant aggregation analysis focussed exclusively on the identification of rare, damaging variants in the CO/HP-genes. One method of aggregation is grouping all rare variants within small samples of patients from the larger cohort. A sample of 100 exomes was extracted from the Swedish Exomes project. This approach ultimately aims to unveil potential groupings of rare variants with substantial impact on similar biological processes highly relevant to brain regions implicated in SCZ consistently across multiple groups.

It is important to note that CO/HP refers to the GTEx bulk RNA-Seq samples 'Frontal Cortex' and 'Hippocampus' where the genes were identified as highly expressed; and the sample size of 100 exomes was chosen to strike a balance between capturing enough genetic variation within the CO/HP-genes and maintaining a manageable group size for in-depth analysis in subsequent chapters.

Main aim:

- To detail the methods and outputs that lead to the identification of a set of biological pathways and biological processes enriched with variants within the CO/HP-genes within a small sample of patients.

The objectives are:

1. Randomly select 100 SCZ samples from the Swedish Exomes project and retain only variants in the CO/HP-genes.
2. Analyse the coding consequences of all variants, identify potential multinucleotide variants (MNVs), and quantify the types of coding consequences across the 100 samples.
3. Annotate variants with allele frequencies, filter for rare variants (MAF<0.01 and gnomAD AC <= 5), and categorise them (for example, SNVs, indels, singletons, not-in-gnomAD).
4. Use VEP annotation to predict functional impact (missense or worse) and categorise variants based on pathogenicity scores (deleterious or pathogenic).
5. Aggregate variants into genes and subregions (20kb, 50kb, 100kb) and prioritise genes/subregions with high counts of rare variants.
6. Aggregate variants into pathways and perform gene enrichment analysis using GO, PANTHER, and Reactome databases to identify enriched biological processes.

The final list of biological processes and molecular functions provided an initial indication of potential Cortex- and Hippocampus-relevant pathways enriched for RVs among these 100 SCZ patients. This analysis lays the groundwork for further investigation into the prevalence and impact of these variants between multiple case and control groups of 100.

4.3 Methods

4.3.1 Sample Population and Data

In the following sections the Swedish Exomes data and gnomAD data were downloaded and the CO/HP-gene regions were converted from GRCh38 to GRCh37. These were performed once only for this chapter and were also used in chapter 5.

4.3.1.1 Swedish Exomes

The Swedish SCZ Population-based case control Exome Sequencing ('Swedish Exomes') Project (dbGaP accession: phs000473.v2.p2) recruited 4,877 unrelated SCZ patients, 6,242 case-matched controls and 1,144 other psychiatric illnesses (mainly bipolar disorder) all with Swedish parents only (Genovese et al., 2016; Purcell et al., 2014).

4.3.1.1.1 Swedish Exomes – Phenotype and Metadata files

Phenotype and Metadata were obtained from dbGaP with permission granted for general research use (GRU). Subject phenotypes data (phs000473.v2.pht002600.v2.p2.c1. Schizophrenia_Swedish_Subject_Phenotypes.GRU.txt.gz, referred to as Phenotypes.txt.gz) and Sample data (phs000473.v2.pht002599.v2.p2.Schizophrenia_Swedish_Sample.MULTI.txt.gz, referred to as Sample.txt.gz) were downloaded from dbGaP via the NCBI's Sequence Run Archive (SRA) dbGaP File Selector (Leinonen et al., 2011; NCBI, 2022). The SRA File Selector produced a single-use Unix command (below) for use with IBM Aspera Connect (IBM, 2023) to download the files to the registered dbGaP folder (registered in Section 3.2.1.1). A separate text file containing the dbGaP credentials was required in the same folder to authenticate the connection. (This link must be re-generated by the SRA File Selector for each new session.)

```
"%ASPERA_CONNECT_DIR%\bin\ascp" -QTr -l 300M -k 1 -i
"%ASPERA_CONNECT_DIR%\etc\asperaweb_id_dsa.openssh" -
W
AF077DA2247C64B5D8A33DA4094BA0641CFBD9B50E5CD4B748894
387473A97FE2F323067CEBA4486389551C1D7B4E4FB7D
dbtest@gap-
upload.ncbi.nlm.nih.gov:data/instant/<dbgap-
username>/73135 .
```

All downloaded encrypted files were decrypted using '`vdb-decrypt <download directory>`' in sra-toolkit (NCBI, 2019) and then moved to the project directory.

The phenotypes data (Phenotypes.txt.gz) includes a variable that shows whether the sample exceeded 20X coverage at 80% of sites on average. Samples that did not meet this were filtered out ('`Coverage_Pass=="N"`').

4.3.1.1.2 Swedish Exomes – Variant call data

The Swedish Exome VCF data was selected in the NIH SRA Run Selector. The following command was generated to run on the local server:

```
"%ASPERA_CONNECT_DIR%\bin\ascp" -QTr -l 300M -k 1 -i
"%ASPERA_CONNECT_DIR%\etc\asperaweb_id_dsa.openssh" -
W
A2B585439FD154AC3A19BC2D4E77B6F23D882BED11F4C4D4B8F19
55BD843BACD15EA667BD84CF4014A870273014CDD55FD
dbtest@gap-
upload.ncbi.nlm.nih.gov:data/instant/oslay1/73214 .
```

The downloaded, encrypted tarball (`phg000773.v1.SchizophreniaSwedish_Sklar_v2.genotype-calls-vcf.WES_markerset_grc37.c1.GRU.tar.gz.ncb_enc`) file size was 65GB. It was decrypted using '`vdb-decrypt`' in sra-toolkit. The Exomes VCF file (`swe.vcf.gz`) and its index file (`swe.vcf.gz.tbi`) were unpacked from the decrypted tarball using '`tar xvfz <file>`'. A vcf.gz file of this size and format is very slow to query; therefore, the file was converted to binary VCF (BCF) format for a significant improvement in query time. To convert to binary format, the following command was run followed by the generation of a new TBI index file:

```
nohup bcftools view --threads 25 ../phs000473-
SwedishExome/files/Genotype/sub/sub20160809/swe.vcf.g
z -Ob -o res/swe.bcf > res/conv-swe.log 2> res/conv-
swe.err &

bcftools index –threads 25 -c -m 9 swe.bcf
```

4.3.1.2 CO- and HP-genes

The combined list of 1,076 CO- and HP-genes (CO/HP-genes) discovered in Chapter 3 were used to filter the variants in the Swedish Exomes VCF (in 4.3.2) and to aggregate rare variants in the pathway analysis (4.3.8). In Chapter 3, GTEx reads were mapped to the loci associated with the genes in GRCh38 coordinates. Swedish Exome data was in GRCh37 coordinates, therefore, the coordinates of the GRCh38 genes were lifted over to the corresponding regions in GRCh37. The list of genes was loaded into a table in R and annotated with the chromosome, start and end positions, name, and description for each gene in GRCh38 coordinates from Ensembl BioMart using biomaRt (v2.44.4; Durinck et al., 2009) in R with Ensembl release 97 (July 2019) and saved as `merged_regions_38.txt`. The GRCh38 genomic coordinates were lifted over in

Ensembl's Assembly Converter tool (EMBL-EBI, n.d.). The mapping "GRCh38 -> GRCh37" was selected and `merged_regions_38.txt` was selected for conversion. The results were downloaded as a tab-delimited file in a browser-extensible format (BED) file and renamed to `merged.37.bed`. The file was sorted by chromosome and position using the following Unix bash *sort* command and saved as `merged.sort.37.bed`.

```
sort -V -k 1,1 -k2,2n ../data/merged.37.bed > ../data/merged.sort.37.bed
```

One duplicated gene (AHRR) with two Ensembl ids was removed in this process. Another gene, CR589904.2 could not be resolved to GRCh37 coordinates. `merged.sort.37.bed` contained 1074 genes that had lifted over to GRCh37.

4.3.1.3 gnomAD Exomes

The Genome Aggregation Database (gnomAD) is a reference population database of curated variants from 125,748 exomes and 15,708 genomes (Karczewski et al., 2020). This chapter used gnomAD exome data v 2.1.1 which contains about 17.2 million variants representing about 11.5% and 3.5% of all the synonymous and stop-gained possible variants in the human exome. In this chapter, groups of 100 SCZ cases were extracted from the Swedish Exomes, making it difficult to assess the rarity of the variants in the wider population. The population frequency of all variants within the CO/HP-gene regions of the Swedish Exomes were assessed against the latest Exomes dataset (v2.1.1) from gnomAD.

The gnomAD v2.1.1 Exomes VCF was downloaded from gnomAD's Google Cloud Storage repository (https://storage.googleapis.com/gnomad-public/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.vcf.bgz) along with the index (tbi) file (https://storage.googleapis.com/gnomad-public/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.vcf.bgz.tbi). The file is 58.8 GiB and is very slow to query in the bgzip-compressed format, as supplied. Converting to the binary VCF format (BCF) produced very significant performance gains when querying the file. The file was converted using `bcftools view` and re-indexed using `bcftools index` – as used for the Swedish Exomes VCF (see 4.3.1.1.2 above).

4.3.1.4 Human Reference Genome and Genome Annotation

In the Swedish Exomes study, exomes were analysed using Human Reference Genome GRCh37 primary assemblies with decoy sequences (hereafter hs37d5). This was obtained from GENCODE (Frankish et al., 2019)

Coding consequences (4.3.3) leveraged gene annotation data from GENCODE v31 (lift 37) (Jun 2019). The main analysis and all subsequent downstream analyses in this chapter used GENCODE v31 and corresponding data from Ensembl Release 97.

4.3.2 Subsampling of 100 subjects

100 subject IDs were randomly selected from the phenotype and samples data using an R script [Chapter4\01-SwedishExomes-DataSelection.R (Slay, 2021)]. The Phenotypes file and the Samples file (Samples.txt.gz) were opened in R, decompressed in the script using the R function `gzfile()`, and read into respective tables. A new column was added to the phenotypes table to contain the group number, which was composed of the subject group (SCZ) and a three-digit number (for example, SCZ000 as group 0). 100 subjects with `primary_disease="Schizophrenia"` were randomly assigned to the SCZ000 group using the R function, `sample()`.

34 more groups of 100 SCZ patients and 48 groups of 100 controls (labelled with the group number in the format 'CTL000' where `primary_disease="Control"`) for Chapter 5 were processed within the same script. The rest of this chapter describes the processing of the single group (SCZ000).

The phenotypes table linked a subject to their phenotype data and the Samples table links the subject to their sample data (as a sample ID). The phenotypes table was joined to the samples table by subject ID (`subjid`). The 100 sample IDs for the subjects in SCZ000 were recorded to a text file (`SCZ000.sampid.args`) for use in the next step.

Variants for the 100 SCZ subjects were extracted from the Swedish Exomes VCF file (converted to BCF in 4.3.1.1.2) using the `view` command from BCFtools v1.11 (Danecek & McCarthy, 2017) [Chapter4\02-SwedishExomes-SplitVCFtoGroups.sh (Slay, 2021)]. `bcftools view` was run on Swedish Exomes file `swe.bcf` to filter all variants in samples with IDs in `SCZ000.sampid.args`, further filtered to the genomic regions in `merged.sort.37.bed` (see 4.3.1.2) to create `SCZ000.bcf`

and a CSI index `SCZ000.bcf.csi`. The file was further processed to remove all variants where all sample genotypes were the reference. The final BCF file contained all variants in the CO/HP-genes from each of the 100 subjects in group SCZ000. The BCFtools `stats` command and `smpl-stats` plugin were called to examine the extracted variants overall and per sample.

4.3.3 Annotation of variants using BCFtools Consequences (`csq`)

The assessment of variant consequences on the coding sequence involved annotating all variants in the SCZ000 BCF file using the Consequence (`csq`) function from BCFtools v1.11 to determine if the effects of neighbouring variants combined. BCFtools `csq`, a haplotype-aware consequence predictor, was used to identify multinucleotide variants (MNV). An MNV can cause a different change in amino acid to the individual SNVs, and can detect instances, for example, where a variant in one position causes a frameshift in the coding sequence and neighbouring variant (even across exons) restores it. This approach is essential, as conventional software operating site-by-site may overlook these interactions between variants (Wakeling et al., 2019).

BCFtools `csq` was run [Chapter4/03-VariantConsequence.sh (Slay, 2021)] against the group BCF file (`SCZ000.bcf`), supplied with the reference genome, hs37d5, and genome annotation GENCODE v31 (see 4.3.1.4). Because the Swedish Exomes VCF is unphased, which is ideally required for true haplotype-aware consequence prediction, the `-phase` option for handling unphased heterozygous genotypes was set to take the genotypes (GT) as is and create haplotypes regardless of the genotype phase (0/1 -> 0/0; option: `-phase = a`). The annotations were added to a new field, named 'INFO/BCSQ'. The annotated output binary file (`SCZ000.csq.bcf`) was subsequently indexed. Finally, the worst consequence for any variant (in any transcript of the gene) with a consequence of missense or worse (option: `-s worst:missense+`) was extracted from the BCSQ field (option: `-a BCSQ`) to a tab-delimited text file (`SCZ000.worst.missense.txt`) using the BCFtools plugin `+split-vep`. This function was supplied with a format string specifying which fields to extract and tab delimiters:

`%CHROM\t%POS\t%REF\t%ALT\t%gene\t%strand\t%biotype\t%amino_acid_change\t%dna_change\t%Consequence\n`

`SCZ000.worst.missense.txt` was examined for MNVs, which may be missed when aggregating variants based on counts of genotypes.

4.3.4 Annotation of variants using gnomAD Exomes v2.1.1

Annotation of the patient exome data with allele frequencies from a large, healthy population in gnomAD Exomes v2.1.1 (see 4.3.1.3 above) helps to identify variants with allele frequencies as low as 1 in 128,000. To explore the range of allele frequencies, annotating all variants with allele counts and frequencies was preferred over filtering out common variants. All variants from the 100 samples in the VCF file were annotated with a set of metrics (Table 4.1) from gnomAD Exomes using BCFtools `annotate` function [Chapter4/04-gnomadAnnotation.sh (Slay, 2021)]. This function was supplied with the gnomAD binary VCF file (see 4.3.1.3 above), the list of metrics (Table 4.1) to extract for each allele in gnomAD Exomes, and the `SCZ000.csq.bcf` output from the previous step. Records were annotated where the position of the variant was the same in both datasets (option: `-collapse all`) and output to a new binary VCF file, `SCZ000.gnomad.bcf`. The output file was indexed.

Table 4.1 gnomAD Metrics used for annotating variants in each group

Population	**Allele Count (AC)**	**Allele Number (AN)**	**Allele Frequency (AF)**	**Homozygous ALTs (nhomalt)**
Popmax	popmax_AC	popmax_AN	popmax_AF	popmax_nhomalt
Global	gnomAD_AC (AC)	gnomAD_AN (AN)	gnomAD_AF (AF)	gnomAD_nhomalt (nhomalt)
Non-Finnish European	AC_nfe	AN_nfe	AF_nfe	nhomalt_nfe
NFE Controls only	controls_AC_nfe	controls_AN_nfe	controls_AF_nfe	controls_nhomalt_nfe
Swedish (subgroup of NFE)	AC_nfe_swe	AN_nfe_swe	AF_nfe_swe	nhomalt_nfe_swe
Swedish Controls only	controls_AC_nfe_swe	controls_AN_nfe_swe	controls_AF_nfe_swe	controls_nhomalt_nfe_swe
Other fields: popmax, rf_tp_probability, vep				

Sample-level statistics relating to genotype and variant type were extracted from the VCF file using the BCFtools plugin `+smpl-stats`. Statistics were extracted for all variants that do not appear in gnomAD or those that appear in gnomAD with an ALT allele count of 5 or less (option: `-i'GT="alt" && (gnomAD_AC<5 || gnomAD_AC=".")'`) and stored as `SCZ000.smpl.stats`. Simultaneously, statistics for all variants in the annotated BCF file were extracted using `+smpl-stats` and stored as `SCZ000.all.smpl.stats`. The total number of variants per sample, number of samples heterozygous or homozygous for the alternate allele and singletons were extracted and analysed together in box plots in R.

4.3.5 Normalisation of variants

Variants in the VCF file were normalised using BCFtools `norm`. This realigns variants and splits multiallelic sites from a single line in the file into biallelic sites. For example,

if a position has more than two ALT alleles that are found in the genotypes of different samples (for example 0/1, 0/2, 1/2, 2/2), the alleles are split across multiple lines. This ensures that each line represents a single genotype. The BCFtools `norm` function [Chapter4/05-VariantRare.sh (Slay, 2021)] was run on the gnomAD-annotated `SCZ000.gnomad.bcf` file with the multiallelic option set to split all multiallelic sites into true biallelic sites (option: `-m -`).

4.3.6 Filtering to rare variants

The results of the normalisation step were piped into BCFtools `view` command with a filter set to include only rare variants (allele frequency in Swedish Exomes < 1%) that had a near zero allele count in gnomAD (gnomAD allele count < 5). The filter also included sites where 99% or more of gnomAD Exomes had the ALT allele and all but between 1 and 5 samples in Swedish Exomes had the REF allele. In the latter case, the reference allele was the rare variant. The `view` command outputs a final VCF file (`SCZ000.rare.bcf`). The BCFtools `stats` command and `smpl-stats` plugin were called to examine the extracted rare variants overall and per sample.

4.3.7 Pathogenicity of Remaining Variants – Ensembl Variant Effects Predictor (VEP)

The VCF file with rare variants (`SCZ000.rare.bcf`) was annotated with the results of Ensembl's Variant Effect Predictor (VEP) (McLaren et al., 2016). Using VEP in a cached and offline mode eliminated network connections and allowed larger numbers of queries to be performed in faster time without exposing private data to a public database. Ensembl VEP Cache version 101 virtual machine was downloaded from the VEP website and run offline. VEP was run against `SCZ000.rare.bcf` from the previous step with cache and offline options (`--vcf --cache --offline`). The `--dir_cache` option was set to the vep_cache folder containing the cached databases created during VEP setup. This produced a final `SCZ000.vep.vcf` file. The file was small enough to output in plain text rather than binary VCF format for ease of use.

Both BCFtools csq and VEP provided SIFT and PolyPhen-2 scores against gene transcripts within the annotation ('INFO/BCSQ' for csq and 'INFO/vep' for VEP). Variants with a high PolyPhen-2 score (0.85-1.0) cause an amino acid change that has been more confidently predicted to lead to a damaging impact on protein function or

structure. Variants with a SIFT score below 0.05 lead to an amino acid change that has been predicted to lead to a deleterious change in protein function. The scores were extracted using the BCFtools `+split-vep` function with the following format string (to include genomic position, allele counts, allele numbers and allele frequencies in both the Swedish Exomes dataset and from gnomAD, the gene affected, the consequence calculated and scores for SIFT and PolyPhen):

```
%CHROM\t%POS\t%AC\t%AN\t%gnomAD_AC\t%gnomAD_AN\t%gnomAD_A
F
\t%gnomAD_nhomalt\t%SYMBOL\t%ID\t%Consequence\t%SIFT\t%PolyPhen\
n
```

Variants were filtered to those with a consequence of 'missense or worse' and with a SIFT value of 'deleterious' or a PolyPhen value of 'damaging' (>= 0.99). The results were piped into an `awk` command to filter allele frequencies in gnomAD of less than 0.01. The results were appended to a tab-delimited text file, `SCZ000.worst.missense.polyphen-sift.txt`. This was sorted and duplicate variants were removed.

4.3.8 Binning variants to Genes and Pathways

To analyse whether rare variants occur in specific genes or specific regions within a gene, rare variants, both SNPs and INDELs, were aggregated into bins for each gene and into subregions within the gene of sizes 10 kilobases, 20 Kb and 100 Kb. BEDtools `annotate` function was used to annotate a copy of the `merged.sort.37.bed` regions file with the count of variants in each region from the `SCZ000.vep.vcf` file [Chapter4/07-regions-10k-bins.sh (Slay, 2021)]. From this the gene symbol or regions and count were extracted to a tab-separated file. Four files were created – one for counts (SNPs, indels and total) per gene (`SCZ000.gene.txt`), and one each for bin regions of 10 Kb (`SCZ000.10kanno.txt`), 20 Kb (`SCZ000.20kanno.txt`) and 100 Kb (`SCZ000.100kanno.txt`).

This list of genes with variant counts was analysed for overrepresentation in PANTHER (Mi et al., 2021) following their updated protocol (v14.0) (Mi et al., 2019). The statistical overrepresentation test determined whether genes with variants were over- or underrepresented compared to the proportion of all human reference genes

assigned to any biological process, molecular function or pathway tested. For example, in the human genome reference list (20,595 genes) there were 5 genes (0.0243%) in the GO Biological Process 'trigeminal nerve structural organisation (GO:0021637)'. The expected number of genes with variants in the CO/HP-genes would be 0.11 (0.0243% of 444). The observed number of genes was 3 indicating that the input gene list was 27.83-fold overrepresented for genes involved in trigeminal nerve structural organisation.

An overrepresentation binomial test was run against the PANTHER GO Biological Process Complete dataset, corrected for multiple testing using Benjamini-Hochberg false discovery rate (FDR) of 0.05. As a comparison, the same test was run on GO-Slim biological processes, GO molecular functions, PANTHER Pathways and Reactome Pathways, producing a single list per test.

4.4 Results

In this chapter, the Swedish Exomes project provided whole exome sequencing (WES) data from 12,380 individuals, including 4969 with SCZ, 1166 with BPD, and 6245 controls. Exomes with over 20X coverage in 80% of sequenced sites were selectively included. From the remaining 3582 SCZ samples, 100 subjects were randomly sampled. Variant sites were filtered to focus on the 1074 CO/HP-gene regions identified in Chapter 3 (as detailed in Section 4.3.1.2). Subsequently, variants were annotated for coding consequences, enabling the identification of multinucleotide variants that are often overlooked in variant analysis. Additionally, variants were annotated with allele frequencies from gnomAD, a large population study. Finally, rare variants (RVs) were grouped into gene regions and 10/20/100 kilobase regions. The resulting genes and RV counts underwent analysis through gene set enrichment analysis in PANTHER.

4.4.1 Filtering of Swedish Exome data to CO/HP-genes in 100 SCZ patients

In this section the Swedish Exomes data was filtered targeting the CO/HP-gene regions (as detailed in Section 4.3.1.2) in a subset of 100 patients with SCZ. The original Swedish Exomes VCF contained 1,811,206 variant sites in the 12,380 individuals, comprising both biallelic and multiallelic variants. Following filtering to the CO/HP-

gene regions and selection of 100 SCZ samples, the dataset was reduced to 12,084 sites. Of these, 11,033 were single nucleotide variants (SNVs), and 1,099 were INDELs (summarised in Figure 4.1).

Among the variant sites, 358 were multiallelic, with 30 being SNVs. Some multiallelic sites were intronic and were considered likely to be artifacts, potentially stemming from PCR or sequencing processes near exon extraction kit boundaries. 3,420 SNPs and 447 INDELs were identified as singletons within the 100 samples. Additionally, 4,238 SNPs and 541 INDELs had an allele frequency lower than 0.01.

At the time of the Swedish Exomes study, 8905 of the 11,033 SNPs were marked as known and 2128 were marked as novel, while 376 of the 1,099 INDELs were marked as known and 723 as novel. Approximately half (1,705) of the singletons were marked as known, with the remaining 1,715 marked as novel.

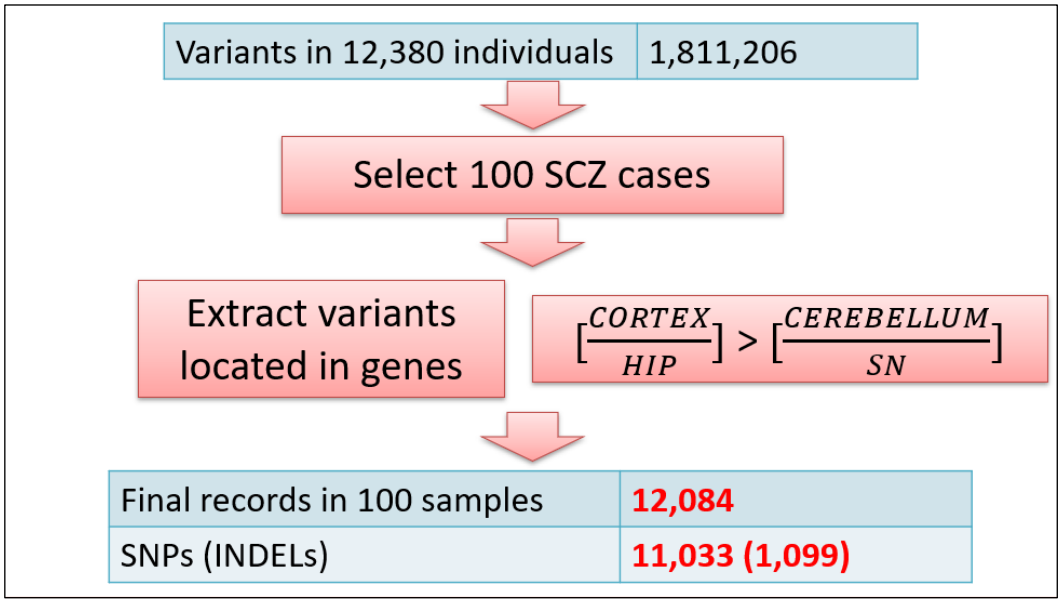

Figure 4.1 Workflow for selection of variants in the CO/HP-genes from 100 SCZ patients

Each SCZ sample, on average, had a recorded genotype at 11,770 (SD 188.6) of the 12,084 sites, with an average of 2,757 (SD 86.6) non-reference genotypes. This included 1,025 homozygous alternate alleles and 1,732 heterozygous genotypes. Samples, on average, carried 2,470 SNVs (Figure 4.2) and 287 INDELs (Figure 4.3). Approximately 38 of these variants were identified as singletons relative to the broader

Swedish Exomes population (within a range of 26-74) (Figure 4.4). An outlier sample was identified with 2,093 missing genotypes, having genotypes at 9,991 of the 12,084 sites, including 2,097 SNVs and 203 INDELs. This sample also exhibited the highest number of singletons (74). This sample passed the quality threshold therefore it was retained in the group of 100 SCZ samples. Without this outlier, the range of singletons narrows to 26-58.

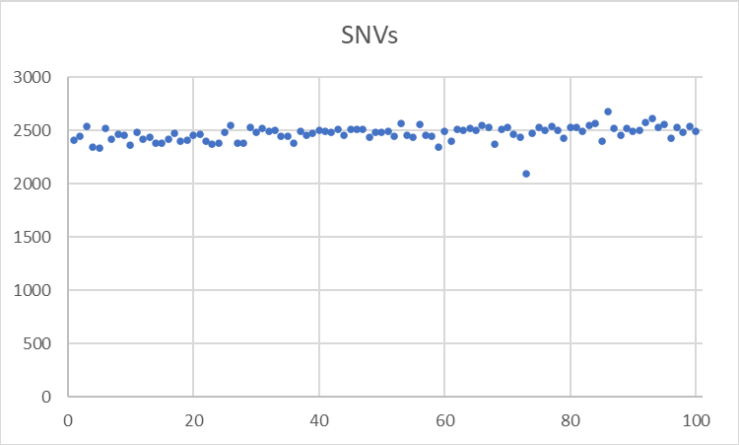

Figure 4.2 SNVs per sample in SCZ000

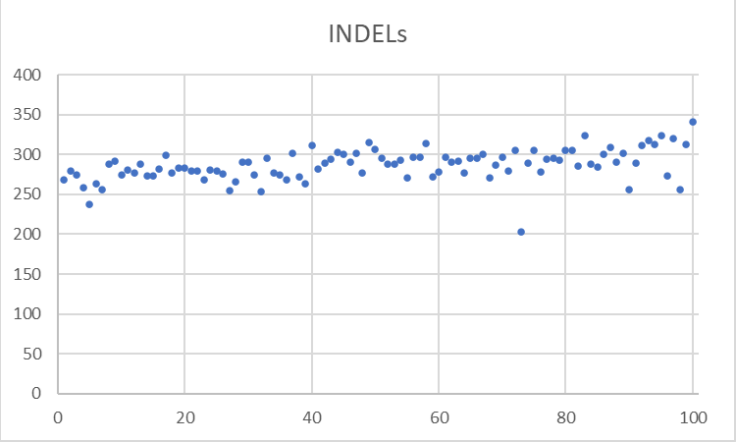

Figure 4.3 INDELs per sample in SCZ000

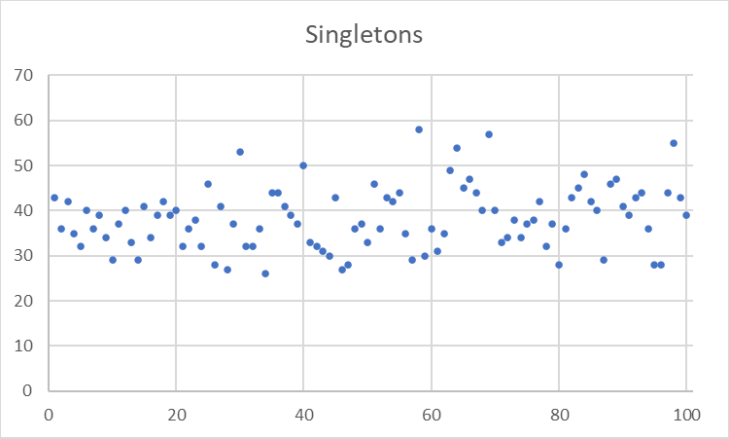

Figure 4.4 Singletons per sample in SCZ000

4.4.2 Annotation with Coding Consequences considering multinucleotide variants (MNVs)

In this section, the coding consequences of variants in the 100 SCZ samples were examined, with consideration of multinucleotide variants (MNVs). Using GENCODE gene annotation and the human reference genome (GRCh37), variants were annotated using BCFtools *csq* to explore potential interactions between neighbouring SNVs and SNVs separated by exons, which could collectively influence the coding sequence as an MNV.

Table 4.2 presents a breakdown of variant types with consequences of missense or worse from the 100 SCZ samples. While most variants were missense (3,126), including 28 potentially downstream from a gained stop codon, several other variant types were observed. These included 63 combined variants leading to a potential frameshift in the coding sequence, with 3 being downstream from gained a gained stop codon. There were 57 in-frame deletions and 31 in-frame insertions, including two potentially downstream from a gained stop codon. Genes with trinucleotide repeats, such as DLX6, ASCL1, FAM171B and POU6F2, exhibited variations in polyglutamine repeat lengths. Other variant types included 37 stop-gain, 3 stop-lost, 5 start-lost, and one start-lost/start-retained variant. Additionally, 35 splice site variants were identified, with 18 affecting the splice donor and 17 the splice acceptor.

Table 4.2 Breakdown of variant types with consequence of missense or worse from the 100 SCZ samples

Type	Count
missense	3126
frameshift	63
inframe_deletion	57
stop_gained	37
inframe_insertion	31
splice_donor	18
splice_acceptor	17
start_lost	5
stop_lost	3
synonymous&splice_acceptor	1
synonymous&frameshift&splice_donor	1
synonymous&frameshift&splice_acceptor	1
start_lost&start_retained	1
splice_acceptor&coding_sequence	1
missense&inframe_altering	1
frameshift&stop_retained	1

Exploration of MNVs was conducted, with acknowledgment of the limitations posed by the unphased nature of the Swedish Exomes data. A total of 116 variants were identified to be potentially involved in an MNV. Specific instances, such as variants in the genes exostosin-like glycosyltransferase 1 (EXTL1) on chromosome 1 and lipoprotein A (LPA) on chromosome 6, were highlighted. The two chromosome 1 variants in Table 4.3 in gene EXTL1 combined in 3 SCZ samples to produce a missense variant leading to an amino acid change of non-polar, aliphatic proline to the positively charged, hydrophilic histidine (p.P621H). The MNV in chromosome 6 in Table 4.4 in gene LPA combined in one sample to produce a novel missense variant that causes an amino acid change of proline to phenylalanine (p.P355F).

Table 4.3 Rare multinucleotide variant in EXTL1 in the 100 SCZ samples
Two rare variants in gene EXTL1 that interact in 3 samples. The consequence is listed against the first base indicating that the gene is on the forward strand.

Chrom	Pos	REF	ALT	AC	gnomAD AF	Gene	AA change	DNA change	Consequence
1	26361669	C	A	3	0.003129	EXTL1	621P>621H	CG>AT	missense
1	26361670	G	T	3	0.003125	EXTL1	.		

Table 4.4 Rare multinucleotide variants in LPA in the 100 SCZ samples
Two rare variants in gene LPA that interact in 1 sample. The consequence is listed against the second base pair indicating that the gene is on the reverse strand.

Chrom	Pos	REF	ALT	AC	gnomAD AF	Gene	AA change	DNA change	Consequence
6	161056166	G	A	1	0.001531	LPA	.	@161056167	
6	161056167	G	A	1	0.001535	LPA	355P>355F	GG>AA	missense

Table 4.5 lists the top 100 genes with the greatest number of variants with a consequence of missense or worse across all 100 SCZ samples. Each MNV is counted once, regardless of the number of variants that combine within the MNV.

Table 4.5 Top 100 genes ordered by counts of variants with worse consequences
Variants with a consequence of missense or worse were assigned to each gene. The top 100 genes were included.

Gene	Count	Gene	Count	Gene	Count	Gene	Count
GPR98	40	CCDC108	14	DRC1	11	CNGB1	9
DNAH7	33	CNTN5	14	EPHA10	11	DGKZ	9
FRAS1	29	SPATA31A6	14	FMN1	11	KIAA1217	9
DNAH11	26	SPEF2	14	OR2T29	11	MAPK15	9
FAT3	23	TTC40	14	PLXNA4	11	NRP1	9
DNAH3	21	ADAMTS8	13	SEMA5B	11	PKP2	9
FNDC1	20	GPR123	13	SLC12A8	11	RBBP8NL	9
LPA	20	KALRN	13	WDR65	11	SKA3	9
NLRP2	18	MAST4	13	ZBBX	11	SLC22A25	9
OTOGL	18	OR2T34	13	CCDC74B	10	SLC24A3	9
COL5A3	17	PLA2G4D	13	CD101	10	ABCA12	8
LMO7	17	WDR66	13	HS6ST1	10	ABCG5	8
LTK	17	CYP4A22	12	IGSF3	10	ADAD2	8
MYO16	17	DUSP27	12	KIAA1549	10	ANKRD24	8
ROS1	17	FHOD3	12	KLHL38	10	C4orf45	8
TLN2	17	PKD2L1	12	KRT84	10	C6orf118	8
EXO1	16	SEC16B	12	PLA2G4E	10	CCDC65	8
MSLNL	16	VWA3A	12	PPL	10	CCDC74A	8
TCHH	16	ABCC12	11	PTPRN2	10	CHL1	8
ZNF831	16	ACE	11	STK32C	10	CIDEA	8
CSMD1	15	B4GALNT3	11	THEG	10	COL5A2	8
MYH13	15	BAIAP3	11	ZFR2	10	KNG1	8
MYO5B	15	CNGB3	11	C19orf26	9	LRRK1	8
SYTL2	15	COL24A1	11	C4orf50	9	MAP2	8
ZNF804B	15	DNAAF1	11	CLSPN	9	MMP17	8

4.4.3 Annotation with gnomAD Exomes and filtering to Rare Variants

All variants within the 100 SCZ samples VCF were annotated with allele frequencies and counts from gnomAD Exomes v2.1.1. Out of the total 12,084 variants, 67 variants were not found in gnomAD Exomes, including individual variants in ELAVL4, LPPR5, SCN1A, COL5A2, DNAH7, ABCA12, and GRM7. Additionally, 837 variants were identified in gnomAD Exomes with an allele count of 5 or less, giving a total of 904 sites with rare variants.

The dataset was further filtered for RVs with allele frequencies within the Swedish Exomes of less than 0.01 and allele counts of 5 or less in gnomAD. Individual sites in the file were normalised to reduce multiallelic sites to biallelic sites spread over several lines in the VCF, resulting in a set of 17,070 rare variants distributed across the 904 sites. On average each of the 100 SCZ patients had 170 (SD 15.5) variants. All genotypes were either homozygous (mean 30; SD 4.5) or heterozygous (mean 140; SD

16.5) for the alternate allele. The 100 SCZ patients exhibited an average of 27 rare SNVs (SD 3.8) (Figure 4.5) and 144 rare INDELs (SD 13.7) (Figure 4.6), with approximately 7 of these being rare singletons within a range of 2-13 (Figure 4.7). Following the filtering process, the outlier sample with missing genotypes (identified in Section 4.4.1), showed a consistent number of homozygous ALT genotypes but half the number of heterozygous genotypes compared to SNVs and INDELs in other samples.

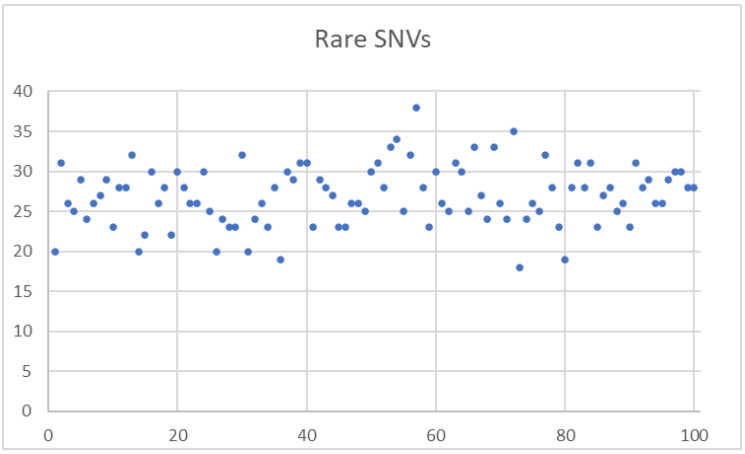

Figure 4.5 Rare SNVs (AF<0.01 and AC<=5 in gnomAD) per sample in SCZ000

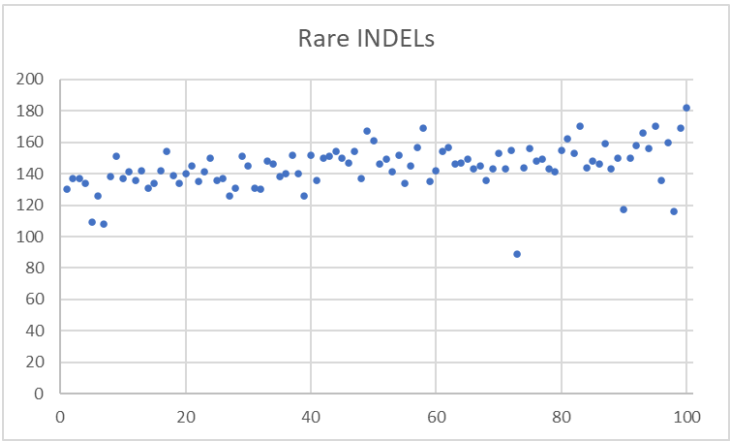

Figure 4.6 Rare INDELs (AF<0.01 and AC<=5 in gnomAD) per sample in SCZ000

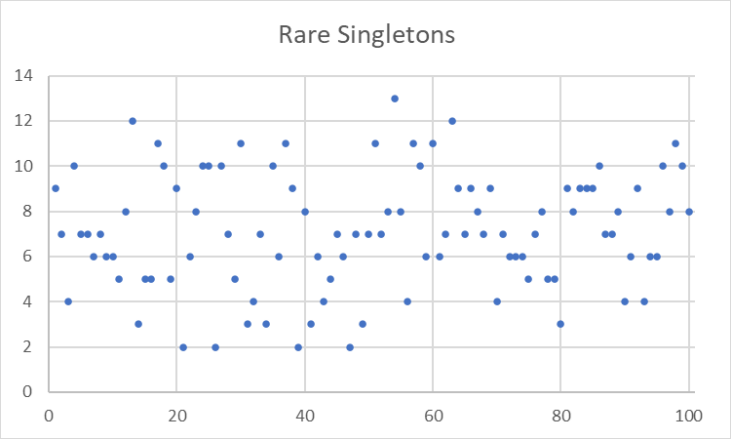

Figure 4.7 Rare singletons (AF<0.01 and AC<=5 in gnomAD) per sample in SCZ000

4.4.4 Variant Effect Predictor

Table 4.6 and Figure 4.8 show the summary results of analysing the normalised, biallelic VCF for SCZ000 using Ensembl VEP. 2.0% of the variants are novel and the rest are known. Since we have identified variants restricted to gene boundaries, the high number of variants overlapping a gene was expected. Additionally, just under one third of the variants overlapped regulatory sequences, including splice regulatory regions. The variant consequences were diverse with most variants identified as intronic (34%) and missense (12%). A smaller number of variants were predicted to cause nonsense mediated decay (NMD – 6%), were synonymous or in regulatory regions. The remainder were predicted to be upstream, downstream, or non-coding. For the variants affecting the coding sequences of genes, 60% were missense variants and 28% synonymous. The remainder were predicted to cause frameshifts (6%), in-frame deletions and insertions (4%), or gain of a stop codon (2%).

Table 4.6 VEP summary data for SCZ000

Category	Count
Variants processed	751
Variants filtered out	0
Novel / existing variants	15 (2.0) / 736 (98.0)
Overlapped genes	646
Overlapped transcripts	2936
Overlapped regulatory features	215

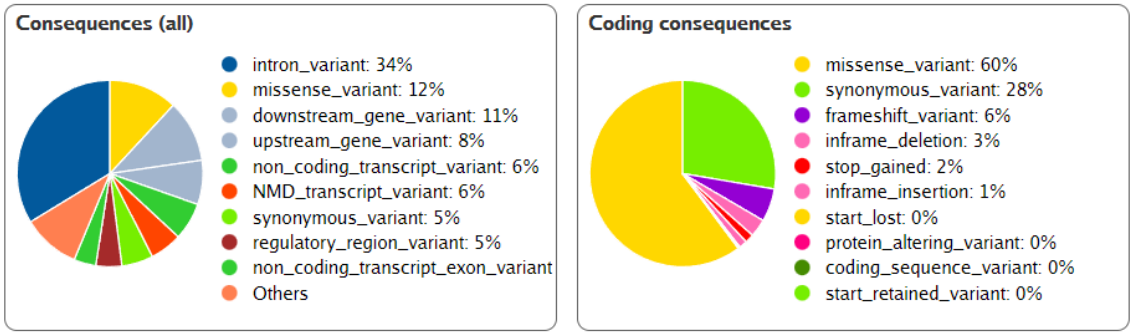

Figure 4.8 Ensembl VEP summary of variants in group SCZ000

There were 92 rare variants predicted to be deleterious by SIFT or damaging by PolyPhen-2 (listed in **Appendix 8.3**) among which 5 have an entry in dbSNP

IGSF3 - rs141947907

MAST4 - rs17221458

CSMD1 - rs11984691

CHST15 - rs75589650

OR2AG1 - rs2659879)

4.4.5 Aggregation to Gene Regions and Subregions

To uncover potential patterns and understand the collective impact of rare variants, a systematic aggregation process was undertaken. Rare variants were binned into the whole gene region as well as into three distinct window sizes: 10 Kb, 20 Kb and 100 Kb. The count and frequency of variants assigned to these bins were further divided into SNPs, INDELs, and singletons. This aggregation aimed to shed light on the possibility

that disease-causing variants may be influencing the same gene or region of a gene in different patients.

4.4.5.1 Gene aggregation

At the level of aggregation to the gene, Table 4.7 shows the first 20 genes ordered by count of rare SNVs in the CO/HP-genes. The first two variants had higher expression in the GTEx Frontal Cortex (Region 1) than in the control regions, but not in the Hippocampus. The complement-related gene CUB and Sushi multiple domains 1 (CSMD1) had the highest count of 9 RVs in this group of 100 patient samples. This was followed by talin-1 (TLN1) with 7 SNVs. Both also had 3 and 2 indels. The list includes two dynein axonemal heavy chains 3 and 11 (DNAH3, DNAH11) with 4 and 3 RVs respectively; the metabotropic glutamate receptor (GRM7) in the hippocampus with 5 RVs and the chromodomain helicase DNA binding protein 5 (CHD5) a regulator of chromatin remodelling with 3 RVs affecting the frontal cortex (region 1).

Table 4.7 Top 20 Genes by count of rare SNVs
Genes are ordered by the total number of single nucleotide rare SNVs.

CHROM	Start	End	Gene	Region	Strand	Count	P/cent	SNVs	SNV_freq	indels	indel_freq
8	2792874	4852494	CSMD1	1	-	12	0.000009	9	0.000004	3	0.000004
15	62682724	63136830	TLN2	1	+	9	0.000026	7	0.000015	2	0.000011
3	123798869	124445172	KALRN	3	+	7	0.000046	5	0.000008	2	0.000042
15	41795834	41806085	LTK	1	-	6	0.000585	5	0.000488	1	0.000098
11	92085261	92629636	FAT3	1	+	5	0.000009	5	0.000009	0	0
16	20944432	21170762	DNAH3	2	-	5	0.000027	4	0.000018	1	0.000009
1	156495196	156542396	IQGAP3	1	-	4	0.000085	4	0.000085	0	0
16	1383601	1399442	BAIAP3	3	+	4	0.000253	4	0.000253	0	0
2	219867567	219906249	CFAP65	2	-	4	0.000103	4	0.000103	0	0
6	160952514	161085291	LPA	1	-	4	0.00003	4	0.00003	0	0
12	80493316	80774659	OTOGL	1	+	7	0.000064	3	0.000011	4	0.000053
7	21582656	21941457	DNAH11	2	+	7	0.000114	3	0.000008	4	0.000106
3	6811687	7783220	GRM7	3	+	5	0.000013	3	0.000003	2	0.000011
10	106401047	107025000	SORCS3	3	+	4	0.000016	3	0.000005	1	0.000011
10	98757794	98945716	SLIT1	3	-	4	0.000021	3	0.000016	1	0.000005
3	166958074	167099134	ZBBX	3	-	4	0.00005	3	0.000021	1	0.000028
5	128795973	129074376	ADAMTS19	1	+	4	0.000018	3	0.000011	1	0.000007
6	159590431	159693140	FNDC1	2	+	4	0.000068	3	0.000029	1	0.000039
1	81771844	82458120	ADGRL2	1	+	3	0.000004	3	0.000004	0	0
1	6161846	6240381	CHD5	1	-	3	0.000038	3	0.000038	0	0

4.4.5.2 Aggregation to 10k, 20k, and 100k bins

Table 4.8, Table 4.9 and Table 4.10 show the top genes by count of variants in the 10k bins, 20k bins and 100k bins respectively. The bin names were compared in Chapter 5

between groups of 100 SCZ and 100 healthy subjects to identify common subregions among patients that are not affected in healthy subjects.

An illustrative example is found in SKA3, where 9 rare variants were distributed across the first 10 Kb, with an additional three variants in the following 10 Kb. 10 of the 12 variants are intronic INDELs, which could be due to technical errors from PCR or due to a low complexity base sequence.

Table 4.8 Top 20 10Kb gene subregions by count of rare variants

CHROM	Start	End	Gene10kbin	VarCount	VarFreq	SNPCount	SNPFreq	IndelCount	IndelFreq	Gene
13	21727733	21737733	SKA3_1	9	0.001000	2	0.000200	7	0.000800	SKA3
15	42293779	42303779	PLA2G4E_3	8	0.002000	NA	NA	8	0.002000	PLA2G4E
12	80723316	80733316	OTOGL_24	5	0.001600	1	0.000100	4	0.001500	OTOGL4
15	41805834	41806085	LTK_2	5	0.019920	4	0.015936	1	0.003984	LTK
11	26571145	26581145	ANO3_25	4	0.002300	NA	NA	4	0.002300	ANO35
2	166891150	166901150	SCN1A_6	4	0.001000	1	0.000100	3	0.000900	SCN1A
3	124058869	124068869	KALRN_27	4	0.002700	2	0.000200	2	0.002700	KALRN7
4	160244164	160254164	RAPGEF2_23	4	0.000800	2	0.000600	2	0.000200	RAPGEF23
5	176794924	176799597	RGS14_2	4	0.011770	2	0.000428	2	0.011556	RGS14
1	214454444	214464444	SMYD2_1	3	0.001100	1	0.000100	2	0.001000	SMYD2
11	72287184	72297184	PDE2A_1	3	0.000300	3	0.000300	NA	NA	PDE2A
12	130920681	130930681	RIMBP2_5	3	0.000200	1	0.000100	2	0.000100	RIMBP2
13	21737733	21747733	SKA3_2	3	0.000500	NA	NA	3	0.000500	SKA3
16	1383601	1393601	BAIAP3_1	3	0.000300	3	0.000300	NA	NA	BAIAP3
16	84198891	84208891	DNAAF1_3	3	0.001400	NA	NA	3	0.001400	DNAAF1
17	80332105	80333462	UTS2R_1	3	0.011054	2	0.001474	1	0.009580	UTS2R
19	10090236	10100236	COL5A3_3	3	0.000500	NA	NA	3	0.000500	COL5A3
2	101576600	101586600	NPAS2_15	3	0.000400	1	0.000100	2	0.000300	NPAS25
20	57828435	57834168	ZNF831_14	3	0.000349	1	0.000174	2	0.000349	ZNF8314
3	7491687	7501687	GRM7_69	3	0.001100	1	0.000100	2	0.001100	GRM79
4	30722036	30732036	PCDH7_1	3	0.000300	3	0.000300	NA	NA	PCDH7

Table 4.9 Top 20 20Kb gene subregions by count of rare variants

CHROM	Start	End	Gene20kbin	VarCount	VarFreq	SNPCount	SNPFreq	IndelCount	IndelFreq	Gene
13	21727733	21747733	SKA3_1	12	0.000750	2	0.000100	10	0.000650	SKA3
15	42293779	42313779	PLA2G4E_2	8	0.001000	NA	NA	8	0.001000	PLA2G4E
12	80713316	80733316	OTOGL_12	6	0.000850	2	0.000100	4	0.000750	OTOGL2
15	41795834	41806085	LTK_1	6	0.000585	5	0.000488	1	0.000098	LTK
11	26571145	26591145	ANO3_13	5	0.001200	1	0.000050	4	0.001150	ANO33
16	1383601	1399442	BAIAP3_1	4	0.000253	4	0.000253	NA	NA	BAIAP3
2	166881150	166901150	SCN1A_3	4	0.000500	1	0.000050	3	0.000450	SCN1A
3	124058869	124078869	KALRN_14	4	0.001350	2	0.000100	2	0.001350	KALRN4
4	160244164	160264164	RAPGEF2_12	4	0.000400	2	0.000300	2	0.000100	RAPGEF22
5	176784924	176799597	RGS14_1	4	0.003748	2	0.000136	2	0.003680	RGS14
7	81695759	81715759	CACNA2D1_7	4	0.000550	1	0.000050	3	0.000500	CACNA2D1
1	214454444	214474444	SMYD2_1	3	0.000550	1	0.000050	2	0.000500	SMYD2
11	72287184	72307184	PDE2A_1	3	0.000150	3	0.000150	NA	NA	PDE2A
12	130920681	130940681	RIMBP2_3	3	0.000100	1	0.000050	2	0.000050	RIMBP2
16	4932507	4952507	PPL_1	3	0.000150	3	0.000150	NA	NA	PPL
16	84198891	84212373	DNAAF1_2	3	0.001038	NA	NA	3	0.001038	DNAAF1
17	61554421	61574421	ACE_1	3	0.000150	3	0.000150	NA	NA	ACE
17	80332105	80333462	UTS2R_1	3	0.011054	2	0.001474	1	0.009580	UTS2R
19	10090236	10110236	COL5A3_2	3	0.000250	NA	NA	3	0.000250	COL5A3
19	19322766	19342766	NCAN_1	3	0.000150	2	0.000100	1	0.000050	NCAN
19	55484497	55504497	NLRP2_2	3	0.000300	2	0.000100	1	0.000200	NLRP2

Table 4.10 Top 20 100 Kb gene subregions by count of rare variants

CHROM	Start	End	Gene100kbin	VarCount	VarFreq	SNPCount	SNPFreq	IndelCount	IndelFreq	Gene
13	21727733	21750691	SKA3_1	12	0.000653	2	0.000087	10	0.000566	SKA3
15	42273779	42343388	PLA2G4E_1	9	0.000302	1	0.000014	8	0.000287	PLA2G4E
12	80693316	80774659	OTOGL_3	6	0.000209	2	0.000025	4	0.000184	OTOGL
15	41795834	41806085	LTK_1	6	0.000585	5	0.000488	1	0.000098	LTK
15	62982724	63082724	TLN2_4	6	0.000090	4	0.000040	2	0.000050	TLN2
3	123998869	124098869	KALRN_3	6	0.000290	4	0.000040	2	0.000270	KALRN
8	2792874	2892874	CSMD1_1	6	0.000070	5	0.000050	1	0.000020	CSMD1
11	26531145	26631145	ANO3_3	5	0.000240	1	0.000010	4	0.000230	ANO3
19	10070236	10121180	COL5A3_1	5	0.000137	2	0.000039	3	0.000098	COL5A3
4	160224164	160281326	RAPGEF2_3	5	0.000192	2	0.000105	3	0.000087	RAPGEF2
1	156495196	156542396	IQGAP3_1	4	0.000085	4	0.000085	NA	NA	IQGAP3
10	98757794	98857794	SLIT1_1	4	0.000040	3	0.000030	1	0.000010	SLIT1
12	130880681	130980681	RIMBP2_1	4	0.000030	2	0.000020	2	0.000010	RIMBP2
16	1383601	1399442	BAIAP3_1	4	0.000253	4	0.000253	NA	NA	BAIAP3
16	20944432	21044432	DNAH3_1	4	0.000050	3	0.000030	1	0.000020	DNAH3
2	166841150	166941150	SCN1A_1	4	0.000100	1	0.000010	3	0.000090	SCN1A
2	219867567	219906249	CFAP65_1	4	0.000103	4	0.000103	NA	NA	CFAP65
5	89925160	90025160	ADGRV1_2	4	0.000050	1	0.000010	3	0.000040	ADGRV1
5	176784924	176799597	RGS14_1	4	0.003748	2	0.000136	2	0.003680	RGS14
6	159590431	159690431	FNDC1_1	4	0.000070	3	0.000030	1	0.000040	FNDC1
7	21582656	21682656	DNAH11_1	4	0.000380	1	0.000010	3	0.000370	DNAH11

4.4.6 Aggregation in Pathways

To understand the collective impact of RVs within the context of functional pathways, the count of RVs against each gene in the CO/HP-gene list was determined. Genes with

non-zero counts were subjected to an overrepresentation GO analysis using PANTHER version 16.0 (release 20210224) (Mi et al., 2021). The analysis encompassed various sets of functional annotations, including both GO and PANTHER GO-Slim biological processes, 'GO molecular function complete', and PANTHER Pathways and Reactome Pathways. PANTHER Protein Classes were also considered.

In Table 4.11 and Table 4.12 the top 20 enriched GO and GO-Slim biological process terms are provided for genes in the CO/HP-genes set that have rare variants. Among genes with RVs, the median number of RVs per gene is 1. Genes with more than one variant often have INDELs that might be sequencing errors. Therefore, 'genes that have variants' were used in the overrepresentation test against all genes in the human reference genome.

A common thematic thread emerged across the top 20 results of both tests, focussing on axon extension, projection, and guidance; synaptic membrane adhesion and signalling; and cilium movement. The GO biological processes (GO BP) delved into specific aspects such as axon guidance, emphasising various neural structures like the trigeminal nerve, retinal ganglion cell, cerebral cortex neurons, olfactory bulb neurons, and motor neurons. GO-Slim BP terms highlighted overarching processes like axon genesis and development, but specific cell types were not emphasised. Within the processes of synaptic adhesion and transmission, GO BP terms included synaptic membrane adhesion and distinct processes such as GABA signalling, semaphorin-plexin signalling, and serotonin receptor signalling. GO-Slim BP terms emphasised regulation and modulation of glutamatergic signalling and chemical synaptic transmission. For processes related to cilia movement, the both GO BP and GO-Slim BP terms included epithelial cilia movement involved in extracellular fluid movement, and axonemal dynein complex assembly. The overrepresentation test returned 192 significant terms overall for GP BP and 72 for GO-Slim BP.

Table 4.11 Top 20 GO biological processes overrepresented for genes with RVs in SCZ000

Overrepresentation test of genes that have RVs carried out against the GO biological processes complete dataset via PANTHERdb.org (release 20210224) using Fisher's Exact test with FDR < 0.05. Results were ordered by fold enrichment. Indented processes are parent categories that also have a fold enrichment within the top 20 values. Overall, 171 terms were significantly overrepresented, and 21 terms were underrepresented. 431 gene IDs were mapped out of 444 genes with an RV.

GO biological process complete	fold Enrichment	raw P-value	FDR
trigeminal nerve structural organization (GO:0021637)	27.83	4.83E-04	4.10E-02
trigeminal nerve morphogenesis (GO:0021636)	27.83	4.83E-04	4.08E-02
regulation of retinal ganglion cell axon guidance (GO:0090259)	27.83	4.83E-04	4.06E-02
regulation of axon guidance (GO:1902667)	16.87	2.22E-04	2.24E-02
axon extension involved in axon guidance (GO:0048846)	23.19	8.47E-05	1.08E-02
axon extension (GO:0048675)	7.73	2.35E-04	2.36E-02
neuron projection extension involved in neuron projection guidance (GO:1902284)	23.19	8.47E-05	1.09E-02
cerebral cortex neuron differentiation (GO:0021895)	12.10	2.69E-05	4.31E-03
synaptic membrane adhesion (GO:0099560)	9.66	3.20E-04	3.03E-02
olfactory bulb development (GO:0021772)	9.28	9.72E-05	1.18E-02
olfactory lobe development (GO:0021988)	8.70	1.33E-04	1.47E-02
motor neuron axon guidance (GO:0008045)	8.98	1.14E-04	1.34E-02
gamma-aminobutyric acid signaling pathway (GO:0007214)	8.92	4.42E-04	3.86E-02
epithelial cilium movement involved in extracellular fluid movement (GO:0003351)	7.95	2.05E-04	2.10E-02
extracellular transport (GO:0006858)	7.32	3.06E-04	2.95E-02
regulation of neuron migration (GO:2001222)	7.55	8.11E-05	1.08E-02
semaphorin-plexin signaling pathway (GO:0071526)	7.32	3.06E-04	2.93E-02
axonemal dynein complex assembly (GO:0070286)	7.32	3.06E-04	2.91E-02
forebrain neuron differentiation (GO:0021879)	7.28	3.19E-05	5.07E-03
serotonin receptor signaling pathway (GO:0007210)	7.14	3.46E-04	3.20E-02

Table 4.12 Top 20 GO-Slim biological processes overrepresented for genes with RVs in SCZ000

Overrepresentation test of genes that have RVs carried out against the PANTHER GO-Slim biological processes dataset via PANTHERdb.org (release 20210224) using Fisher's Exact test with FDR < 0.05. Results were ordered by fold enrichment. Indented processes are parent categories that also have a fold enrichment within the top 20 values. Overall, 65 terms were significantly overrepresented, and 7 terms were underrepresented. 431 gene IDs were mapped out of 444 genes with an RV.

PANTHER GO-Slim Biological Process	fold Enrichment	raw P-value	FDR
regulation of synaptic transmission, glutamatergic (GO:0051966)	9.28	1.49E-03	4.57E-02
modulation of chemical synaptic transmission (GO:0050804)	5.03	1.41E-04	6.57E-03
regulation of trans-synaptic signaling (GO:0099177)	5.03	1.41E-04	6.71E-03
cilium movement (GO:0003341)	6.63	4.95E-04	1.83E-02
blood circulation (GO:0008015)	5.68	1.03E-03	3.36E-02
circulatory system process (GO:0003013)	6.49	1.89E-04	8.42E-03
axon guidance (GO:0007411)	5.35	6.38E-06	6.96E-04
axonogenesis (GO:0007409)	4.64	5.08E-06	6.16E-04
neuron projection morphogenesis (GO:0048812)	3.67	5.77E-05	3.81E-03
cell projection morphogenesis (GO:0048858)	3.65	6.10E-05	3.80E-03
neuron projection development (GO:0031175)	3.65	1.01E-05	9.19E-04
neuron development (GO:0048666)	3.75	2.12E-06	2.73E-04
neuron differentiation (GO:0030182)	3.77	2.59E-08	4.71E-06
axon development (GO:0061564)	4.54	6.36E-06	7.30E-04
cell morphogenesis involved in neuron differentiation (GO:0048667)	4.11	1.81E-05	1.36E-03
neuron projection guidance (GO:0097485)	5.35	6.38E-06	6.63E-04
chemical synaptic transmission (GO:0007268)	4.27	1.02E-08	3.17E-06
anterograde trans-synaptic signaling (GO:0098916)	4.27	1.02E-08	2.78E-06
trans-synaptic signaling (GO:0099537)	4.23	1.17E-08	2.83E-06
synaptic signaling (GO:0099536)	4.08	2.26E-08	4.49E-06

Table 4.13 shows the top 20 results for an overrepresentation test against the GO molecular function complete dataset in PANTHER 16.0. The term with the highest fold enrichment is related to axon extension and guidance (semaphorin receptor activity). Following this, are signal transduction terms (cAMP binding and G-protein coupled amine receptor (GPCR) activity), synaptic transmission terms (postsynaptic neurotransmission receptor activity) and several channel and transporter terms (voltage- and ligand-gated cation channels and potassium ion transmembrane transporter activity). Lastly, heparin binding relates to genes involved cell adhesion, migration, and extracellular matrix organisation.

Table 4.13 Top 20 GO molecular functions overrepresented for genes with RVs in SCZ000

Overrepresentation test of genes that have RVs carried out against the GO molecular function complete dataset via PANTHERdb.org (release 20210224) using Fisher's Exact test with FDR < 0.05. Results are ordered by fold enrichment. Indented processes are parent categories that also have a fold enrichment within the top 20 values. Overall, 43 terms are significantly overrepresented, and 2 terms are underrepresented. 431 genes are mapped out of 444 genes with an RV.

GO molecular function complete	fold Enrichment	raw P-v	FDR
semaphorin receptor activity (GO:0017154)	14.27	3.74E-04	4.25E-02
cAMP binding (GO:0030552)	10.08	2.69E-04	3.37E-02
cyclic nucleotide binding (GO:0030551)	7.52	2.69E-04	3.45E-02
G protein-coupled amine receptor activity (GO:0008227)	6.37	2.11E-04	2.79E-02
postsynaptic neurotransmitter receptor activity (GO:0098960)	5.8	1.36E-04	2.01E-02
neurotransmitter receptor activity (GO:0030594)	5.34	2.72E-06	1.20E-03
cytoskeletal motor activity (GO:0003774)	4.36	8.69E-05	1.32E-02
voltage-gated cation channel activity (GO:0022843)	4.22	2.80E-05	6.50E-03
voltage-gated ion channel activity (GO:0005244)	3.69	1.61E-05	4.91E-03
voltage-gated channel activity (GO:0022832)	3.67	1.71E-05	4.62E-03
gated channel activity (GO:0022836)	4.05	4.57E-10	2.23E-06
channel activity (GO:0015267)	3.23	7.36E-09	1.20E-05
passive transmembrane transporter activity (GO:0022803)	3.22	7.72E-09	9.42E-06
cation channel activity (GO:0005261)	3.42	2.70E-07	1.32E-04
ion channel activity (GO:0005216)	3.4	6.55E-09	1.60E-05
ligand-gated cation channel activity (GO:0099094)	4.03	3.24E-04	3.76E-02
ligand-gated ion channel activity (GO:0015276)	4.13	3.42E-05	6.67E-03
ligand-gated channel activity (GO:0022834)	4.13	3.42E-05	6.95E-03
potassium ion transmembrane transporter activity (GO:0015079)	3.87	6.44E-05	1.05E-02
heparin binding (GO:0008201)	3.57	1.37E-04	1.96E-02

In Table 4.14 and Table 4.15 the genes are tests for overrepresentation against two pathway datasets in PANTHER 16.0, PANTHER Pathways and Reactome. Only two PANTHER Pathways survived multiple testing, both of which are related to GPCR signalling pathways, Gq/Go and Gi/Gs alpha mediated pathways. 352 genes were not classified in PANTHER Pathways. In Reactome, 5 pathways survived multiple testing (FDR-adjusted p value<0.05). As for PANTHER Pathways, genes involved in GPCR Gq alpha subunit signalling are overrepresented. Like GO molecular functions, genes involved in cyclic nucleotide binding, amine ligand-binding receptors, and synaptic transmission were overrepresented.

Table 4.14 PANTHER Pathways overrepresented for genes with RVs in SCZ000
Overrepresentation test of genes that have RVs carried out against the PANTHER Pathways dataset via PANTHERdb.org (release 20210224) using Fisher's Exact test with FDR < 0.05. Results are ordered by fold enrichment. Overall, 23 terms are significantly overrepresented. 352 genes are unclassified in PANTHER Pathways.

PANTHER Pathways	fold Enrichment	raw P-value	FDR
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P00027)	4.11	1.40E-04	1.17E-02
Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)	3.35	4.15E-04	2.31E-02
Unclassified (UNCLASSIFIED) - 352 genes	0.91	2.83E-06	4.73E-04

Table 4.15 Reactome Pathways overrepresented for genes with RVs in SCZ000
Overrepresentation test of genes that have RVs carried out against the Reactome Pathways dataset via PANTHERdb.org (release 20210224) using Fisher's Exact test with FDR < 0.05. Results are ordered by fold enrichment. Overall, 5 terms are significantly overrepresented and 1 underrepresented. 430 genes of 444 genes were mapped in PANTHER Pathways.

Reactome Pathways	fold Enrichment	raw P-value	FDR
Nitric oxide stimulates guanylate cyclase (R-HSA-392154)	14.76	1.86E-06	2.12E-03
Amine ligand-binding receptors (R-HSA-375280)	7.55	8.11E-05	3.09E-02
G alpha (q) signalling events (R-HSA-416476)	3.9	2.35E-06	1.79E-03
Neuronal System (R-HSA-112316)	3.46	1.32E-08	3.02E-05
Transmission across Chemical Synapses (R-HSA-112315)	3.39	8.31E-06	4.74E-03
Metabolism of RNA (R-HSA-8953854)	(0.07)	2.24E-05	1.02E-02

4.5 Discussion

Chapter 2 identified two target GTEx brain regions impacting SCZ (Frontal Cortex and Hippocampus) and two contrast regions with less evidence for impact in SCZ (Cerebellum and Substantia nigra). Chapter 3 RNA-Seq differential expression identified highly expressed genes between target and control brain regions in 25 healthy donors. CO-Genes were genes with greater than 1.5 times expression in the Frontal Cortex compared to both control regions. HP-Genes were genes with greater than 1.5 times expression in the Hippocampus compared to both control regions. CO-/HP-Genes combined both lists with each gene labelled with its primary region of increased expression (Frontal Cortex/Hippocampus/Both). These genes were chosen as candidates for further study due to their potential increased impact in these regions, not as definitive determinants of SCZ.

This chapter described the set of filtering and aggregation methods and outputs applied to whole exome sequencing (WES) data from an example group of 100 patients with SCZ randomly sampled from the Swedish Exomes data. Variants from the 100

patients with SCZ were first restricted to those located in the 1,074 CO-/HP-genes identified in Chapter 3. For these remaining variants, the following outputs were produced:

- Metrics of variant burden in the CO-/HP-genes – number of variants and number of SNPs, indels and singletons (relative to the Swedish Exomes study size).
- Coding consequences predicted using BCFtools csq function – including multinucleotide variants (MNVs) – number of variants that cause missense, synonymous, stop gain, etc.
- Variants not found in or with low allele frequencies in gnomAD Exomes.
- Metrics of rare variants (RV not in gnomAD) burden in the CO-/HP-genes – number of RVs and number of SNPs, indels and singletons.
- SIFT and PolyPhen2 scores for RVs in the CO-/HP-genes provided by Ensembl's VEP.
- Lists of CO-/HP-genes with counts of RVs per gene, or counts of RVs in 10Kb, 20Kb and 100Kb windows in the subregions of these genes.
- Enrichment of rare variants in GO terms, Reactome and PANTHER pathways most significantly associated with CO-/HP-genes.

4.5.1 Variant Burden in the CO-/HP-genes in an example group of 100 cases with SCZ

The variant burden in the 1,074 CO-/HP-genes, previously prioritised for their increased expression in SCZ-relevant brain regions, were investigated to explore their potential involvement in SCZ pathogenesis. 100 SCZ samples were extracted from the Swedish Exomes dataset and filtered to the CO-/HP-genes. 12,084 biallelic and multiallelic (n=358) variants sites (11,033 SNPs and 1,099 indels) were identified in the CO-/HP-genes in these samples (Section 4.4.1). Each sample had an average coverage of 97.4% for these sites, ensuring reliable genotyping for samples within the dataset.

Each sample carried an average of 2,470 SNVs (including 38 singletons) and 287 indels within the CO-/HP-genes, establishing a baseline for further analysis. Rare variants, often enriched for functional consequence (Tennessen et al., 2012), could not be definitively identified without comparison to a larger dataset. Attempts in the past (Genovese et al., 2016; Purcell et al., 2014) assessed SNVs against 1000 Genomes (1000 Genomes Project Consortium et al., 2015) and ExAC (Lek et al., 2016) leading to

overestimation of the number of rare variants, that recent larger population datasets can ameliorate. However, given the gene selection, some of these variants might contribute to SCZ pathogenesis or susceptibility.

Functional annotation and population genetics analysis will further refine the candidate variants. However, a control-based analysis is needed to distinguish disease-associated variants from variants that are also present in healthy individuals within the Swedish Exomes population.

Whole exome sequencing data is limited in the first instance by the exome extraction kit. The Swedish Exomes project used two different kit versions (Agilent SureSelect Human All Exon Kit (n=132) and v2 (n=12,248)) (Purcell et al., 2014), potentially affecting coverage in the 1,074 CO-/HP-genes. All groups were processed with these same methods in Chapter 5. Samples that did not have over 80% coverage for all 1.8 million sites in the Swedish Exomes dataset were removed before sampling each group of 100.

To further prioritise candidate variants, functional annotation and population genetics analyses were conducted, followed by variant aggregation and enrichment analyses.

4.5.2 Coding consequences

The diverse array of variant types within the 1,074 CO-/HP-genes revealed a spectrum of potential functional implications. The BCFtools csq (Danecek & McCarthy, 2017) analysis assessed the variants in the context of a gene annotation model (Ensembl 97) and the human reference genome (GRCh37) and identified missense variants as the most prevalent (3,163; ~89% of 'missense or worst' consequences), followed by frameshift (63), in-frame deletion (57), insertion (31), stop-gained (37) and splice-region variants (35). These alterations can potentially disrupt protein structure and function to varying degrees, potentially contributing to SCZ susceptibility.

Multinucleotide variants (MNVs) composed of neighbouring variants with combined effects deserve further attention. Two illustrative examples identified in this study highlight their potential importance. The first, in EXTL1 (gnomAD: 1-26361669-CG-AT), involves a missense and a synonymous variant modifying the same codon, highlighting the how synonymous variants can alter coding consequences beyond individual predictions. The second, in LPA (gnomAD: 6-161056166-GG-AA), combines two coding

changes, further reinforcing the complex nature impacts of MNVs. The EXTL1 MNV has been observed in 752 exomes and 66 genomes with an allele frequency of 0.3% in over 131,000 individuals. The LPA MNV has been observed in 358 exomes and 37 genomes in gnomAD with an allelic frequency of 0.15%. These two MNVs in gnomAD were recently analysed using GATK's popular HaplotypeCaller (Wang, Q. et al., 2020); the authors found on average that approximately 60% of MNVs had a different coding consequence to the constituent variants in the MNV. The gnomAD Exomes are also unphased and the authors referenced a dataset of MNVs called from trios to verify most of the MNVs. In further studies, the MNV outputs from this analysis could be verified against the later releases of gnomAD for verification.

Not all 116 MNVs were thoroughly examined. To calculate the MNVs requires a haplotype-aware software that can detect neighbour variants even across exons on the same strand. In this chapter, because the Swedish Exomes data is unphased, BCFtools csq was set to assume that the genotypes for each variant were phased which would place all variants on the same chromosome (in the same haplotype). All genotypes were listed as REF/ALT, for example 0/1, 0/2, 1/2. In a phased VCF file, genotypes listed as 0|2 and 1|0 would inform that the reference in the first is in phase with the alternate allele in the second. This affects the accuracy of the assignment variants to MNVs, and the coding consequences require very careful assessment to avoid misannotating the variants, which can lead to a misdiagnosis (Wakeling et al., 2019). For example, if the two variants in the EXTL1 were from separate strands, calling an MNV in this case would lead to a misinterpretation of their coding consequence. MNVs are understudied (Wang, Q. et al., 2020) because they require correctly phased sequenced data to analyse them with accuracy. However, this output may reveal common MNVs when compared between SCZ groups and controls.

The main limitation for Bcftools csq was the requirement of a phased VCF input file or an up-to-date database of MNVs to verify against. Other haplotype-aware callers can call MNVs more accurately from unphased data, but they require access to the alignment file (BAM) so that variants cane be verified against the original reads (Srinivasan et al., 2021). The recognition of the limitations of unphased data has evoked interest in developing the analysis of MNVs. Future studies employing phased data or advanced haplotype-aware callers will be crucial to accurately capture the full

spectrum of MNVs and their potential contributions to complex genetic pathogenesis, such as in SCZ.

4.5.3 Identifying Rare Variants in CO-/HP-genes - not found in gnomAD Exomes

Analysing variants absent or almost absent from gnomAD Exomes can reveal potentially rare or ultra-rare mutations in the 100 SCZ samples. All variants in the CO-/HP-genes were annotated with allele counts and frequencies from European and global populations in gnomAD. Rare variants were then filtered to a frequency of less than 1% (AF<0.01) in the Swedish Exomes data and with a count of 5 or less in gnomAD Exomes.

Of the 12,084 variant sites, 67 variants were not found in gnomAD Exomes and 837 had an allele count of 5 or less. Each sample in the group had, on average, 170 rare variants spread across the 904 sites: approximately 144 were indels and 27 were SNVs of which 7 were singletons. The majority of the 144 indels were multiallelic and located within 20-40 bases from an exon boundary. The Exome extraction kit used for Swedish Exomes captured approximately 20 bp of flanking intronic sequence, suggesting that the large number of indels could be technical variation at the ends of the extraction kit primers. Some indels may be true variants – however no attempt here was made to extract true indels and the focus was on rare coding SNVs. Outputs from this analysis include total number of rare variants and counts of indels, SNVs, and singletons. In Chapter 5, only the counts for SNVs and singletons will be useful for analysis between SCZ and control groups.

The two earlier studies on the Swedish Exomes data had assessed rare variants using 1000 Genomes (Purcell et al., 2014) and ExAC (Genovese et al., 2016) population data. Purcell et al. (2014) found that 56% of rare variants not seen in 2,504 samples from the 1000 Genomes project were singletons. Using data from 45,376 individuals in ExAC, Genovese et al. (2016) were able to reclassify approximately two thirds of those singletons as less rare variants. In their larger cohort, Genovese et al. identified 244,246 rare coding and splice site variants in 4,877 SCZ and 6,203 control samples with an allele count of 1 in the Swedish Exomes study and not found in ExAC – equating on average to approximately 22 singletons per sample. In this chapter, variants were filtered through approximately 125,000 in gnomAD Exomes (Karczewski et al., 2020), further eliminating false RVs and singletons from the results. The filtering step was

carried out after filtering to the CO-/HP-genes and to 100 SCZ samples and rare REF alleles were counted if 5 or fewer individuals in gnomAD carried the REF allele. The CO-/HP-genes represent approximately 1/20th of the number of genes. Due to codon degeneracy, and evolutionary conserved sequences, genes differ on their ability to tolerate genetic variation. It is therefore challenging to make a direct comparison with these two previous studies. Further studies on Swedish Exomes will use gnomAD 4.0 or later, which includes variation from over 1.4 million individuals.

This step filtered by an allele frequency (AF) of less than 1% in the Swedish Exomes dataset. This could be considered too high as a credible population AF for SCZ (Whiffin, Nicola et al., 2017) where a range of penetrance values may be seen. However, this was ameliorated by included the allele count filter of 5 or less in gnomAD. Those RVs with a count of 5 in gnomAD Exomes will have a much lower AF than 1%. This focused specifically on ultra-rare variants and singletons, which are thought to have larger effects on gene function and are potentially more relevant to complex diseases like SCZ.

4.5.4 Variant Effect Prediction including SIFT and PolyPhen2 scores

Assessing the functional impact of variants using computational prediction tools, providing additional insights into their potential pathogenicity. The gnomAD-filtered rare variants were normalised to produce a list of biallelic sites and these were processed in a cached, offline version of Ensembl's Variant Effect Predictor (VEP) (McLaren et al., 2016) to include SIFT and Polyphen-2 scores for variant tolerance and predicted pathogenicity. For the coding sequence of the genes, 60% of the 751 RVs were predicted to be missense variants, and the remainder were synonymous (28%), frameshift (6%), in-frame deletions and insertions (4%), or gain of a stop codon (2%).

VEP uses a plug-in architecture and can be used to assess variants using various pathogenicity prediction tools, including SIFT, PolyPhen-2, CADD, REVEL. The databases are very large, and this study was limited for diskspace to include as multiple databases. Ideally, rare variants could be annotated with metrics from tools that predict, for example, functional changes, or conformational changes in the protein, or alteration of conserved sites. As these tools are testing for different aspects of the variant, the predictions from these tools often disagree with each other. A grid of metrics would facilitate a Bayesian analysis of the variants. This study intended to

include CADD and other scores but was finally restricted to SIFT and PolyPhen-2 included in the main VEP version. 92 of the variants were identified as deleterious or pathogenic by SIFT and PolyPhen-2. Notably, several of these variants were in the known SCZ-gene CSMD1, which is among the 287 GWAS loci with genome-wide significant association with SCZ risk (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Trubetskoy et al., 2022).

It is not uncommon for predictors of pathogenicity to disagree in their predictions. Selection of a wide range of predictors is important for covering many aspects of the physiochemistry of variants and their effects.

4.5.5 Aggregation to Gene Regions and Subregions

Examining patterns of variant distribution within specific gene regions or subregions, potentially reveals functional hotspots or critical domains. Rare variants were summed per gene in the CO-/HP-genes, 10/20/100 kilobase subregions of these genes, and were further subdivided into counts of SNVs and indels. Aggregating to genes, highlighted CSMD1 and TLN1 in the GTEx Frontal Cortex with 9 and 7 SNVs each and 3 and 2 indels, respectively. There was a total of 7 SNVs in the dynein axonemal subunits.

CSMD1 was the only gene in the top 20 genes to contain a genome-wide risk allele in the top 128 GWAS SNPs of the first GWAS study by the Schizophrenia Work Group of the PGC (Sahraeian et al., 2017). It is involved in the complement system, including the degradation of C3 and C4 (Ermis Akyuz & Bell, 2022; Parker et al., 2022; Woo et al., 2020). DNAH3 and DNAH11 are subunits of the axonemal dynein motor complex involved in the movement of microtubules in the primary cilia in neurons. Primary cilia are involved in signalling in the CA1 region of the hippocampus and various other roles in neuronal development and migration (Park, S. M. et al., 2019; Vien et al., 2023). There are 7 axonemal dynein subunit genes in the 1074 CO-/HP-genes, and there may be several more genes associated with primary cilia within curated biological pathways or GO Cellular Compartments. This drives the case for aggregating variants within pathways to draw out potential underlying pathophysiology in SCZ.

Gene subregion aggregation was based on window sizes of 10, 20 or 100 kilobases. The initial idea was to use these values to optimise the window size, however there was not enough time to complete this. The last window in each gene region is

truncated, making the comparison between windows challenging without normalising counts within all windows. Subregions with large indel counts were more likely to contain exon boundaries and manually examining indels with probably technical causes for removal counters the benefits of a computational pipeline. However, subregions and indel counts can be compared between groups of 100 samples (performed in Chapter 5) to investigate any differences between groups of cases and groups of controls.

The analysis in Chapter 4 did not consider potential variations in mutation tolerance (Karczewski et al., 2020) across different gene and their subregions. This could have influenced the observed variant counts and enrichment patterns, with highly conserved regions potentially harbouring fewer variants regardless of disease status. While this chapter provides valuable insights into the variant landscape within the studied SCZ cases, definitive conclusions about disease-specific patterns will require the upcoming control group analysis in Chapter 5. However, incorporating measures of mutation tolerance or sequence conservation in future analyses could offer a more refined understanding of the observed variation and its potential role in SCZ pathogenesis.

4.5.6 Aggregation in Pathways

The collective impact of variants in functional pathways can aid in highlighting important pathways in the pathogenesis of SCZ. Genes and their variant counts in the CO-/HP-genes for the group of 100 SCZ patients were analysed using overrepresentation analyses using Fisher's Exact test with FDR < 0.05 against PANTHER and Reactome pathways databases, and against PANTHER GO-Slim databases (Biological Function, Cellular Compartment, and Molecular Function). A ranked list of GO terms overrepresented for 'genes with SNVs' for each test was output from this step.

Overrepresentation analysis revealed a common thematic thread across the top 20 enriched terms in each database, converging on three key functional domains:

- Neuronal development and Axon Guidance: Processes like axon extension, projection, and guidance, as well as specific neural structures like the trigeminal nerve and retinal ganglion cells, emerged as significantly enriched.

This finding aligns with established knowledge implicating abnormal neuronal development and connectivity in SCZ pathogenesis (Bernardin et al., 2020; Birnbaum & Weinberger, 2017; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Trubetskoy et al., 2022)

- Synaptic function and Transmission: Terms related to synaptic membrane adhesion, GABA signalling, and glutamatergic signalling were prominently enriched, highlighting the role of disrupted synaptic communication in SCZ.
- Cilia movement: Both GO and Reactome pathways associated with epithelial cilia movement and axonemal dynein complex assembly were overrepresented. This intriguing finding suggested a potential role of ciliary dysfunction in SCZ, warranting further investigation.

The enriched pathways identified in this analysis resonate with existing research implicating these same processes in SCZ. Numerous studies have documented abnormalities in neuronal development and axonal connectivity in patients with SCZ, often linked to specific genes like neuregulin (Chang, X. et al., 2018; Hatzimanolis et al., 2013; Wright, C. et al., 2015). Similarly, dysfunctions in synaptic transmission, particularly glutamatergic signalling, have been extensively studied in SCZ, with NMDA receptor hypofunction as a prominent hypothesis (Trubetskoy et al., 2022).

The link between ciliary dysfunction, primarily in the hippocampus region, initially less explored, is gaining increasing attention. Recent studies have identified rare variants in ciliary genes in SCZ patients, and animal models with disrupted cilia exhibit behavioural and cognitive deficits resembling the disease (Munoz-Estrada et al., 2018). These findings bolster this emerging connection, highlighting the importance of investigating ciliary function in a larger SCZ population.

While the findings in this sample of 100 patients offer an insight into pathway enrichment associated with rare variants in SCZ, it is important to acknowledge certain limitations. This analysis in this group of 100 patients focused on 'genes with RVs' without weighting, therefore all genes were treated equally in the test regardless of the total number of variants in each gene. Pathways and GO terms that include CSMD1, with 9 variants in this group of 100 patients, need to be given slightly more weighting than genes with only one variant. This was further addressed in Chapter 5 by switching to EnrichR to produce lists of enriched pathways and GO terms. While

the overrepresentation analyses do highlight Cortex and Hippocampal SCZ-relevant pathways and biological functions (synaptic maintenance, axon guidance, glutamatergic signalling, etc), enrichment tests in specific gene sets, as well as disease gene sets in DisGeNet, may have been more appropriate. Testing for enrichment in specific gene sets highlights specific pathways. enrichR, used in Section 3.2.6.9, was performed after processing the data in this chapter. enrichR provided a more convenient workflow in R and the outputs were more amenable to the comparisons for chapter 5.

PANTHERdb.org (Mi et al., 2021) makes available the curated PANTHER family (GO-Slim) datasets, complete GO annotation datasets, as well as curated pathways from Reactome and PANTHER Pathways, which are all updated monthly. GO Complete datasets tend to rely on computational analysis. PANTHER datasets are manually curated (and the whole dataset was re-curated between release 14.0 and 16.0) (Mi et al., 2021). Therefore, while the GO Complete datasets are comprehensive covering many more genes per database, they may be less reliable than the PANTHER GO-Slim datasets. The overrepresentation tests were performed against a background that included all genes in the genome. The test compares the genes in the list against those expected from the background. However, having restricted the variants analysed in this chapter to the CO-/HP-genes, they would not be expected to be seen in the results. It would be more appropriate to use either all genes with some expression in the Frontal Cortex and Hippocampus from the RNA-Seq analysis or the full CO-/HP-gene list as the reference list.

4.5.7 Limitations

This study has a number of limitations.

- Whole Exome data is restricted to exons +- 20bp – some SCZ relevant variants could affect promoter/enhancer regions.
- Missing Exons. The Swedish Exomes used an extraction kit based on an old gene annotation model. Variants in exons that have been discovered since this study will not be seen in any analysis of the data. WGS far superior to WES but requires much more storage – and processing power.

4.5.8 Further work

Future research can expand on these findings by:

- Utilizing network analysis: Integrating pathway data with protein-protein interaction networks to identify key hub genes and potential therapeutic targets within these pathways.
- Investigating the role of non-coding variants: Analysing the impact of regulatory elements and non-coding RNAs on the expression of genes involved in enriched pathways.

5 Grouped Variant Case-Control Analysis

5.1 Introduction

5.1.1 Schizophrenia

SCZ is a highly polygenic disease. The genetic contribution is thought to come from rare variants with large effects with few cases sharing the exact same pathogenic variants (Singh et al., 2016, 2017, 2022). The genes affected however are likely to converge on several cellular processes and biological pathways, for example, neuronal development, signalling, and synaptic plasticity.

Polygenic – Can be understood in two ways:

- that multiple genes are affected within a single patient.
- that each patient may have only one or a few damaging variants, but across a population, variants are found in many different genes throughout the genome.

5.1.2 Whole Exome Sequencing data

The Swedish Exomes project contains 12,380 samples, of which there are 4,969 patients with SCZ and 6,245 controls. The data was analysed after Phase I when approximately 2,500 cases had been sequenced (Purcell et al., 2014), when the project was completed in 2016 (Genovese et al., 2016) and the data has been included in a few other studies (Curtis et al., 2018; Nguyen et al., 2017; Singh et al., 2016, 2022)

5.1.3 Main Aim and Objectives

Main aim:

- Application the method from Chapter 4 to further groups of 100 patients and controls sampled from the Swedish Exomes project, to prioritise rare, likely-damaging variants that are unique to SCZ cases in the study. And, to identify common pathways affected by these variants.

The objectives are:

1. Filter all variants to the variants in CO/HP-genes analysed from Chapter 3.
2. Assign WES sample to random groups of 100 SCZ or 100 CTL groups.

3. Apply the steps described in Chapter 4 including predicting the coding consequence of the variants, filtering out common variants from gnomAD Exomes.
4. Identify consistent biological pathways enriched for rare variants across SCZ groups.

5.2 Methods

The methods for processing the data were followed as outlined in Chapter 4, with the following changes for handling multiple groups of 100 cases and groups of 100 controls.

5.2.1 Data

5.2.1.1 Swedish Exomes – Phenotype and Metadata files

The Swedish Exomes data was obtained as a compressed VCF file from dbGaP study phs000473.v2.p2 with permission granted for general research use (GRU). Subject phenotypes data (phs000473.v2.pht002600.v2.p2.c1. Schizophrenia_Swedish_Subject_Phenotypes.GRU.txt.gz, referred to as Phenotypes.txt.gz) and Sample data (phs000473.v2.pht002599.v2.p2.Schizophrenia_Swedish_Sample.MULTI.txt.gz, referred to as Sample.txt.gz) were downloaded from dbGaP via the NIH Run Selector (NCBI, n.d.) as previously described in 4.3.1.1.

5.2.1.2 Swedish Exomes – Variant call data

The Swedish Exome variant call data (VCF) was obtained via the NIH SRA Run Selector and converted to binary VCF format (`swe.bcf`) as previously described in 4.3.1.2.

5.2.1.3 GRCh37 Genomic Regions from genes discovered in Chapter 3

The list of common genes discovered in Chapter 3 was prepared as in 4.3.1.3. The merged BED file was used to supply the genomic regions for filtering the Swedish Exomes VCF data (in 4.3.2).

5.2.1.4 gnomAD Exomes

The gnomAD v2.1.1 Exomes (Karczewski et al., 2020) VCF was downloaded from gnomAD's Google Cloud Storage repository and converted to binary VCF format as previously described in 4.3.1.4.

5.2.2 Sample selection and subsampling into groups of 100 subjects and controls

The phenotypes data (Phenotypes.txt.gz) includes a variable that shows whether the sample exceeded 20X coverage at 80% of sites on average (Table 5.1). Samples that did not meet this were filtered out ('`Coverage_Pass=="N"`').

Table 5.1 Subjects passing the Coverage criterion

	Number of Subjects	Coverage_pass > 80% at 20X	Number of groups	Coverage_pass < 80% at 20X
Schizophrenia	4969	3582	35	1387
Controls	6245	4879	48	1366

35 SCZ groups were randomly sampled from 3500 of the remaining SCZ subjects (including the group of 100 SCZ subjects from Chapter 4). 48 CTL groups were sampled from 4800 of the CTL subjects. All groups contained exactly 100 subjects. Groups of 100 subjects were generated from the phenotype and samples data using an R script (GitHub: Chapter4\01-SwedishExomes-DataSelection.R). The Phenotypes file and the Samples file (Samples.txt.gz) were opened in R, decompressed in the script using `gzfile()`, and read into tables: phenotypes and samples. A new column was added to the phenotypes table to contain the group number, which was composed of the subject group (SCZ or CTL) and a three-digit number (for example, SCZ000 as group 0 and SCZ034 as group 34). 100 subjects with `primary_disease="Schizophrenia"` were randomly assigned to each SCZ group using the `sample()` command, and 100 subjects with `primary_disease="Control"` were randomly assigned to each CTL group. The phenotypes table was joined to the samples table by subject ID (`subjid`). Sample IDs in each group were written to a file (for example, `SCZ001.sampid.args` and `CTL047.sampid.args`) to be used in the next step. The phenotypes table, updated phenotypes table with group assignment and samples table were saved in an RData file. The `sampid.args` files containing the samples assigned to each group were stored for reference.

The Swedish Exomes VCF file (converted to BCF in 4.3.1.1.2) was partitioned into the respective groups of 100 subjects using the `view` command from bcftools v1.11 (Danecek & McCarthy, 2017) within a bash script to cycle through each group (GitHub: Chapter4\02-SwedishExomes-SplitVCFtoGroups.sh). `bcftools view` was run specifying the `sampid.args` files containing the 100 Sample IDs to be extracted, and the genomic regions in `merged.sort.37.bed` (see 4.3.1.2) to create a `bcf` and `bcf.csi` index file for each group. For example, using `SCZ001.sampid.args`, the files `SCZ001.bcf` and `SCZ001.bcf.csi` were created. This BCF file contain all the variants in the CO/HP-genes from each of the 100 exome samples assigned to group SCZ001.

5.2.3 Annotation of variants using BCFtools Consequences (csq)

All variants in each group's VCF file were annotated using BCFtools' Consequences (csq) function as previously described in 4.3.3 (GitHub: Chapter4/03-VariantConsequence.sh). BCFtools `csq` adds the BCSQ annotation containing the predicted consequences to each variant. For each input group's `bcf` file BCFtools `csq` created an annotated output binary file (for example, `SCZ000.csq.bcf`). A list of the worst consequences per group including potential multinucleotide variants (MNV) for each variant (`SCZ000.worst.missense.txt`) was extracted using BCFtools plugin `+split-vep`.

5.2.4 Annotation of variants using gnomAD v2.1.1

The variants in each group's VCF file were annotated with a set of metrics from gnomAD Exomes v2.1.1 (see 4.3.1.3 above) using BCFtools `annotate` function (GitHub: Chapter4/04-gnomadAnnotation.sh) as described previously in 4.3.4. For each group's `csq.bcf` input file a new binary VCF file annotated with gnomAD Exome allele frequencies and counts was output and indexed (for example, `SCZ000.gnomad.bcf`).

The general landscape of sample genotypes (for all alleles and for rare alleles) was analysed for variance and outliers in the groups. Statistics of variants that do not appear in, or with an allele count of 5 or less in gnomAD Exomes (option: `-i'GT="alt" && (gnomAD_AC<5 || gnomAD_AC=".")'`) were extracted from each group's annotated `gnomad.bcf` file using the BCFtools plugin `+smpl-`

`stats`. Statistics of all variants in the annotated BCF file were also extracted using `+smpl-stats`. The total number of variants per sample, number of samples heterozygous or homozygous for the alternate allele were analysed and plotted together in box plots in R.

The following code output the number of the group and all sample table data to a single text file, `SCZ.all.smpl.stats`. The file was opened in Excel and a new column was added to label each line with the corresponding group ID (for example, "SCZ000" in column 1):

```
for f in {000..034}
  do echo ${f} >> SCZ.all.smpl.stats
  grep '^FLT0' SCZ${f}.all.smpl.stats >>
SCZ.all.smpl.stats
done
```

This was repeated for all 48 CTL groups.

5.2.5 Normalisation of variants

Normalisation of variants in a VCF file, realigned variants, split multiallelic sites into biallelic sites and was achieved using BCFtools `norm` function (GitHub: Chapter4/05-VariantRare.sh) as described previously in 4.3.5. In this step, for each group's gnomAD annotated `gnomad.bcf` file, variants were normalised, and common variants filtered out producing a `vcf` file of rare variants (`SCZ000.rare.vcf`).

5.2.6 Pathogenicity of Rare Variants – Ensembl Variant Effect Prediction

The BCF files with rare variants (`SCZ000.rare.vcf`) were annotated with the results of an offline, cached GRCh37 version of Ensembl's Variant Effect Predictor (VEP; v101) (McLaren et al., 2016) as described previously in 4.3.7 with additional prediction tools described in Table 5.2.

Table 5.2 Plugins used from Ensembl VEP
Variants were scored using plugins within VEP. The Threshold value was used to filter damaging variants in R (see section 5.2.8).

Plugin	Description	Threshold
SIFT	SIFT (Sorting Intolerant From Tolerant) is a computational tool that predicts the functional impact of amino acid substitutions on proteins. Using sequence homology and protein properties, SIFT identifies potentially damaging variants. With a reported accuracy of 82%, SIFT is suitable for assessing both rare and common protein-coding variants (Kumar et al., 2009; Ng, P. C. & Henikoff, 2001)	< 0.05
PolyPhen2 HumDiv	PolyPhen-2 (Polymorphism Phenotyping v2) is a computational tool that predicts functional effects of human non-synonymous SNVs. The HumDiv model of PolyPhen-2 is trained on damaging alleles causing Mendelian diseases and distinguishes severe mutations from benign variants. HumDiv provides a probability score and categorizes variants as benign, possibly damaging, or probably damaging. PolyPhen-2 exhibits high accuracy, with a 92% true positive rate on the HumDiv dataset at a 20% false positive rate. HumDiv is particularly suitable for assessing rare damaging alleles due to its focus on severe disease-causing mutations (Adzhubei, Ivan A. et al., 2010)	> 0.9
REVEL	REVEL (Rare Exome Variants Ensemble Learner) is an ensemble computational method that predicts the pathogenicity of rare missense variants. It combines scores from 13 individual prediction tools, including MutPred, FATHMM, VEST, PolyPhen-2, SIFT, and others. REVEL's ensemble approach enhances the accuracy of predicting variant pathogenicity, making it a valuable resource for assessing the impact of rare variants on protein function and their potential contribution to disease phenotypes (Ioannidis et al., 2016).	> 0.5

Plugin	Description	Threshold
dbscSNV v1.1	dbscSNV (Database of Splicing Consensus SNVs) is a database for predicting the splicing impact by Ada Boost and Random Forest methods focussing on potential human SNVs within splicing consensus regions of coding exons. These regions include the 5' splice site (from -3 to +8) and the 3' splice site (from -12 to +2).	Ada: > 0.9 RF: > 0.9

For each group's input `rare.vcf` file, VEP was supplied the FASTA for GRCh37 primary assembly to predict the codon and amino acid changes. The chromosome and position were kept in the output rather than converting to a RefSeq transcript reference (`--use_given_ref`). For each file processed, a summary file was produced (`--stats_text -sf`) and VEP output a new plain text VCF file (for example, `SCZ000.vep.vcf`). The following command and switches were run:

```
/VEP/vep -i SCZ000.rare.vcf --vcf --cache --offline --dir_cache $VEP --merged
      --fasta 101_GRCh37/Homo_sapiens.GRCh37.75.dna.primary_assembly.fa.gz -
      -assembly GRCh37
      --plugin dbscSNV,$VEP/dbscSNV/dbscSNV1.1_GRCh37.txt.gz
      --plugin REVEL,file=$VEP/REVEL/new_tabbed_revel.tsv.gz,no_match=1
      --sift b --polyphen b --humdiv --symbol –biotype
      --use_given_ref --stats_text --fork 10
      -o SCZ000.vep.vcf -sf SCZ000.vep.html --force-overwrite
```

Properties (position, the reference and alternative alleles, the gene name, Ensembl IDs, allele frequencies and counts from gnomAD, the exon, intron, position in the coding sequence and protein sequence, codon, and amino acid changes) and predicted scores for each biallelic variant with a VEP (`-a CSQ`) prediction of splicing region or worse were extracted using BCFtools' `+split-vep` plugin to a tab-delimited file (`SCZ000.vep.worst.txt`):

```
fStr='%CHROM\t%POS\t%REF\t%ALT\t%IMPACT\t%Allele\t%Gene\t%Featu
      re\t%Feature_type\t%AC\t%AN\t%gnomAD_AC\t%gnomAD_AN\t%gno
      mAD_AF\t%gnomAD_nhomalt\t%SYMBOL\t%ID\t%BIOTYPE\t%EXON\t%
      INTRON\t%cDNA_position\t%CDS_position\t%Protein_position\
      t%Amino_acids\t%Codons\t%DISTANCE\t%Consequence\t%ada_sco
      re\t%rf_score\t%SIFT\t%PolyPhen\t%REVEL\n'
```

```
bcftools +split-vep -f$fStr SCZ000.vep.vcf -a CSQ -s
      worst:splice_region+ -d
      > SCZ000.vep.worst.txt
```

5.2.7 Pre-processing of Variants in R

Statistics from summary files were imported into R for analysis.

- Ingest all VEP files with custom VEP reader. (Add R file to GitHub)
- Extract summary values from sections in the VEP statistics file
- Plot graphs of statistics across groups

All `vep.worst.txt` files (35 SCZ files and 48 CTL files) were imported into a single data table (`scores.all.vep`) in R for pre-processing, with the filename added into a new id column.

```
files.scz.vep <- list.files(path = "05_vep_worst_plus/",
      pattern = "SCZ[0-9]{3}.vep.worst.txt", full.names = T)
scores.scz.vep <- rbindlist(sapply(files.scz.vep, fread,
      simplify = F), idcol = 'filename')
files.ctl.vep <- list.files(path = "05_vep_worst_plus/",
      pattern = "CTL[0-9]{3}.vep.worst.txt", full.names = T)
scores.ctl.vep <- rbindlist(sapply(files.ctl.vep, fread,
      simplify = F), idcol = 'filename')
scores.all.vep <- rbind(scores.scz.vep, scores.ctl.vep)
```

The following processing steps were carried out on the variant data:

1. Paths and suffixes were removed from the filename column to produce the group Id (for example, `CTL023.vep.worst.txt -> CTL023`).
2. A Group column was added as a factor (SCZ or CTL) derived from the Group ID.
3. Column names were set to appropriate names.
4. Chromosome X entries were edited to ‘23’ and the column was converted to a numeric column.
5. The following columns were converted to factors: Impact, Filter, Culprit, gene_name, Biotype, and Consequence.

6. Variant quality fields were converted to numeric if required (NA values were introduced if values are missing): QUAL, quality of depth (QD), Fisher's Strand bias (FS), VQSLOD, and mapping quality (MQ).
7. gnomAD AF/AC/AN and nhomalt fields were converted to numeric with NAs for missing values.
8. Many sites passed the filters with low allele counts in the groups, but existed in all groups, CTL and SCZ. The difference between gnomAD allele count and number was calculated into a new column gn_diff and variants were removed where there were excessive numbers in gnomAD. For example, all variants with rsIDs were mostly homozygous for the ALT allele - rs77262002, rs4962173, rs5951328 - were removed. This removed 171 rows for the three variants.
9. All variants with an allele count greater than 11 (per group) and a difference between gnomAD AC and AN of 35 or greater were removed from the dataset.
10. Two new columns for minor allele count and allele frequency were calculated where either the REF or the ALT allele had the lower frequency in the population, using the following formulae (median AC = 1.000):
    a. minor_AC: if AC < 11, minor_AC = AC, else minor_AC = AN – AC.
    b. minor_AF: if AC < 11, minor_AF = gnomAD_AF, else minor_AF = 1 – gnomAD_AF
11. A new column was created, has_rsID with 1 if the rsID column contained an rsID or 0 if rsID = "." (missing value).
12. The column, Exon, had either "." For intronic variants or "n/m" for the *n*th exon of m exons. A new column exon_type was created with values of 'dot', 'AofB' for each of the value types. The column, intron, holding the nth value of m introns for intronic variants, was similarly used to encode intron_type. For AofB values, Exon values were split into two new numeric columns, Exon1 and NumExons and Intron values were split into two new numeric columns, Intron1 and NumIntrons.
13. The columns cDNA_position, CDS_position and Protein_position contained either single positions or start-end ranges. Some values are only contained ".". A new factor column for each (cDNA, CDS and Protein) was created with

the levels: “dot”, “range”, and “single”. Two further numeric columns were created for each, with suffixes “_start” and “_end”. For ranges, start and end values were parsed from the original fields. For single positions, start and end were set to the same value.

14. A new factor column was created, AA_effect to store a single interpretation of the Amino_acids column. If the value of Amino_acids was “.” then AA_effect = “none”. If the value of Amino_acids began or ended with “-“, then AA_effect = “insertion” or “deletion”, respectively. If Amino_acids ended with a “*” then AA_effect = “truncation”. If there was exactly one base change, for example A/G, then AA_effect = “missense” and if there was only a single base, AA_effect = “synonymous”. Any remaining values were ascribed as “indel”.
15. dbscSNV scores Ada_score and RF_score and the REVEL column were converted to numeric columns. Missing values were set to NA.
16. SIFT values were in the format “deleterious(0.01)” and were split into new columns: a factor, SIFT_level (“deleterious”) and a numeric SIFT_score (0.01). Rows with no entry were assigned “unknown” and 1.
17. PolyPhen values were in the format “pathogenic(1)” and were split into new columns: a factor, PolyPhen_level (“pathogenic”) and a numeric PolyPhen_score (1). Rows with no entry were assigned “unknown” and 0.

5.2.8 Filter variants to SCZ-only, predicted-pathogenic variants

Variants were analysed using various aggregation methods – by group Id (SCZ000, CTL000), by gene, by coding consequence, or by predicted pathogenicity scores. The full list of variants included several rare variants that were found in almost equal numbers across control groups as SCZ groups. On inspection, many of these appeared to be potential false positives with poor depth or quality, with sequencing errors, alignment errors, or other technical errors. Some may be true positives and could be rare variants that predispose towards a SCZ phenotype depending on the existence of another mutation (epistasis). Here the focus was on damaging variants that only affect patients with SCZ. All variants that affected CTL groups and SCZ groups were removed from the data. This also had the effect of removing many of the variants with lower scores for Fisher’s strand bias, read depth and other technical errors.

All variants in *scores.all.vep* were grouped by chromosome, position, REF and ALT and the value of minor allele count was totalled across all CTL groups. Variants with zero records in CTL groups were assigned to *scores.all.scz* using the following R code:

```
scores.bypos.all.vep <- as.data.table(scores.all.vep %>% group_by(Chrom,
Position, REF, ALT) %>%
    summarise(total_AC = sum(minor_AC),
        AC_CTL = sum(minor_AC[Group=="CTL"]),
        AC_SCZ = sum(minor_AC[Group=="SCZ"]),
        AC_SCZ_wt = as.integer(sum(minor_AC[Group=="SCZ"]) * 48 / 35)))
scores.bypos.all.vep[, AC_diff := AC_SCZ_wt - AC_CTL]
scores.bypos.all.vep.scz <- scores.bypos.all.vep[AC_CTL==0,]
scores.all.scz <- scores.all.vep[scores.bypos.all.vep.scz, on = c("Chrom",
"Position", "REF", "ALT")]
(scores.all.scz <- droplevels(scores.all.scz))
```

Variants were retained if they were predicted to be pathogenic by one or more predicted predictive tools from VEP. SIFT scores below 0.05 were considered as evidence for being deleterious; PolyPhen2 HumDiv scores above 0.9 were considered pathogenic; REVEL scores above 0.5, and scores for RF-score and Ada-score from dbcsSNV above 0.9 were considered as disrupting consensus splice sites. Furthermore, variants that led to an introduction of the premature stop codon, removed a start codon, or introduced a frameshift were considered damaging and retained.

R code for filtering pathogenic variants:

```
scores.all.scz.pathogenic <- scores.all.scz %>%
        filter(PolyPhen_score > 0.9 |
            SIFT_score < 0.05 |
            REVEL > 0.5 |
            Ada_score > 0.9 |
            RF_score > 0.9 |
            grepl("stop_gain", Consequence) |
            grepl("start_lost", Consequence) |
            grepl("frameshift", Consequence))
```

This final set of variants contained ultra-rare variants, with an allele count under 5 (median 1) in gnomAD, and either predicted to be pathogenic/deleterious or whose coding consequence would lead to a loss of function or haploinsufficiency though nonsense-mediated decay.

The number of variants remaining per group was calculated and plotted to show the distribution across the SCZ groups.

5.2.9 Comparison with Genome-Wide Association Studies (GWAS)

To assess the overlap of the filtered rare variants in the SCZ groups with the loci associated with SCZ from the PGC Schizophrenia Working Group's latest GWAS (Trubetskoy et al., 2022), loci data were extracted from the Supplementary Materials. These were imported into R as a data table with 5 columns (Chrom, Start, End, odds ratio (OR), and P value (P)). For each GWAS locus, an expanded locus was computed with a buffer of 200 Kb before and after. For each locus and buffered locus, the total number of RVs (minor_AC) were totalled, and the genes affected were collected. The study identified 628 loci with a p value less than 1×10^{6}, including 287 genome-wide significant ($p < 5 \times 10^{8}$) loci. Histograms per chromosome were plotted in R of total minor_AC within bins of 200,000 bases along the length of each chromosome. Genome-wide significant loci were shaded as areas extending the full height of the graph on the graph in light red; non-significant loci were shaded in light blue.

5.2.10 Variant Pathway Analysis – consistent pathways across groups

- All genes with variants were collated as Ensembl gene accession numbers (without version).
- All GO IDs for these genes were retrieved using Ensembl biomaRt in R (version 101).
- For each GO ID the name, description and namespace were retrieved.
- Summary statistics per group were assigned to each GO ID – total minor_AC, a list of genes covered by the GO term.
- The top 50 GO terms in each namespace (Biological Process, Molecular Function, Cellular Component) were extracted containing the most variants / highest percentage of variants per SCZ case.

In individual GO terms with many genes, there were often subsets of genes from more specific biological pathways. Genes from individual GO terms were used as an input into STRING-DB v12 (Szklarczyk et al., 2021). STRING provides an interactive network graph of the genes, and further gene enrichment analyses results from GO, KEGG, Reactome, WikiPathways, and the protein domain databases, InterPro and Pfam.

5.3 Results

The Swedish Exomes project sequenced the whole exomes from 12,380 individuals, including 4969 SCZ, 1166 BPD and 6245 controls. In this study, SCZ and CTL exomes with more than 20X coverage in 80% of the sequenced sites were included. 35 groups of 100 SCZ subjects were sampled at random from 3582 SCZ samples with no sample appearing more than once. 48 groups of 100 controls were sampled at random from 4879 control samples. Variant sites were filtered to the 1074 CO/HP-gene regions from Chapter 4. Variants were annotated with data from variant effect prediction tools and with allele frequencies from gnomAD, a large population study.

Methods detailed in Chapter 4, with some modifications, were applied to all groups and compared.

5.3.1 Analysis of SCZ and CTL groups

Figure 5.1 shows a plot of average per-sample genotype counts and average counts of SNVs and indels across all SCZ and CTL groups. Among the SCZ groups there were an average of 11,916 genotyped sites per sample (blue in Figure 5.1) that passed the quality filter (80% of the genotypes had greater than 20X coverage). There were many SNV sites at which one or more individuals carried an alternate allele, but the remainder of the cohort did not. There were on average 9,166 sites (76.9%) homozygous for the REF allele (dark blue in Figure 5.1) per group. Homozygous REF sites were strongly influenced by the number of RVs and singletons, increasing by 99 for each singleton in a group. On average per SCZ sample across all groups, there were, 2,749 non-REF genotypes (23.1%; yellow), of which 1,030 were homozygous for the ALT allele (8.6% of all genotyped sites; red in Figure 5.1) and 1,719 were heterozygous (14.4%; grey in Figure 5.1), and 342 missing genotypes (`./.`, `.`, or `./0`) (2.9%; data not shown in Figure 5.1). There was an average of 288 indels (light blue) and 40 singletons (green) per SCZ sample per group.

Among the 48 CTL groups (right-hand side of Figure 5.1) there were, on average, 11,863 genotypes per sample that passed the quality filter. On average per CTL sample across all groups, there were 9,112 homozygous REF genotypes (76.8%; dark blue), 2,751 non-REF genotypes (23.2%; yellow), of which 1,028 were homozygous for the ALT allele (8.7%; red) and 1,724 were heterozygous (14.5%; grey), and 320 missing genotypes (2.7%; data not shown in Figure 5.1). There was an average of 289 indels (light blue) and 40 singletons (green) per CTL sample per group.

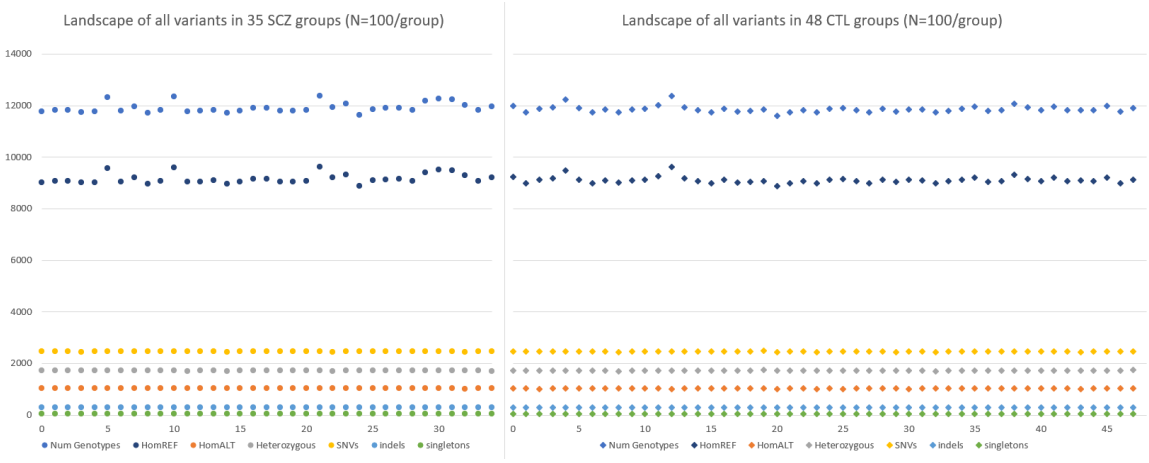

Figure 5.1 Plot of average counts of genotypes per subject across 35 SCZ groups and 48 CTL groups
For each of the 35 SCZ and 48 CTL groups of 100 subjects, average counts were plotted for the number of genotypes (blue), the number of genotypes homozygous for the REF allele (dark blue), non-REF sites (yellow), heterozygous (grey) and heterozygous (red) non-REF, indels (light blue) and singletons (green).

Overall, each sample has approximately 2,750 common or rare SNVs or indels within the 1074 CO/HP-genes. At this resolution it was not possible to see large differences between SCZ samples and CTL samples. These SNVs and indels could include many common variants seen in all patients, technical errors from the sequencing processes or misalignments during the mapping stage.

5.3.2 Effect of filtering with gnomAD Exomes

Variants were annotated with allele frequencies from gnomAD Exome v2.1.1 data. Figure 5.2 shows the per-sample statistics for the remaining variants after filtering all variants with an allele count of more than 5 in gnomAD. In CO/HP-genes, among the SCZ groups there remained an average of 170 genotyped sites per sample (dark blue

in Figure 5.2) that passed the quality filter (80% of the genotypes had greater than 20X coverage) and all were non-REF genotypes. The remaining variants were either homozygous (mean: 29; red in Figure 5.2) or heterozygous (mean: 142; grey in Figure 5.2) for the ALT allele. The variants were either indels (mean: 145; light blue) or SNVs (mean: 26; yellow in Figure 5.2). There was an average of 7 singletons (range: 0 to 60; green in Figure 5.2) per SCZ sample per group after filtering, with only 4 SCZ samples with no singletons and 4 SCZ samples with over 50 singletons each.

Among the CTL groups there remained an average of 172 genotyped sites per sample (dark blue in Figure 5.2) that passed the quality filter (80% of the genotypes had greater than 20X coverage) and all were non-REF genotypes. The remaining variants were either homozygous (mean: 29; red in Figure 5.2) or heterozygous (mean: 143; grey in Figure 5.2) for the ALT allele. The variants were either indels (mean: 147; light blue) or SNVs (mean: 26; yellow in Figure 5.2). There was an average of 7 singletons (range: 0 to 66; green in Figure 5.2) per CTL sample per group after filtering, with 12 CTL samples with no singletons and 4 CTL samples with over 50 singletons each.

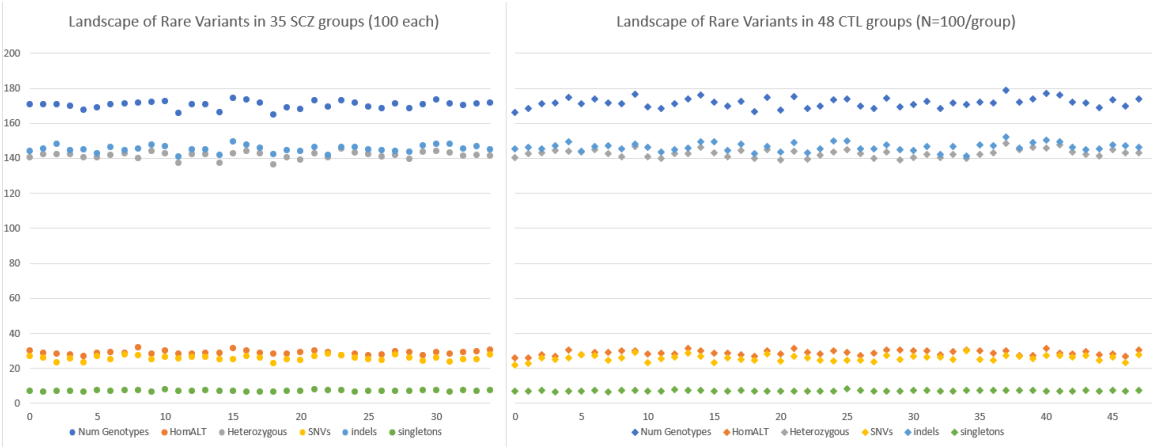

Figure 5.2 Plot of average counts of genotypes for rare variants only across all groups
For each of the 35 SCZ and 48 CTL groups of 100 subjects filtered to rare variants (and AC<5 in gnomAD), average counts were plotted for the number of non-REF genotypes (dark blue). These genotypes were split between homozygous (red) and heterozygous (grey) non-REF genotypes, or between indels (light blue) and SNVs (yellow). Singletons are in green.

Filtering common variants removed all homozygous REF alleles and left mainly heterozygous indels and on average 26 SNVs with a range of 13 to 76 SNVs per SCZ

sample. Figure 5.3 through Figure 5.6 depicts the distribution of RVs per chromosome in each group.

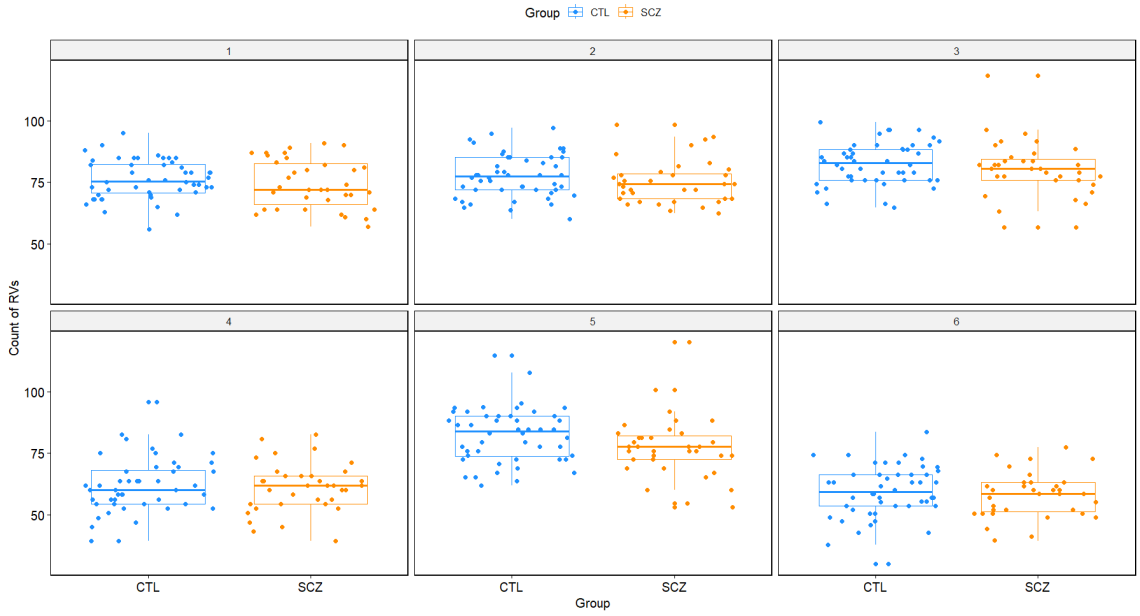

Figure 5.3 Box plot of SCZ and CTL groups in chromosomes 1-6
Individual points are plotted to help identify outlier groups that have a higher burden of rare variants. Counts are normalised to the approximate number of genes relative to chromosome 1. SCZ: Orange; CTL: Blue

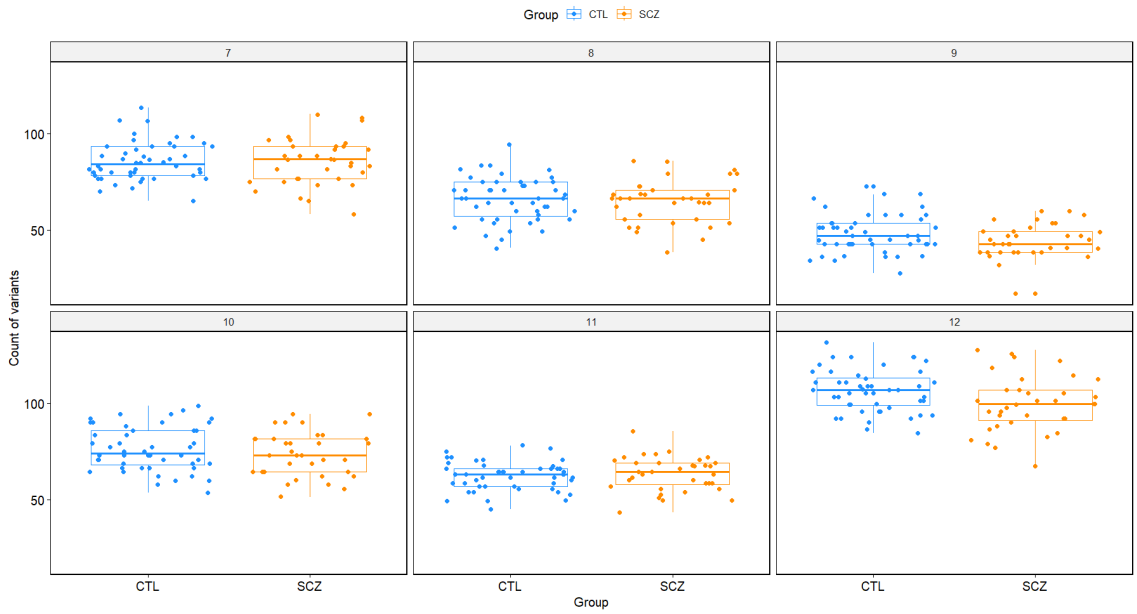

Figure 5.4 Box plot of SCZ and CTL groups in chromosomes 7-12
Individual points are plotted to help identify outlier groups that have a higher burden of rare variants. Counts are normalised to the approximate number of genes relative to chromosome 1. SCZ: Orange; CTL: Blue

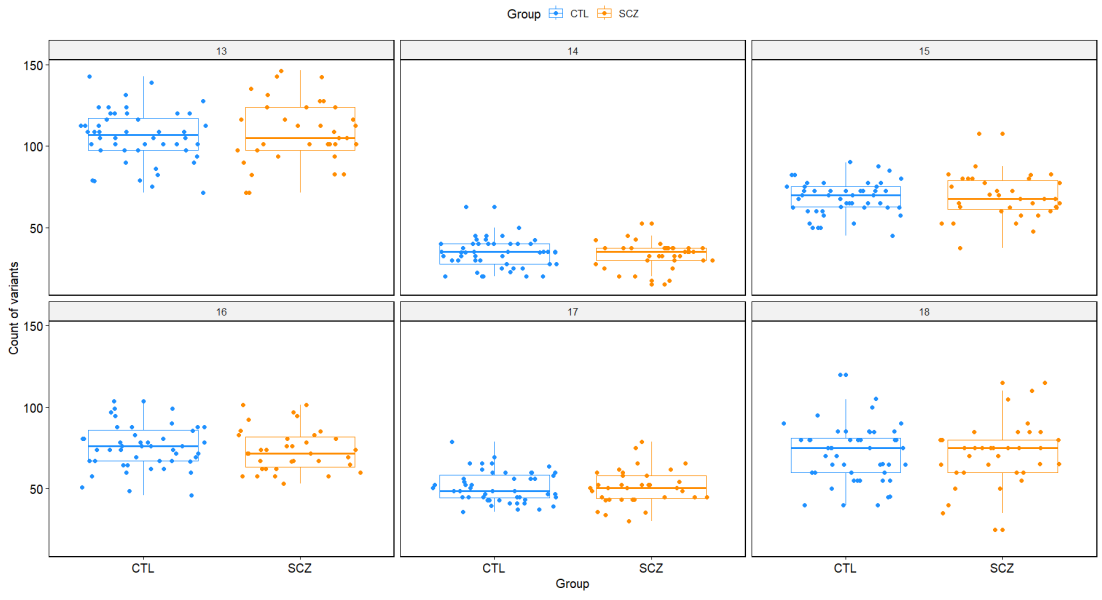

Figure 5.5 Box plot of SCZ and CTL groups in chromosomes 13-18
Individual points are plotted to help identify outlier groups that have a higher burden of rare variants. Counts are normalised to the approximate number of genes relative to chromosome 1. SCZ: Orange; CTL: Blue

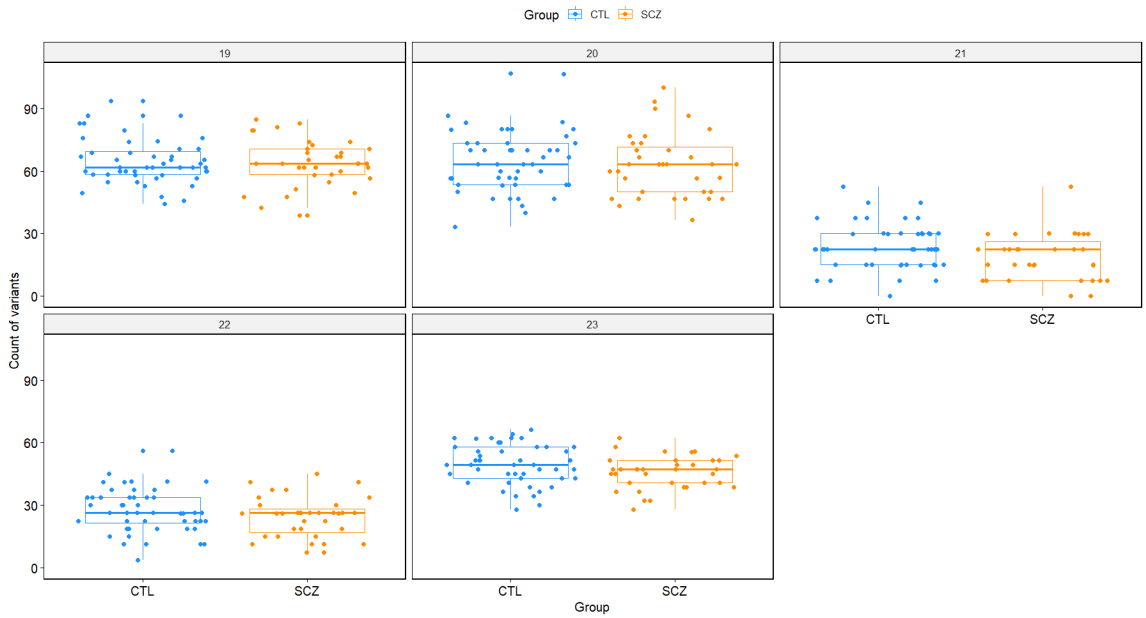

Figure 5.6 Box plot of SCZ and CTL groups in chromosomes 19-22 and X
Individual points are plotted to help identify outlier groups that have a higher burden of rare variants. Counts are normalised to the approximate number of genes relative to chromosome 1. SCZ: Orange; CTL: Blue

5.3.3 Analysis of Consequences for all groups

All variants were analysed using BCFtools *csq* function to identify potential multinucleotide variants in the complete dataset. There were approximately 119 MNVs detected per group. These were stored for reference when assessing the predicted effect of RVs in the following sections.

- Across all groups there were few differences of type of consequence / group that can be spotted at this stage
- It looks like there are more missense variants in CTL groups than SCZ – but missense may be interpreted differently if the missense variants are harmful in SCZ and not harmful in CTL.

5.3.4 Variant Effect Predictor

Ensembl's VEP was used to annotate rare variants from biallelic sites in all groups with scores from predictive tools, PolyPhen-2, SIFT, REVEL to assess the pathogenicity of missense variants in the coding sequence and dbscSNV's Adaptive Boosting (AdaBoost) and random forest (RF) scores for the effect of missense variants on consensus splicing

sites, as described in the following sections. VEP also provided the predicted effect on different transcripts and listed the codon changes and amino acid changes caused by the variant. VEP calculated these as stand-alone without considering any involvement in MNVs. The output of VEP included effects on all known transcripts in the gene. The worst effect on any transcript was extracted for each variant from the VEP output for analysis in R – including splice region variants, missense variants, stop gain, start lost, frameshift, in-frame insertions and deletion, and splice site variants.

Overall, there were 26,976 unique SNVs and indels with a total allele count of 29,199 among the 8,300 individuals in the 48 CTL groups and 35 SCZ groups.

5.3.4.1 PolyPhen2 scores

Ensembl VEP assigned PolyPhen-2 HumDiv scores to all missense variants in each group. The HumDiv score derives from a homology-based comparison of the amino acid sequence with a wide range of homologous proteins. A higher HumDiv score indicates a more conserved position, suggesting that a substitution may be more likely to disrupt function. PolyPhen-2 HumDiv combines this score with other physiochemical properties of the amino acids being substituted, and potential structural changes (Adzhubei, Ivan A. et al., 2010). PolyPhen-2 assesses the functional impact an amino acid substitution using a score from 0 (benign) to 1 (damaging). Scores closer to 1 (0.9 and above) represent the highest confidence in a damaging prediction by PolyPhen-2 but does not guarantee the variant is damaging. The current version of PolyPhen-2 (2.2) HumDiv scores have a false positive rate of 5-10%. Figure 5.7 depicts the number of missense variants in each SCZ and CTL group predicted by PolyPhen-2 HumDiv to be damaging (scores above 0.9 threshold). The CTL groups exhibited a wider range of variants predicted to be damaging (76 to 129) compared to SCZ groups (73 to 118).

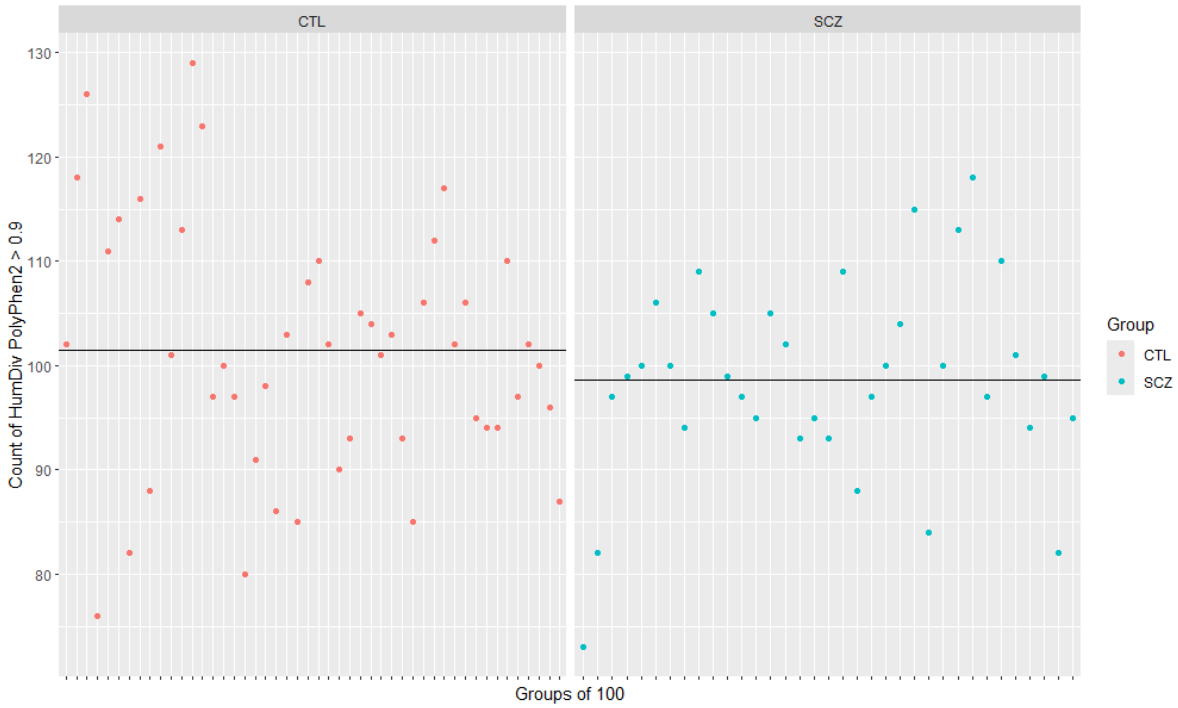

Figure 5.7 Count of pathogenic PolyPhen-2 HumDiv scores across all groups
Plot showing the number of variants per group predicted to be pathogenic using PolyPhen-2 HumDiv in Ensembl's VEP. A threshold of 0.9 was chosen for HumDiv scores. The 48 control groups on the left (red) run from CTL000 to CTL047. The 35 SCZ groups on the right (blue) run from SCZ000 to SCZ034. The mean number of variants is shown for CTL and SCZ (horizontal line).

5.3.4.2 SIFT (Sorting Intolerant from Tolerant) scores

SIFT scores were also assigned to all rare missense variants in each CTL and SCZ group. SIFT also evaluates the functional impact of missense variants by comparing the physiochemical properties of the original and substituted amino acids at the variant site. Scores closer to 0 (less than 0.05) indicate a higher predicted deleteriousness, suggesting a stronger potential for the protein to disrupt protein function. Figure 5.8 depicts the number of variants in each CTL and SCZ group predicted by SIFT to be deleterious (scores below 0.05 threshold). CTL groups exhibited a wider range of variants predicted to be deleterious (92 to 144) compared to SCZ groups (92 to 137).

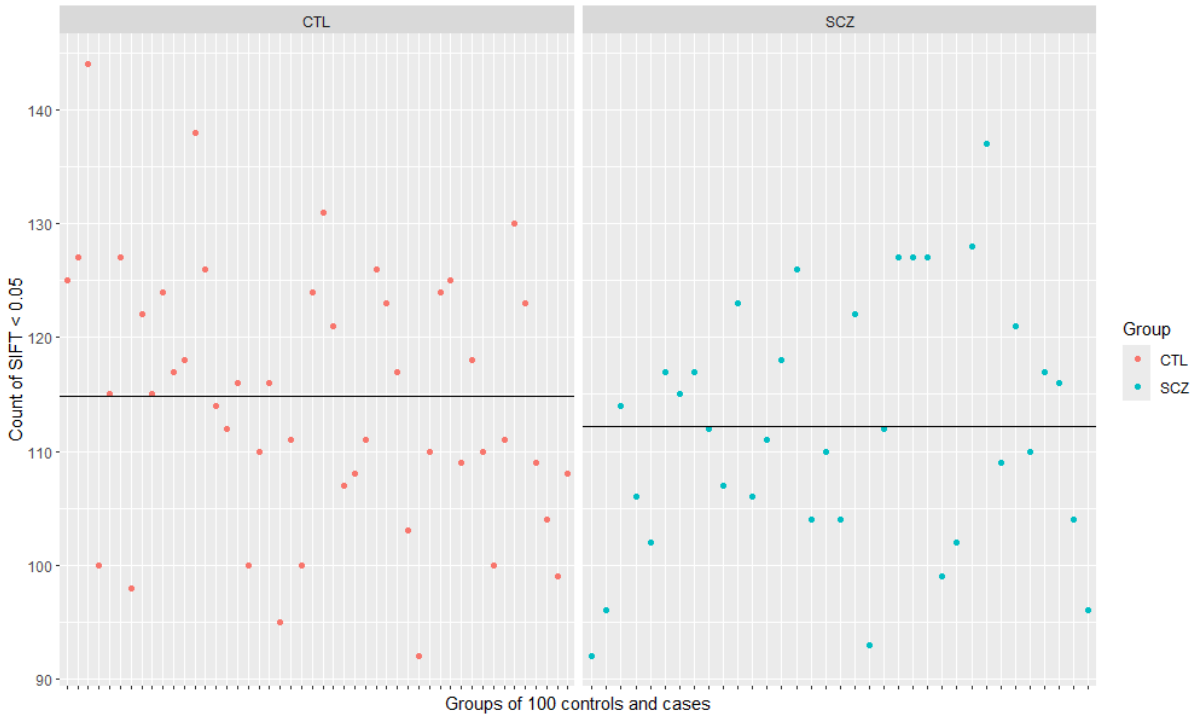

Figure 5.8 Count of SIFT scores per group
Plot showing the number of variants per group predicted to be pathogenic using SIFT in Ensembl's VEP. An upper threshold of 0.05 was chosen. The 48 control groups on the left (red) run from CTL000 to CTL047. The 35 SCZ groups on the right (blue) run from SCZ000 to SCZ034. The mean number of variants is shown for CTL and SCZ (horizontal line).

5.3.4.3 Rare Exome Variant Ensemble Learner (REVEL) scores

Ensembl VEP assigned REVEL scores to all rare missense variants in the CTL and SCZ groups. REVEL uses a machine learning model that combines evidence from several sources into a single score, ranging from 0 (neutral) to 1 (deleterious). Figure 5.9 shows the number of variants with a score above a threshold of 0.5 in CTL and SCZ groups. Scores were generally similar between CTL (mean: 44.1; range: 28 to 56) and SCZ groups (mean: 42.4; range: 24 to 53).

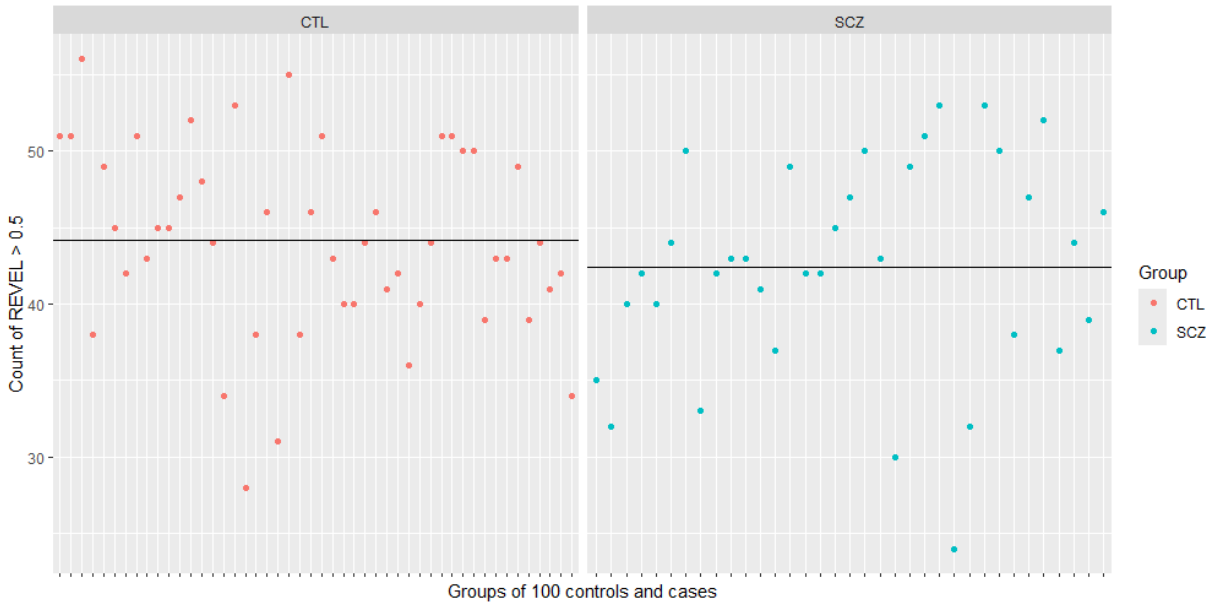

Figure 5.9 Count of REVEL scores per group
Plot showing the number of variants per group predicted to be pathogenic (> 0.5) using REVEL in Ensembl's VEP. The 48 control groups on the left (red) run from CTL000 to CTL047. The 35 SCZ groups on the right (blue) run from SCZ000 to SCZ034. The mean number of variants is shown for CTL and SCZ (horizontal line).

5.3.4.4 Database of Splicing Consensus SNVs (dbscSNV) scores

Ensembl VEP assigned both Ada scores and RF scores from dbscSNV to all CTL and SCZ groups. dbscSNV uses ensemble machine learning methods to produce scores between 0 and 1 for SNVs in splicing consensus regions flanking exon-intron boundaries that are critical for proper splicing of the pre-mRNA into mature mRNA (Jian et al., 2014). A strict threshold of 0.9 was selected for both Ada and RF scores. As shown in Figure 5.10 and Figure 5.11, CTL and SCZ groups had almost equal numbers of SNVs predicted to disrupt splicing sites by AdaBoost (CTL: mean = 10.2; range = 3 to 22, and SCZ: mean = 10.1; range = 1 to 17) and by Random Forest (CTL: mean = 5.85; range = 1 to 13, and SCZ: mean = 5.97; range = 0 to 12).

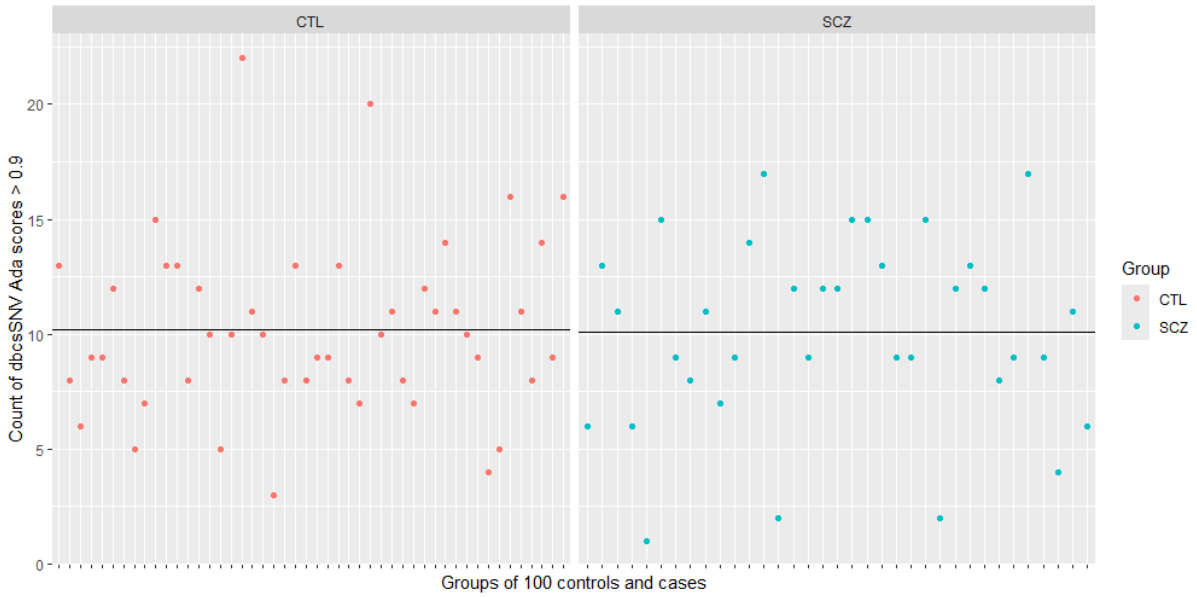

Figure 5.10 Count of dbscSNV AdaBoost scores per group
Plot showing the number of variants per group predicted to disrupt splicing consensus regions (> 0.9) using dbscSNV's Ada scores in Ensembl's VEP. The 48 control groups on the left (red) run from CTL000 to CTL047. The 35 SCZ groups on the right (blue) run from SCZ000 to SCZ034. The mean number of variants is shown for CTL and SCZ (horizontal line).

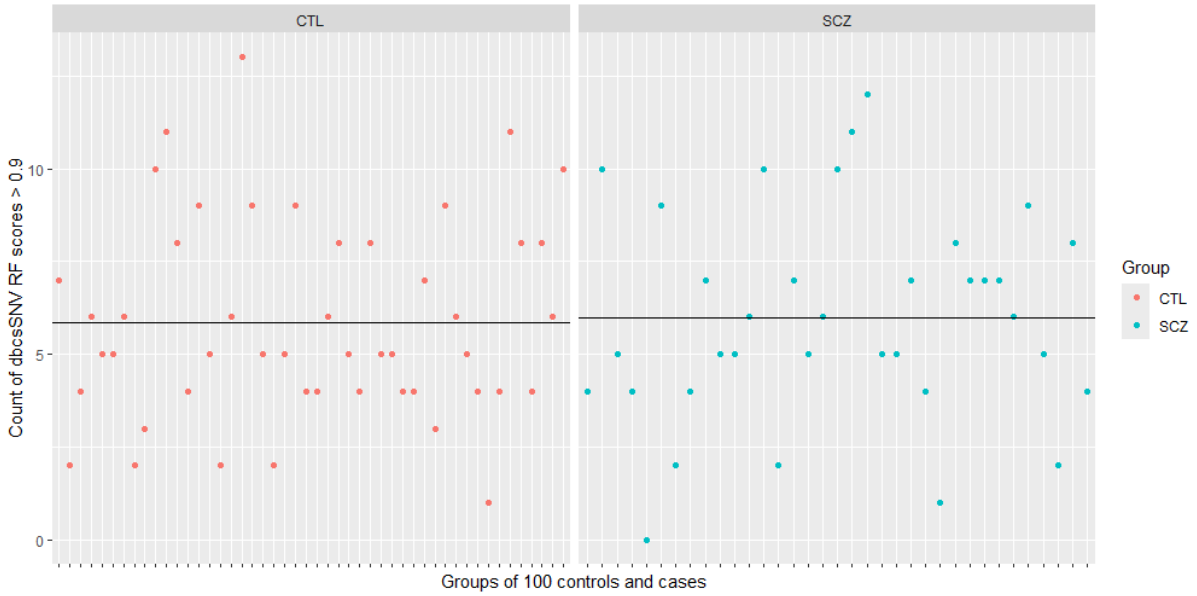

Figure 5.11 Count of dbscSNV Random Forest scores per group
Plot showing the number of variants per group predicted to disrupt splicing consensus regions (> 0.9) using dbscSNV's RF scores in Ensembl's VEP. The 48 control groups on the left (red) run from CTL000 to CTL047. The 35 SCZ groups on the right (blue) run from SCZ000 to SCZ034. The mean number of variants is shown for CTL and SCZ (horizontal line).

5.3.5 Reduction to ultra-rare, predicted damaging variants only affecting SCZ cases

VEP results classified as "splice region variant" or more severe consequences were initially imported into R software for all subject groups. This resulted in a total of 26,976 imported variants.

Following individual examination, a high number of variants in the VEP output were identified across many CTL groups that exceeded the individual group and gnomAD threshold of 5 for allele count. To enrich the dataset for variants potentially associated with schizophrenia, a two-step filtering process was implemented. First, variants observed exclusively in CTL groups or present in both CTL and SCZ groups were excluded. This process aimed to identify variants potentially unique to the disease state. Subsequently, the remaining 7,401 variants were filtered based on pathogenicity scores for missense or splicing variants by at least one of the aforementioned *in silico* prediction methods (SIFT, PolyPhen-2, REVEL, dbscSNV), or based on predicted effects that could disrupt protein translation (stop gain, start loss, frameshift).

This filtering process yielded a final list of 4,398 ultra-rare, predicted damaging variants (RDVs) affecting 4,459 SCZ cases. 940 of the 1074 CO/HP-genes with high expression in the Frontal Cortex and Hippocampus had at least one RDV. The distribution of these variants across all 35 SCZ groups, each containing 100 individuals, is depicted in Figure 5.12. Notably, all groups harboured at least 100 of these RDVs.

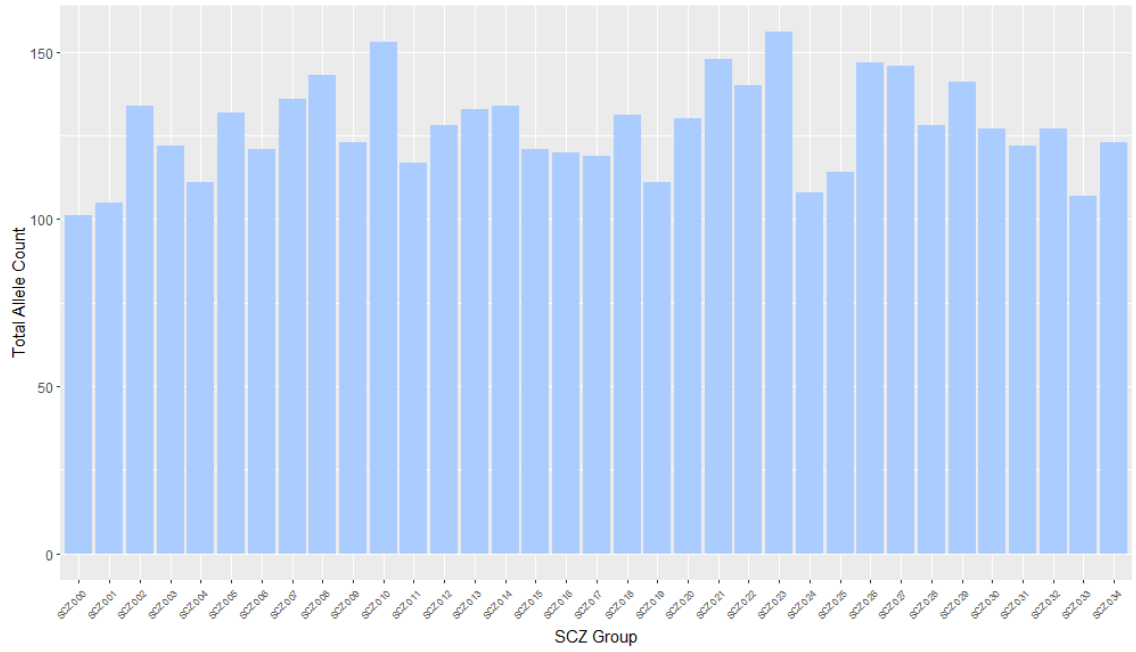

Figure 5.12 Bar chart showing the distribution of rare damaging variants per SCZ group
The total allele count was plotted for all SCZ groups.

The median allele count of the variants was 1, hinting that each individual within the SCZ group could harbour at least one such variant in a gene with high expression in the frontal cortex or hippocampus. Further analysis of these RDVs within the CO/HP-genes revealed an interesting pattern. While most SCZ subjects harboured at least one RDV, they rarely possessed multiple RDVs within the CO/HP-gene set. In 31 males there was an erroneous homozygous genotype for genes on the X chromosome (ARX, CITED1, DACH2, DRP2, FGF13, FRMPD4, GABRA3, GRIA3, LPAR4, MUM1L1, NUDT10, NUDT11, PCDH19, PDK3, PNMA5, PRRG3, RS1, SHROOM2, SLC7A3, SYTL5). Males have a single X chromosome, and the true genotype would be hemizygous with a single rare variant. This inflated the total allele count on chromosome X by 31.

RDVs were summarised per gene to analyse whether individual genes harbouring RDVs were consistent across the SCZ groups. Individual CO/HP-genes harbouring RDVs in the 3500 SCZ cases did not affect more than 1.2% of all cases. Genes with 1 to 17 RDVs across all 3500 cases included most of the CO-HP-genes with each gene being found in under 0.5% of cases (range 0.086% to 0.486%).

Table 5.3 shows the list of total minor allele counts (in descending order) and percentage of cases harbouring RDVs for each gene listed in the corresponding row. The known SCZ risk gene, CSMD1, harboured RDVs in 1.2% of all 3500 cases. DNAH11 harboured RDVs in 1.17% of all cases. GPR98 (ADGRV1), DNAH3, and TCHH harboured RDVs in more than 1% all cases each. At the lower end with only one RDV seen in all 3500 cases, the genes are still potentially important within specific functional networks. For example, the basic helix-loop-helix, BHLHA15, is a transcription activator that affects the expression of GRB14. Reduction of GRB14 is known to be a negative regulator of the insulin receptor. Variants in BHLHA15 leading to reduced GRB14 or variants in GRB14 could lead to lower blood glucose levels (Choi et al., 2023; Pillinger et al., 2017). A single RDV was identified in BHLHA15, and a further 3 RDV in GRB14 supported evidence of these variants affecting these genes involved in a pathway related to insulin/glucose homoeostasis. Genes with 1 to 17 RDVs across all 3500 cases included most of the CO-HP-genes with each gene being found in under 0.5% of cases (range 0.086% to 0.486%).

A full list of genes divided into groups by whether the RDVs are exclusive to 3500 cases or by how many more RDVs are in the genes compared to controls can be found in **Appendix 8.5**.

Table 5.3 CO/HP-genes with rare damaging variants by allele count
Variants were grouped by gene and summarised with total minor allele count, and the percentage of cases harbouring variants in these genes. AC: Allele Count.

AC	% cases	Genes
42	1.200	CSMD1
41	1.171	DNAH11
38	1.086	GPR98
37	1.057	DNAH3
36	1.029	TCHH
33	0.943	FRAS1
31	0.886	DNAH7
26	0.743	MAST4
24	0.686	LMO7
23	0.657	KALRN
22	0.629	TENM2
21	0.600	CHD5, FHOD3
20	0.571	FAT3

AC	% cases	Genes
19	0.543	CACNA1E, KIAA1549, LPA, SLIT1, TLN2
18	0.514	EPHA5
17	0.486	CIT, KIF21B
16	0.457	BAIAP3, KIAA1549L, MYO5B, ROS1, SYTL2
15	0.429	ARHGAP32, CCDC108, ITPR1, PTPRN2
14	0.400	DAPK1, FMN1, IQGAP3, OTOGL, PLXNA4
13	0.371	ABCA12, ATRNL1, COL24A1, COL5A3, DCDC1, EPHB6, FNDC1, GAS2L2, KCNMA1, LPHN2, PSD, SHROOM2
12	0.343	CHL1, CPNE7, CUX2, DBH, NRP1, PDE4C, TANC2, TMEM132D
11	0.314	ABLIM2, ACE, ADAMTS3, CNGB1, CNTN5, CNTNAP2, DAGLA, DGKZ, DUSP27, GRM5, KL, LPAR4, LRRK1, MYH13, NRG3, PRSS12, PTPRO, RHOBTB2, SHANK2, SORCS1, TENM4, WDR96
10	0.286	ABCG8, ADAD2, ASXL3, CLSTN2, CNTNAP5, DLGAP2, DRP2, FSTL4, GLDC, ITGA9, KIAA1217, MAP2, MCF2L2, MET, NRCAM, PALM, PDE2A, PPFIA2, RBBP8NL, TSHZ3, VWA3A
9	0.257	ADAMTS8, ASTN1, B4GALNT3, BAI2, BICC1, CDH12, CDH13, CDH4, CTNNA2, EPHA10, GPR123, GRIA3, GRM2, HUNK, KATNB1, KIF17, MSLNL, NLRP2, PCDH7, PCDH8, PRDM8, PTPN3, RAPGEF4, RASAL2, SEMA3A, SLIT2, UNC5D, ZNF804B
8	0.229	ABR, ANO3, C6orf222, CDH9, CYP4A22, DCLK3, DLGAP4, DSC1, ELFN2, GDF7, GRM7, IL1RL2, KCNH5, KIAA1045, KIAA1522, LINGO1, LTK, NAGPA, NCAN, NELL2, NUAK1, PDZRN3, POSTN, POU6F2, PPL, PTPRT, RAPGEF2, RASAL1, SNTG1, SORCS3, TRABD2A, TRPA1, TTC40, ZFR2
7	0.200	ACTN1, ADAMTS19, ALOX15, BRIP1, CACNB1, CAPN11, CCDC74A, CLSPN, CNGB3, DCAF6, DMRT3, DNAAF3, FAM65B, GIT1, JAKMIP1, KCNH4, KCNQ5, KNG1, KRT84, LRRC7, MAEL, MAN1A1, MAP3K10, MAPK15, MAST3, MYLK3, MYO16, OLFM1, PCSK1, PHACTR1, PRICKLE2, PRSS16, R3HDM1, RIMBP2, SEMA5B, SGSM2, SLC25A23, SLCO5A1, SPEF2, THEG, UTS2R, VGF, WDR65, ZNF831
6	0.171	AHSG, ANKRD24, ANKRD30B, ATP12A, C1QTNF9B, CABP1, CCDC37, CD101, CHRM4, CHST15, CHSY3, CILP2, CNIH2, COL5A2, CSRNP3, CYP4Z1, DRC1, FKBP11, FOXP1, GALNTL5, GDA, GPR150, GUCY1B3, HIPK4, HTR2A, HTR4, ISLR2, KCNA4, KCNIP3, KMO, LNX1, LRIT1, MCOLN3, MTA3, NEK10, NEK2, NIPAL2, NNT, NPAS1, P2RX2, PAX7, PCDHGC5, PNPLA3, PPP4R4, PPP6R1, PRB2, PTGFR, SH3RF2, SLC12A8, SLC22A9, SLC45A4, SNPH, SNX32, SPAG17, SPAG6, SPATA31A6, TBR1, THRB, THSD1, TOLLIP, TRIML2, WDR66, WDR86, WIPF3
5	0.143	AHRR, ARHGAP28, AZU1, BTBD10, C19orf26, CACNA2D1, CACNG8, CDKL5, CIDEA, CPNE4, CPO, CRACDL, DIO2, DPP4, EPS8L3, EXO1, FAM171B, FAR2, FILIP1, GABRA2, GLB1L2, GLRB, GNA14, GRIP1, GUCY1A3, HAPLN1, HCN1, HDGFL1, HTR3A, ISM2, KCNH3, KCNS2, KCTD13, KHDRBS2, LRRC6, MGAT5B, MKL2, MOXD1, NCEH1, NETO1, NIM1K, NOX3, NPFFR2, NPR3, OPN4, PAH, PAX1, PCDH19, PKP2, PLA2G4E, PNMA5, PRB1, PRRG3, RASGRF2, RBP4, RFX6, RXFP1, SATB2, SCN1A, SERPINF1, SH2D4B, SH2D5, SHD, SLC22A10, SLC38A8, SLCO4C1, SLITRK1, SMYD2, SOWAHB, STRA6, STRN4, SUSD4, SYN3, SYNPO, SYTL5, TEKT4, TMEM132B, TULP1, UROC1, USP43, VSTM2A, WDR16, WDR64, WNT16, ZBBX, ZNF365
4	0.114	ABCC12, ADRA1B, ARHGEF4, ASB2, B3GNT5, BBS12, BEND5, BZW2, CABP7, CCDC33, CCNA1, CDK20, CHRM3, CMIP, COL10A1, CRYGA, CXXC11, CYB561, DGKB, DUSP2, EMX1, EPHX4, ESRP1, FAM81A, FAM84A, FBXO10, FBXO34, FHL2, GABRA3, GAD1, GFRA4, GPR153, HS3ST4, IL12RB2, INSM2, ITPKA, KCNQ3, KREMEN2, KRT3, LHX6, LINGO2, LRFN2, LRFN5, LZTS1, MATK, MEIS3, MEPE, MMD, MMP17, MUC15, NELL1, NGEF, NT5DC3, NTNG2, OAF, OPRD1, PCDH20, PDE1B, PDP1, PGM2L1, PIK3R1, PKD2L1, PLA2G4D, PLEKHM2, PLPPR2, PNKD, PNPLA5, PPP1R16B, PRICKLE1, PRSS35, RARB, RGS20, RGS4, RHPN1, RIMS3, ROPN1L, SATB1, SCN3B, SDR16C5, SEC16B, SFRP2, SHC3, SLC24A3, SLC30A3, SLC9C2, SSTR3, ST6GALNAC5, STAT4, STOML1, TESPA1, TMBIM1, TMPRSS13, TNNT2, TOMM34, TRUB1, TSPAN13, TSPAN16, WASF1, WNT10B
3	0.086	AACS, ABCG5, ACPL2, ACTR3B, ADORA2B, ADPRHL1, ADRB1, AFP, AGBL4, AP1S1, AQP9, ARC, ASIC5, ATCAY, ATP6V1H, B3GALTL, BAP1, BCL11B, BHMT, BTBD9, C16orf59, C4orf50, C9orf117, CA7, CACNA2D3, CACNG3, CAMK2A, CCBE1, CCDC3, CCDC60, CCDC74B, CCDC85A, CCR9, CDRT15L2, CHRNA1, CLIC6, CLPSL1, CNTN3, COL25A1, COL2A1, CREB3L3, CREG2, CRHR2, CRLF1, CYP26A1, CYP46A1, DAW1, DCLK1, DCSTAMP, DDN, DGAT2, DPF1, E2F3, ELANE, EXTL1, FAM155A, FAM49A, FBXL2, FBXO40, FRMPD4, GALNT8,

AC	% cases	Genes
		GPR63, GRASP, GRB14, GRM8, HGF, HS3ST3A1, HS6ST1, HS6ST3, HTR3B, ICAM5, IGFBP2, IGSF3, ILDR2, KCNB1, KCNB2, KCNF1, KCNIP2, KCNJ4, KCTD16, KIAA1324, KIF5A, KRT17, LARGE, LCN15, LRRC40, LRRC4C, LRRTM4, LRTM1, MARCH4, MCHR2, METTL21C, MLIP, MORN3, MYB, NDFIP2, NECAB2, NOS2, OCA2, OR2T34, OR2T6, OSBPL3, PARK2, PCDP1, PCSK2, PLCB1, PLEKHS1, PLK2, PNLDC1, PNMAL2, PPP3R1, PREP, PRTN3, PTER, RASD1, RFTN1, RGR, RMI2, RNF208, RTBDN, RTKN2, RTN4RL2, SCG3, SH2D1B, SLC1A2, SLC22A25, SLC26A4, SLC47A1, ST6GAL2, STOML3, STYK1, TCERG1L, TDO2, TFAP2C, TINAG, TMEM200A, TNFAIP8L3, TRIM71, TUBA4B, UPP2, ZDHHC23, ZNF184, ZNRF4
2	0.057	ACRV1, ADRA1A, ADRA1D, ADTRP, AK5, ANKRD2, ANO5, ARF3, ARL9, ARX, ATL1, ATP5S, ATXN7L3, B3GALT2, B9D1, BCL11A, BRINP2, C16orf87, C2orf50, C6orf223, C7orf62, CACNG6, CAMK1G, CAMKV, CCDC173, CDHR4, CDK5, CDKL1, CHN1, CHRDL1, CHST1, CITED1, COL26A1, CPNE5, CPNE8, CTXN1, CX3CL1, CYP24A1, DACH2, DDI1, DDX54, DLX3, DLX5, DMTN, DNAH6, DRD1, DRGX, EFNB3, ELAVL4, EMX2, F12, FAIM2, FAM110C, FAM163B, FAM211A, FAM216B, FAM217B, FAM3C, FGF12, FGF13, FIBCD1, FKBP1C, FLRT2, FRMPD2, FZD8, GABRG3, GABRQ, GAD2, GJD2, GJD4, GPR6, GRIN2B, GSG1, GSTM5, GYPB, HABP2, HMGCS1, HRH1, HS3ST2, HTR1A, HTR6, IGFL4, IL34, KCNN1, KCNS1, KCNV1, KCTD1, KCTD17, KHDRBS3, KIAA1671, KLHL38, KLK5, KLK7, LAMP5, LPAR3, LPPR3, LPPR4, LRRC73, LRRC8B, LRRTM1, LRTM2, LUZP1, MAPK10, MCHR1, MPPED1, MS4A8, MUM1L1, NCALD, NNAT, NPAS2, NPTX2, NPTXR, NPY, NPY2R, NRSN1, NUDT10, NUDT11, OSTN, OVOL2, OXGR1, PANX3, PDK3, PDYN, PLEKHA1, PPP1R3D, PRKCG, PRM2, PTHLH, PTPN5, RAB27B, RDM1, RHEBL1, RS1, RTP1, SERPINB13, SKA3, SLC39A4, SLC7A3, SOHLH1, SOSTDC1, SOWAHA, SPRYD3, STS, STX19, SYT10, SYT16, TBC1D26, TEX29, THEMIS, THY1, TMEM130, TMEM150C, TMEM183BP, TMEM241, TNFRSF11A, TPRG1L, TRIM54, TRIP13, TYRP1, UBP1, ULBP3, VSTM2L, WNT1, WNT10A, WNT2, YBX2, YWHAZ, ZNF215, ZNF697
1	0.029	ABRACL, ACVR1C, AKAP5, ALDH1A2, ALG1L2, ASCL1, BASP1, BCAS4, BEND3, BHLHA15, C16orf74, C19orf60, C1orf115, C1orf194, C1orf216, C1orf95, C1QL2, C22orf15, C6orf118, C9orf16, CABLES1, CARTPT, CASQ1, CBLN2, CCDC34, CCDC65, CD200R1, CDC42EP3, CDCA5, CDK5R2, CDX1, CHML, CHRM1, CLEC2L, COG8, CORT, CPNE3, CRH, CRYGN, CXCL14, DBN1, DLX1, DNAAF1, DNAL4, DUSP19, DYDC2, EFHD2, EFNA3, EGFL6, EHD3, ELMOD1, EMID1, ENC1, FAM19A1, FAM19A2, FAM237A, FATE1, FBXL16, FCN2, FEZF2, FGF16, FGF8, FIP1L1, FOXG1, GABRB3, GAP43, GPR21, GPR52, GREM2, GULP1, HLF, HNF4A, HOMER1, HPCA, HRH2, HSD11B1L, HSD3B1, HSPB3, HTR1D, HTR1E, IDS, IL5RA, IQCF3, KCNA6, KCNC2, KCNG2, KCNJ1, KCTD4, KLF16, LCN1, LIMD2, LIPC, LMO3, LMO4, LPPR5, LY6D, LY6H, LYNX1, LYPD6B, MAS1, MDH1, METTL7B, MGME1, NAT8L, NDNL2, NEGR1, NGB, NLK, NPBWR1, NPBWR2, NPY1R, NPY5R, NR2E1, NSMF, NUDT4, OR2AG1, OR6A2, OTOS, OTUB2, PAQR9, PCDHGA12, PCP4L1, PFN2, PKDCC, PNCK, PNMA2, POU3F1, PPP1R1A, PRKAR2B, PRKCB, PRM3, PTF1A, PTGES3L-AARSD1, PTRH1, RAP2B, RASD2, REEP5, RFPL2, RGS14, RLBP1, RND1, RP11-383H13.1, SCGB3A1, SCTR, SERPINI1, SERTM1, SLC39A10, SP8, SP9, SPC24, STEAP1, STK32C, TAF7L, TESC, TMEM155, TMEM196, TMEM213, TNNI3K, TOMM40L, TOX, TTC24, TUSC5, UBA52, UBE2E2, VIP, VPS36, VSIG1, WDR3, WNT7B, YIF1A, YWHAH, ZMAT4, ZNF214, ZRSR1

5.3.6 GWAS risk loci overlap

The 4,398 RDVs (4,459 in all individuals) were assessed for their overlap across the risk loci associated with SCZ in the PGC's latest GWAS (Trubetskoy et al., 2022). All 624 risk loci with a p value less than 1×10^{-6} were selected, including 287 loci with genome-wide significance ($p < 5 \times 10^{-8}$) for an association with SCZ. A buffer of 200 kilobases was also added to each locus to assess how many RDVs were located by the boundaries of these loci. This helped to capture genes whose coding regions crossed the loci

boundaries where RDVs affecting the gene fell outside of the study's estimated risk loci. GWAS risk loci were based on a standard cut-off value for linkage disequilibrium, and nearby RDVs may still affect the odds ratio at the associated index SNP. It is important here to note that many genes within each GWAS risk loci were not included in the CO/HP-genes set and RDVs were not analysed in those genes.

As summarised in Table 5.4, approximately 5% (AC=219) of the RDVs were in the 287 genome-wide significant risk loci, with a further 4.6% (AC=203) in the 200 Kb flanking regions. 7.0% (AC=312) of the RDVs and a further 7.8% (AC=350) in the 200 Kb flanking regions were found in all 624 GWAS loci with some association with SCZ.

Table 5.4 Variants overlapping GWAS loci
The total number of variants – total number of alleles (all heterozygous) – and overlap with the 624 loci in the PGC study, of which 287 had genome-wide significance.

Total allele count in samples	4,459	
Number of variants within all 624 loci	312	7.0%
- within 287 GW-significant loci	219	4.9%
Total number of variants 624 buffered loci	662	14.8%
- within 287 buffered GW-significant loci	422	9.5%

RDVs were found in or near 170 (27.2%) of all 624 loci from the PGC study. RDVs from the SCZ groups were in 44 (15.3%) of the 287 genome-wide significant loci. A further 40 loci had RDVs within the 200 Kb buffer. RDVs were also found in 43 (12.8%) of the loci that did not reach genome-wide significance in the PGC study, with a further 43 loci having RDVs within 200 Kb before and after. Several non-significant loci were within 200 Kb of a genome-wide significant locus, and in a few cases the same RDV and gene associated with both loci. Whether those variants affected the odds ratio of either locus cannot be determined from the data.

5.3.6.1 Top overlapping genes

Table 5.5 shows the first 30 GWAS loci with the most RDVs in the 200 Kb buffered region around each locus. Loci with a genome-wide significant ($P < 5 \times 10^{-8}$) association with SCZ are shaded in red. The values for start, end, odds ratio (OR) and P value were extracted from the PGC study. The GWAS locus between 90062197 and 90420197 on

chromosome 5 has 10 RDVs from Adhesion G-protein coupled receptor V1 (ADGRV1, previously GPR98) and a further 28 outside of this locus. On chromosome 17, angiotensin-converting enzyme (ACE) has 11 RDVs within the significant loci. In the PGC study, the reference allele at the index SNP associated with a higher ratio of cases to controls at this locus and had a genome-wide significant OR < 1. ACE was also one of the 106 protein-coding genes prioritised in the PGC fine-mapping study (Trubetskoy et al., 2022). The full list of genes in GWAS loci also included many that have been studied in SCZ, including known SCZ risk genes (see Table 5.5).

Table 5.5 Top 30 GWAS loci with variants in CO/HP-genes in SCZ groups
The top 30 PGC GWAS loci ordered by the largest number of variants found within 200 Kb of the GWAS loci. Loci with a genome-wide significant association with SCZ (P < 5E-08) are shaded red; loci with an association with SCZ (P < 1E-06) are shaded blue. Several GWAS loci overlapped and both can contain a common set of variants/genes (for example, chr 6:89402975-90079975 – CPNE7). Odds ratios (OR) and P values were from the original PGC study for reference.

Chr	Start	End	OR	P value	Loci Genes	Buffer Genes	Loci	Buffer
5	90062197	90420197	1.05548	1.77E-09	GPR98	GPR98	10	38
17	61545779	61579579	0.95256	1.23E-09	ACE	CYB561 LIMD2 TANC2	11	27
7	21468640	21564810	1.05454	1.22E-09		DNAH11	0	22
7	138571080	138642480	0.95542	6.58E-08	KIAA1549	KIAA1549, TMEM213	11	20
12	120017106	120209706	1.04917	1.68E-07	CIT	CCDC60 CIT	12	20
6	24983869	30513869	1.22838	1.21E-39	PRSS16 ZNF184	FAM65B, PRSS16, ZNF184	10	17
11	46227013	47634013	0.91567	2.69E-18	CHRM4 DGKZ	CHRM4 DGKZ	17	17
1	200874225	201024025	1.05106	6.91E-09	KIF21B	KIF21B	17	17
3	48182866	52792866	0.93445	1.95E-08	BAP1 CAMKV CDHR4 GRM2 IQCF3	BAP1 CAMKV, CDHR4 GRM2 IQCF3	17	17
19	19324984	19799984	1.06099	1.77E-14	CILP2 NCAN	CILP2, NCAN	14	14
10	104067277	105219277	1.12019	1.25E-23	PSD	PSD	13	13

Chr	Start	End	OR	P value	Loci Genes	Buffer Genes	Loci	Buffer
3	49647395	51635395	1.05327	4.03E-09	CAMKV CDHR4	CAMKV, CDHR4, GRM2	4	13
10	104285676	104995676	1.04404	8.19E-09		PSD	0	13
16	89402975	90079975	1.0518	1.33E-09	CPNE7	CPNE7	12	12
16	89393580	89682380	0.95858	1.11E-07	CPNE7	CPNE7	12	12
10	33657413	33689853	1.04394	1.26E-07		NRP1	0	12
22	43956279	44205879	1.18803	2.93E-07		MPPED1, PNPLA3, PNPLA5	0	12
19	11342572	11430672	0.94923	4.53E-09	TSPAN16	PLPPR2, SPC24, TSPAN16	4	9
4	30477157	30987157	1.06205	9.70E-08	PCDH7	PCDH7	8	9
7	83630770	83799270	0.96194	2.10E-07	SEMA3A	SEMA3A	4	9
1	177082006	177428006	0.93417	8.28E-14	BRINP2	ASTN1, BRINP2	2	8
3	36834122	36998522	0.94762	3.19E-10		DCLK3	0	8
1	35591823	36645823	1.07466	4.61E-10	C1orf216, CLSPN	C1orf216, CLSPN	8	8
13	38291944	38362106	1.04865	8.02E-10		POSTN	0	8
10	106404343	106889343	1.04728	7.45E-09		SORCS3	0	8
10	64061104	64399104	1.12053	8.36E-08	ZNF365	RTKN2, ZNF365	2	8
1	43793530	44480530	1.06599	2.51E-16		WDR65	0	7
17	1983774	2229074	0.95247	7.85E-10		SGSM2	0	7
3	63715618	64115618	0.95495	5.88E-09	PRICKLE2	PRICKLE2	5	7
8	144796453	144940113	1.04352	3.68E-08	MAPK15	MAPK15	7	7

The proximity of nearly 15% of the RDVs in CO/HP-genes in the SCZ groups to all the GWAS loci gave some evidence that the RDVs identified in the rest of the genes may also contribute to the risk of SCZ but were not significantly associated to differences in ratios of cases to controls at the index SNPs selected in the PGC GWAS. There were still many known SCZ risk genes among the CO/HP-genes that have not been detected by GWAS. For example, the group III metabotropic glutamate receptors, GRM2 and GRM8, were within GWAS loci, however other subunits were not. The CO/HP-genes in the SCZ groups also included 8 RDVs in GRM7 and 11 in GRM5. In the following section, a functional analysis was performed on all the genes with RDVs from the CO/HP-genes isolated across the 35 groups of 100 cases.

CSMD1 had the greatest number of RDVs (42 in 1.2% of all cases) but the RDVs did not occur in any GWAS loci or buffered loci. One GWAS risk loci (chr8:4,170790-4,253590), however, with genome-wide significant ($P = 1.450 \times 10^{-11}$) association for SCZ was found within the CSMD1 gene. Table 5.6 shows two loci of containing RDVs within the gene CSMD1 where 'Start' and 'End' are inclusive of the nearest and furthest RDVs from the risk locus in row 2. The RDVs identified were in CSMD1 exons beyond the 200 Kb buffer around the GWAS loci, the closest being 280 Kb from the loci. The 3 RDVs on the other side of the GWAS loci were 240 Kb from the loci. It is possible that RDVs in CSMD1 had a large effect but weakened by distance on the balance of cases to controls around the index SNP at this locus.

Table 5.6 Table showing CSMD1 variants surrounding GWAS loci
The CSMD1 gene is located on chromosome 8 (GRCh37) and consists of 56 or 68 exons in two different transcripts. Rare damaging variants identified in SCZ cases (blue background) were identified in regions surround the GWAS significant loci (white background).

Chrom	Start	End	Variants	
8	2796112	3889561	39	CSMD1
8	4170790	4253590	0	GWAS loci
8	4494960	4495044	3	CSMD1

5.3.7 Gene ontologies (GO) affecting SCZ groups

A functional analysis was conducted to identify the broader GO biological processes (BP), molecular functions (MF), and cellular locations (CC) for the 940 CO/HP-genes affected by the 4398 RDVs. The objective was to discover whether there were consistent GO terms affected by RDVs across all 35 groups of 100 SCZ cases. 4,720 unique GO terms for all namespaces (BP, MF, CC) were retrieved from Ensembl's most recent BioMart (v101) for 928 of the 940 genes. Each unique GO term was annotated with the total minor allele count in each group, the count of groups with at least one RDV in a gene covered by this GO term, and the percentage of the 3500 cases in whom this GO term was affected by an RDV. 80 GO terms were affected by at least one RDV in all 35 groups. The top GO term was the molecular function, "protein binding (GO:0005515)", containing on average 85 RDVs per group (range: 66 to 106) and

affecting approximately 85% of all cases. The next top GO term was the cellular component, “membrane (GO:0016020)” containing on average 70 RDVs per group (range: 52 to 88), with 76% of the RDVs in the child term, “plasma membrane (GO:0005886)” (mean: 53; range 39 to 65), affecting 70% and 53% of all cases respectively. The fourth GO term was “cytoplasm (GO:0005737)” with a total of 1,748 RDVs across all 35 groups of which more than half, 922 RDVs were in the cytosol (GO:0005829) and not associated with organelles. The fifth GO term was the molecular function, “metal ion binding (GO:00046872)” containing on average 30 RDVs per group (range: 20 to 43), affecting approximately 30% of all cases. The following sections focussed on the top 50 terms in each namespace.

5.3.7.1 GO Biological Processes affecting SCZ groups

Figure 5.13 shows the top 50 GO Biological Process terms by total minor AC, 23 of which contained RDVs in all 35 SCZ groups. The largest number of RDVs were in the cellular processes: ‘signal transduction’ (705 RDVs; range: 10 to 28 per group; in 20.1% of cases), ‘cell adhesion’ (306 RDVs; range: 4 to 16 per group; in 8.7% of cases), ‘axon guidance’ (211 RDVs), ‘chemical synaptic transmission’ (211 RDVs), ‘microtubule-based movement’ (141 variants), ‘phosphorylation’ (307 RDVs; range: 2 to 14 per group; in 8.8% of cases). Next were RDVs in developmental processes: ‘nervous system development’ (360 RDVs; range: 3 to 18 per group; in 10.3% of cases), ‘cell differentiation’ (347 RDVs; range: 5 to 17 per group; in 9.9% of cases). This group also included genes involved in ‘spermatogenesis’ (145 RDVs) and ‘heart development’ (114 RDVs). Third, were RDVs in genes involved in localisation and transport processes: ‘monoatomic ion transport’ with 296 RDVs, a third of which were involved in potassium ion transport (107 RDVs in 33 of 35 groups; in 3.1% of cases), ‘transmembrane transport’ (246 RDVs), and in at least 32 of the 35 groups, ‘vesicle-mediated transport’ (97 RDVs in 2.8% of cases) and ‘protein transport’ (95 RDVs in 2.7% of cases). The remaining RDVs, found in fewer than 35 groups, were in genes involved in both positive and negative regulation of gene expression via regulation of DNA-templated transcription or transcription by RNA polymerase II, metabolic processes including lipid metabolism and protein ubiquitination, cellular responses to calcium ions, and genes involved in system processes such as memory, visual and auditory perception.

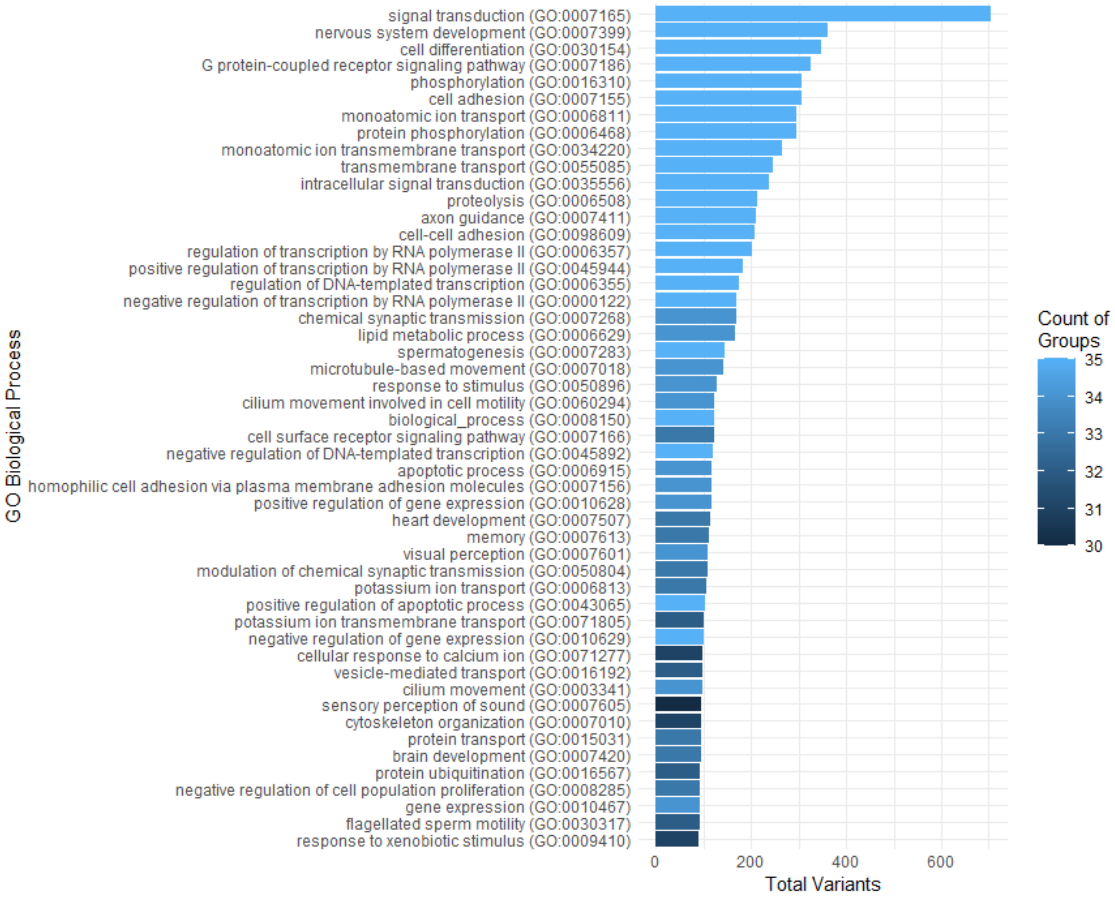

Figure 5.13 Top 50 GO Biological Process terms containing RDVs
Individual GO terms are plotted with the total number of variants per term across all 35 groups.

In the following three sections, three GO BP terms were selected (G-protein coupled receptor signalling, monoatomic ion transport, and axon guidance), and analysed using STRING. The gene names for each of the three terms were plotted as a network graph to identify interaction between the genes. STRING uses various other databases that were checked to annotate genes, such as with common InterPro domains, and pathways from KEGG, Reactome and WikiPathways. STRING can also cluster genes using k-means clustering to identify potential common functionality or domains.

5.3.7.1.1 RDVs affected a wide array of synaptic transmission via G Protein-coupled receptors

Of the 705 RDVs in genes involved in signal transduction, 327 were in the GO BP term 'G protein-coupled receptor signalling pathway'. The 75 genes in this term were plotted (Figure 5.14) as a network graph using STRING to identify common processes that may be affected by the RDVs. Approximately half of the genes were involved in

GPCRs that modulate adenylate cyclase downstream (red in Figure 5.14), and these included subunits of glutamatergic (yellow), neuropeptide (green), cholinergic, serotonergic, adrenergic, and lysophosphatidic acid (LPA) receptors. 13 of these were involved in the cellular response to dopamine (blue). Other genes regulated the activity of GPCRs (RGS14, RGS20), or were GPCRs in other cellular pathways (for example, adhesion, metabolic, chemokine, and olfactory receptors).

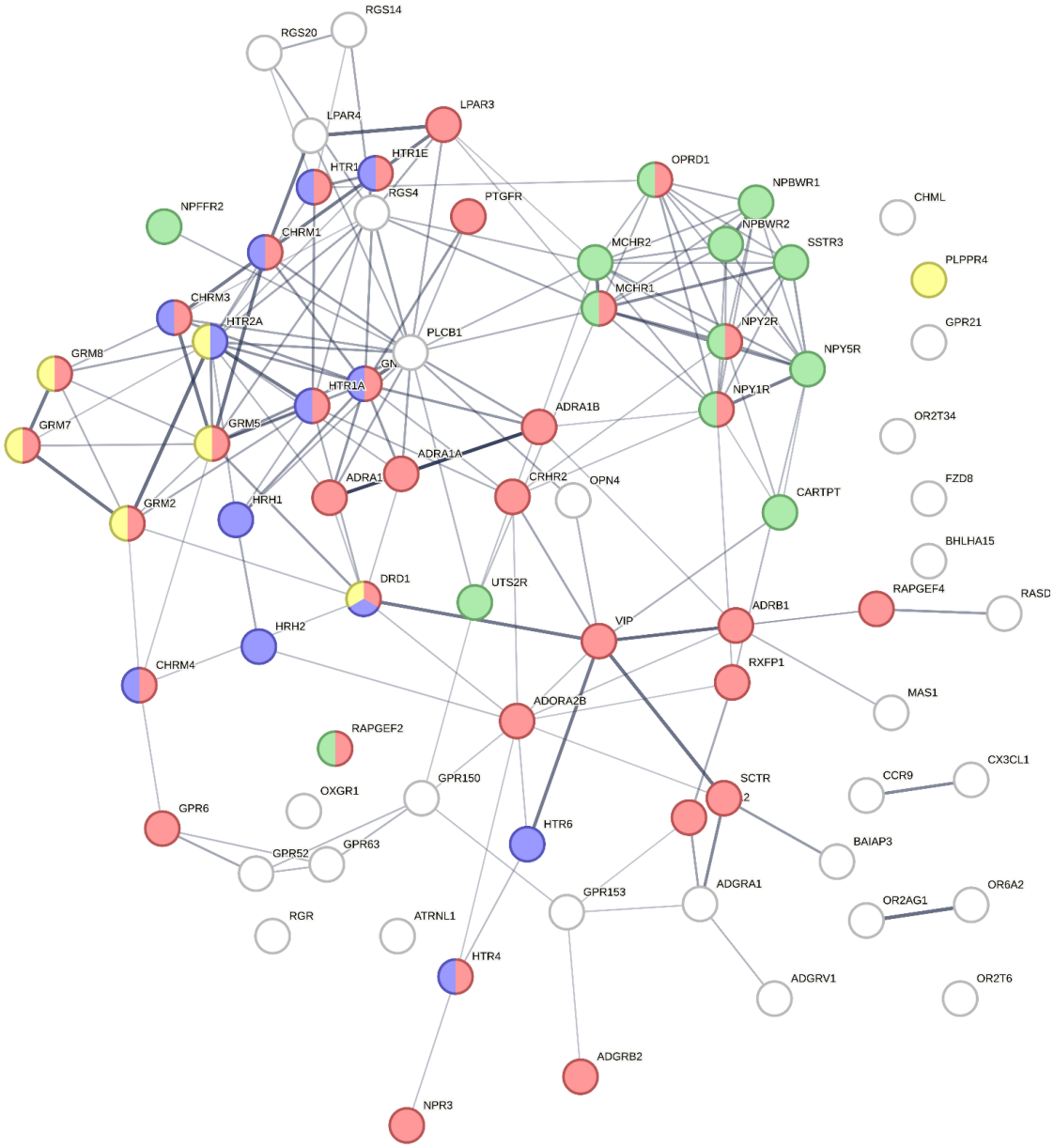

Figure 5.14 Genes with RDVs affecting G-protein coupled receptor pathways
The 75 genes with 309 variants in the GO term 'G-protein coupled receptor signalling pathway' under the term 'signal transduction' were entered into STRING-DB (Szklarczyk et al., 2021). A selection of nodes was highlighted as follows: adenylate cyclase-modifying GPCR signalling (red), cellular response to dopamine (blue), neuropeptide signalling (green) and regulation of glutamatergic signalling (yellow).

5.3.7.1.2 RDVs affected multiple subunits of a wide range of ion transport complexes

Overall, there were 296 RDVs identified in 66 genes in the GO term, 'monoatomic ion transport', involving the directed movement of ions within the cells, or in and out of the cell. Figure 5.15 shows a network graph of the 66 genes clustered by k-means

clustering via STRING, into 4 clusters. Cluster 1 (red) included a wide selection of members of voltage-gated potassium channel subfamilies (KCNA4, KCNA6, KCNB1, KCNB2, KCNC2, KCNF1, KCNG2, KCNH3, KCNH4, KCNH5, KCNIP2, KCNIP3, KCNJ1, KCNJ4, KCNMA1, KCNN1, KCNQ3, KCNQ5, KCNS1, KCNS2, KCNV1) and two sodium channel members (SCN1A, SCN3B, SLC26A4). This cluster included 3 members of the transient receptor potential (TRP) family of cation channels: mucolipin-3 (MCOLN3), which transports Ca^{2+} ions across membranes such as of the early and late endosomes and lysosomes, a regulator of neuronal excitability (PKD2L1), and an endogenous chemosensor for histamines and reactive oxygen species (TRPA1). Cluster 2 (yellow) included several members of the voltage-dependent calcium channels (CACNA1E, CACNA2D1, CACNA2D3, CACNB1, CACNG3, CACNG6, CACNG8), several subunits of the gamma-aminobutyric acid family (GABRA2, GABRA3, GABRB3, GABRG3, GABRQ), a glycine receptor chloride channel (GLRB), two ionotropic glutamate ligand-gated ion channels subunits (GRIA3, GRIN2B), and serotonin receptor subunits (HTR3A, HTR3B). Cluster 3 (green) in Figure 5.15 included three zinc transporters (SLC30A3, SLC39A10, SLC39A4) and two ATPase potassium transporters (ATP12A, ATP6V1H). Lastly, cluster 4 (blue) captured two cyclic-GMP-gated ion channels (CNGB1, CNGB3), which play a role in visual and olfactory signal transduction.

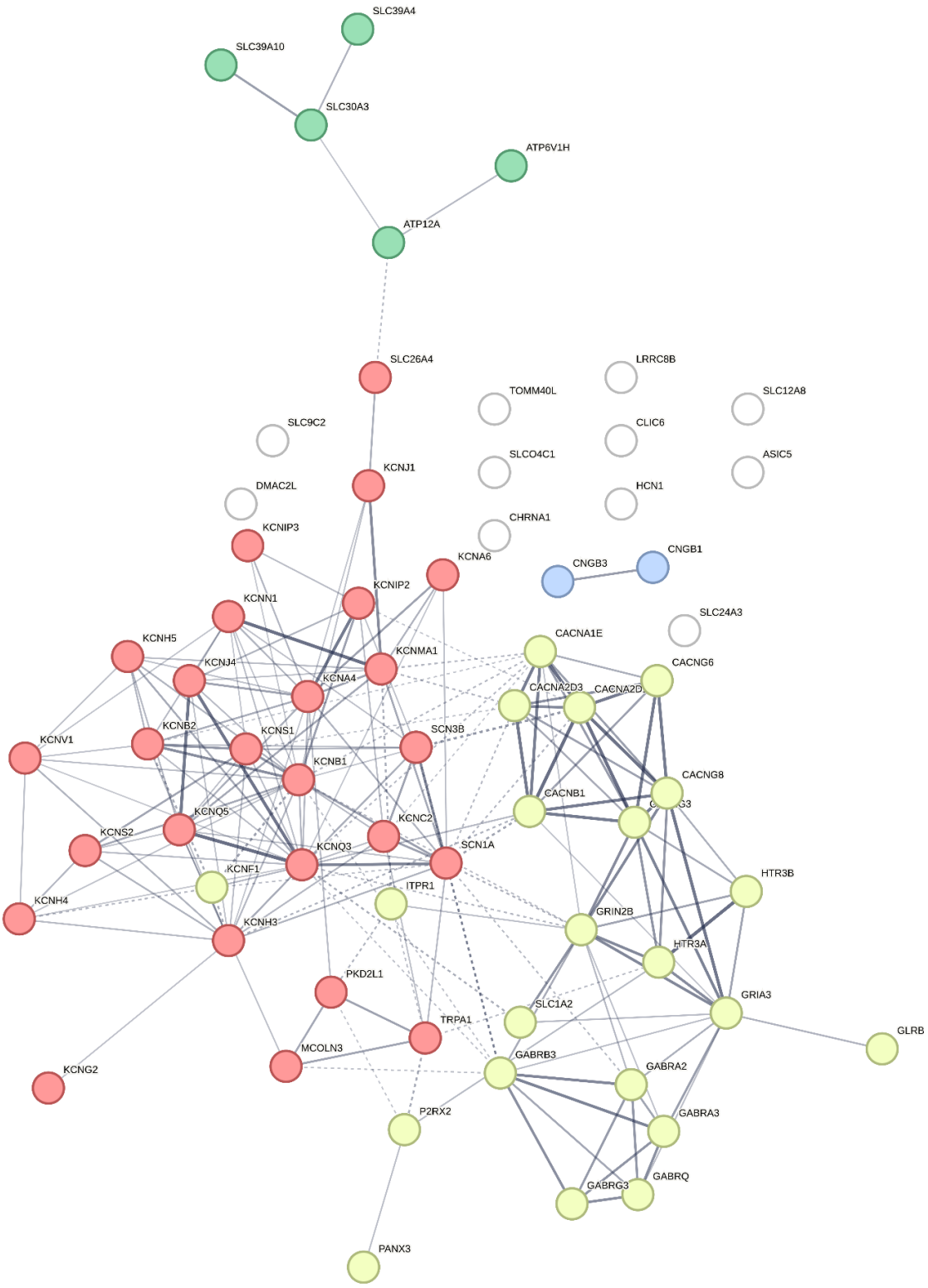

Figure 5.15 Genes with RDVs in Monoatomic ion transport processes
The 66 genes identified with RDVs involved in the GO Biological Process 'monoatomic ion transport' were plotted as a network graph using STRING (Szklarczyk et al., 2021). Terms were clustered using k-means clustering with 4 clusters. The clusters are coloured as follows: genes associated with channels including neurotransmitter receptor subunits, calcium channels or pores (yellow), potassium ion transport (red),

with zinc transport / ATPase (green), and the cyclic nucleotide-gated cation channels (blue).

5.3.7.1.3 Axon guidance genes implicate Ephrin and Semaphorin signalling

There were 27 genes with RDVs in the GO term 'axon guidance', which includes 232 genes overall. The network graph in Figure 5.16 depicted clusters of genes in the semaphorin-plexin signalling pathway (blue) including the transmembrane-ephrin receptors and ephrin (light green) (EFNB3, EFNA3, EPHA5 and EPHA10) and semaphorin-related genes (yellow) (NRP1, PLXNA4, SEMA3A, SEMA5B). Genes shaded red (CNTN5, NRCAM and CHL1) were involved in mixed processes but included neurite outgrowth as a common process, and genes coloured purple (UNC3D and FLRT2) were involved in axon guidance through repulsion of the axon outgrowth. SLIT1 and SLIT2 (green) were molecular guidance proteins involved in neural pathfinding. Other genes also played a role in promoting axon formation, axon guidance, adhesion, and maintenance (for example, neuromodulin (GAP43), KIF5A, DLX5, CDH4).

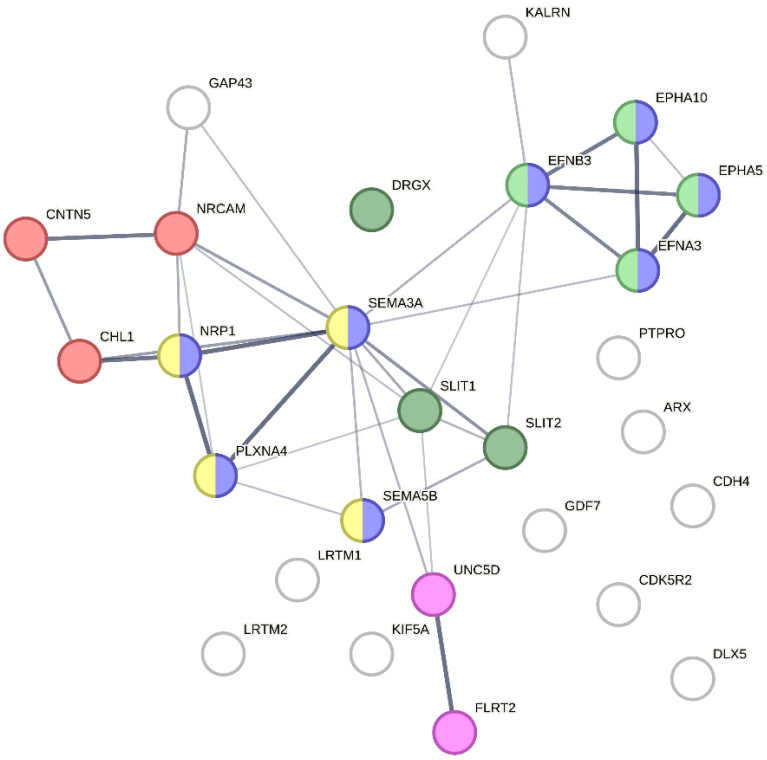

Figure 5.16 Genes with RDVs involved in Axon Guidance
Genes harbouring RDVs involved in axon guidance. Nodes are coloured by STRING clusters: Semaphorin-Ephrin signalling pathway (blue), Ephrin receptor and Ephrin activity (light green), Sema domain (yellow), neurite outgrowth and adhesion (red), axon guidance through repulsion (purple) and chemotaxis (dark green). Individual nodes did not have a confident interaction with the main network. The network graph was produced by entering the genes in the GO term 'axon guidance' into STRING-DB online (Szklarczyk et al., 2021).

5.3.7.2 GO Molecular Functions affecting SCZ groups

Figure 5.17 shows the top 50 GO Molecular Function terms by total minor AC, 21 of which contained RDVs in all 35 SCZ groups. The top GO MF term 'protein binding' (2972 RDVs in 84.9% of all cases) was an umbrella term covering 630 of the 940 genes. The child term 'metal ion binding' included 192 genes with over one third (1042) of the RDVs in 'protein binding'. These genes were predominantly those that bind calcium ions (456 RDVs, 6 to 20 per group; in 57 genes), including CACNA1E, cadherins (CDH12, CDH13, CDH4) and GRM7, or those that bind zinc ions (166 RDVs; 1 to 10 per group; in 38 genes), including ACE and GRIN2B. Several groups of 'protein binding' child terms involved ATP metabolism. 212 RDVs (2 to 13 per group) were in genes that hydrolyse ATP to ADP as a source of energy, including kinesins (KIF17, KIF21B, KIF5A), axonemal dyneins (DNAH11, DNAH3, DNAH6, DNAH7), ATP-binding cassettes (ABCA12, ABCC12,

ABCG5, ABCG8), and ATPases (ATP12A and ATP6V1H), or those kinases and phosphatases that transfer phosphate groups from ATP. 457 RDVs (5 to 23 per group) were in genes involved in the catalysis of the transfer of chemical groups from one compound to another. More than two thirds of these (301 RDVs in 48 genes) were kinases that transfer the phosphate group between ATP and a substrate, the substrates predominantly being serine, serine/threonine, or tyrosine amino acid residues (291 RDVs in 44 genes) in genes involved in signalling pathways.

Another group of terms were related to 'DNA binding' or 'DNA binding activity' revolving around factors, promoters and enhancers, and chromatin. 'DNA binding' had 265 RDVs in 70 genes and of those, 151 RDVs in 47 transcription factors (for example, FOXP1, PAX7, CUX2, and STAT4) and 99 RDVs in genes with 'sequence-specific, double-strand binding' via GC-rich regions or sequence-specific motifs. Many of the genes are found in both and in total there were 80 genes in these categories.

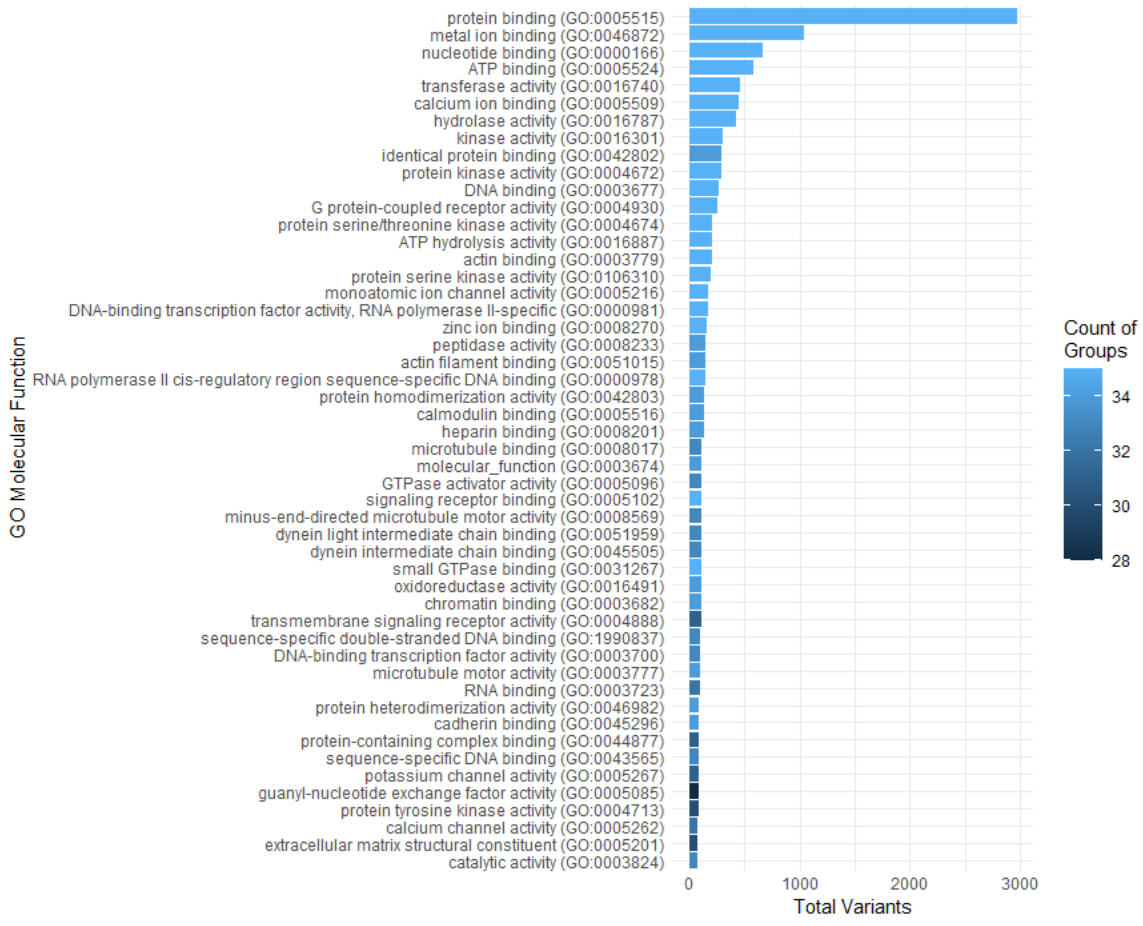

Figure 5.17 Top 50 Molecular Function terms by count of variants

5.3.7.2.1 KEGG enrichment analysis of genes involved in protein binding

Protein binding was a non-specific umbrella term within the Molecular Function results and warranted further analysis. To identify common themes within the 629 genes identified with 2,972 RDVs in total across the 35 SCZ groups, the list of genes was input to STRING and the KEGG pathway enrichment scores were retrieved. Table 5.7 lists the pathways with the strongest enrichment. 'Neuroactive ligand-receptor interactions' was the strongest pathway including genes involved in various neurotransmitter pathways – opioid (OPRD1), acetylcholine (CHRM3), hormones (CRH), adenosine (ADORA2B), epinephrine (ADRA1B, ADRB1, ADRA1A, ADRA1D), neuropeptides (NPY, NPY2R, NPY1R, NPFFR2), dopamine (DRD1), glycine (GLRB), GABA (GABRA3, GABRQ), serotonin (HTR1A, HTR2A, HTR6) and glutamate (GRM5, GRM2, GRIN2B). The second pathway enriched in genes with RDVs was 'Calcium signalling pathway', also highlighted in the previous section. Many of the genes were neurotransmitter-activated G-protein coupled receptors (GPCRs), but also included genes that acted downstream in the signalling pathway. These included Ca^{2+}-dependent kinases (CAMK1G, CAMK2A) and phosphatases (PPP3R1) whose activity can modulate the activity of enzymes (PDE1B) that metabolise cyclic-AMP (cAMP) and c(GMP). The pathway also included phospholipase C beta 1 (PLCB1), inositol 1,4,5-trisphosphate (IP3) receptor (ITPR1), and protein kinase C genes (PRKCB, PRKCG). PLC1B cleaves phosphatidylinositol 4,5-bisphosphate (PIP2) to generate diacylglycerol (DAG), which activates either PRKCB or PRKCG in the presence of Ca^{2+}, and IP3, which binds to ITPR1 in the endoplasmic reticulum membrane causing the release of Ca^{2+} into the cytoplasm. A subset of these genes also was enriched in the pathways 'long-term potentiation' initiated by glutamate receptors and leading to enhanced synaptic transmission, 'glutamatergic synapse', 'amphetamine addiction', and 'WNT signalling pathway'. The IP3 receptor, ITPR1, the protein kinase C, PRKCB, and phospholipase C, PLCB1 were common to most of the pathways enriched in KEGG. There were also 4 RDVs in the gene ITPKA (not included in the enriched pathways) which produces inositol-1,4,5-trisphosphate 3-kinase A that phosphorylates IP3 to IP4. The third strongest enrichment was 'axon guidance' including the semaphorin, ephrin and SLIT genes involved in axon guidance seen in Section 5.3.7.1.3.

Table 5.7 KEGG enrichment analysis of 629 genes with RDVs involved in GO Protein Binding (MF)

Results of a gene enrichment analysis using the KEGG database via STRING. 629 genes (with 2,972 RDVs across all 35 SCZ groups) identified as 'protein binding' in the above GO Molecular Function analysis were analysed.

KEGG pathway	observed genes	background genes	strength	FDR
Neuroactive ligand-receptor interaction	34	329	0.51	3.59E-06
OPRD1, CHRM3, GLRB, CRH, HTR6, ADORA2B, GRM5, CHRM1, ADRA1B, HTR1E, NPFFR2, HTR1A, NPBWR1, NPY2R, CRHR2, NPY1R, VIP, ADRB1, NPBWR2, GABRA3, LPAR3, ADRA1D, ADRA1A, DRD1, GRM2, THRB, NPY, RXFP1, HTR2A, GABRQ, GRIN2B, SSTR3, KNG1, MAS1				
Calcium signalling pathway	22	191	0.56	0.00014
CAMK1G, PPP3R1, PDE1B, CHRM3, PRKCG, HTR6, ADORA2B, GRM5, ITPR1, CHRM1, ADRA1B, NOS2, PLCB1, CASQ1, ADRB1, GNA14, ADRA1D, ADRA1A, DRD1, HTR2A, PRKCB, CAMK2A				
Axon guidance	21	176	0.57	0.00014
EFNB3, PPP3R1, NGEF, SEMA3A, NRP1, SLIT1, GDF7, LRRC4C, RND1, MET, PLXNA4, EFNA3, NTNG2, UNC5D, SEMA5B, CDK5, SLIT2, PIK3R1, EPHA5, EPHB6, CAMK2A				
Salivary secretion	13	89	0.66	0.0013
CHRM3, PRKCG, KCNMA1, GUCY1A1, ITPR1, ADRA1B, PLCB1, ADRB1, ADRA1D, ADRA1A, PRB2, GUCY1B1, PRKCB				
cGMP-PKG signalling pathway	15	163	0.46	0.0247
CREB3L3, PPP3R1, OPRD1, CNGB1, KCNMA1, GUCY1A1, ITPR1, ADRA1B, PDE2A, PLCB1, ADRB1, ADRA1D, ADRA1A, GUCY1B1, KNG1				
Long-term potentiation	9	63	0.65	0.0247
PPP3R1, PPP1R1A, PRKCG, GRM5, ITPR1, PLCB1, GRIN2B, PRKCB, CAMK2A				
Cholinergic synapse	12	109	0.54	0.0247
CREB3L3, CHRM3, PRKCG, ITPR1, CHRM1, KCNJ4, PLCB1, KCNQ5, KCNQ3, PIK3R1, PRKCB, CAMK2A				
Renin secretion	9	66	0.63	0.0247
PPP3R1, PDE1B, KCNMA1, ACE, GUCY1A1, ITPR1, PLCB1, ADRB1, GUCY1B1				
Cushing syndrome	14	153	0.46	0.0266
CREB3L3, RASD1, E2F3, WNT2, CRH, WNT10B, ITPR1, KCNA4, PLCB1, CRHR2, WNT7B, HSD3B1, FZD8, CAMK2A				
Gap junction	10	87	0.56	0.0282
PRKCG, GUCY1A1, GRM5, ITPR1, PLCB1, ADRB1, DRD1, GUCY1B1, HTR2A, PRKCB				
Glutamatergic synapse	11	112	0.49	0.0439
PPP3R1, PRKCG, SLC1A2, GRM5, ITPR1, HOMER1, PLCB1, SHANK2, GRM2, GRIN2B, PRKCB				
Melanogenesis	10	95	0.52	0.0439
CREB3L3, PRKCG, WNT2, WNT10B, PLCB1, WNT7B, FZD8, TYRP1, PRKCB, CAMK2A				
Aldosterone synthesis/secretion	10	94	0.52	0.0439
CAMK1G, CREB3L3, DAGLA, PRKCG, ITPR1, PDE2A, PLCB1, HSD3B1, PRKCB, CAMK2A				
Amphetamine addiction	8	65	0.59	0.0442
CREB3L3, PPP3R1, PRKCG, ARC, DRD1, GRIN2B, PRKCB, CAMK2A				
Wnt signalling pathway	13	154	0.42	0.0476
PPP3R1, SERPINF1, PRKCG, WNT2, SFRP2, WNT10B, PLCB1, WNT7B, PRICKLE1, FZD8, NLK, PRKCB, CAMK2A				

ITPR1 was also one of the top 3 genes with 15 RDVs in all 3500 cases in these enriched KEGG pathways seen in 0.43% of cases across all 35 SCZ groups (see Table 5.8). Diacylglyceride lipase A (DAGLA) catalyses the breakdown of DAG and there were 11 RDVs in all SCZ cases. The genes with the most RDVs in this selection were SLIT1 (19 RDVs; 0.54% of cases) and EPHA5 (18 RDVs; 0.51% of cases) in 'axon guidance'. There was a total of 33 RDVs between SLIT1, SLIT2 and SLITRK1, 40 RDVs between the Ephrin A and B receptors (EPHA5, EPHA10, EPHB6), and 26 RDVs in neuropilins 1 (NRP1) and Plexin A4 (PLXNA4), which binds to the NRP1 and transduces the signal from semaphorin proteins (for example, SEMA3A and SEMA5B with 16 RDVs in the dataset).

The calcium-activated potassium channel subunit A1 (KCNMA1) had 12 RDVs (0.37% of cases) but there were 84 RDVs in all potassium channels. The metabolic glutamate receptor 5 (GRM5) had 11 RDVs and there were 31 RDVs in all 4 GRM subunits (GRM2, GRM5, GRM7, GRM8). The cGMP-gated ion channel, CNGB1, had 11 RDVs. CNGB1 is an ion channel that regulates ion flow into the rod receptors in the retina. Also, in the dataset was CNGB3, which regulates ion flow into the cone receptors. Together they have 18 RDVs, but their function in the brain is not clear.

Also of note, there were 70 RDVs across all cadherins and protocadherins, and 35 RDVs in the WNT signalling pathway genes, across all 35 SCZ groups.

Table 5.8 Top 10 genes with RDVs in cases enriched in KEGG pathways from 'protein binding'
Genes enriched in KEGG pathways in 'protein binding' are shown with the proportion of cases with RDVs.

Ensembl Gene ID	Gene	in % of cases
ENSG00000187122	SLIT1	0.543
ENSG00000145242	EPHA5	0.514
ENSG00000150995	ITPR1	0.429
ENSG00000221866	PLXNA4	0.400
ENSG00000106123	EPHB6	0.371
ENSG00000156113	KCNMA1	0.371
ENSG00000099250	NRP1	0.343
ENSG00000168959	GRM5	0.314
ENSG00000070729	CNGB1	0.314
ENSG00000134780	DAGLA	0.314

5.3.7.3 GO Cellular Components affecting SCZ groups

Figure 5.18 shows the top 50 GO Cellular Components (CC) terms by total minor AC, 36 of which contained RDVs in all 35 SCZ groups. The top GO CC 'membrane' included 2,463 RDVs affecting approximately 70% of all cases. Three quarters of these were in the 'plasma membrane' (1870 RDVs in 53.4% of cases), 255 RDVs were in genes located in the postsynaptic membrane, and 153 RDVs in genes located at the apical plasma membrane. 191 RDVs were in genes associated with the endoplasmic reticulum (for example, ITPR1). 626 RDVs were found in genes associated with the 'synapse', with 340 RDVs at glutamatergic synapses (including PSD, GRM2, GRM5, CAMKV, ELAVL4). 557 RDVs associated with 'cytoskeleton' genes, including 'cilium' (292 RDVs in 40 genes), 'microtubules' (238 RDVs in 24 genes), and 'axoneme' (167 RDVs in 14 genes). 68 genes in 'dendrites' contained 348 RDVs, and 25 genes with 177 RDVS were in 'dendritic spine' with 10 genes shared between the two.

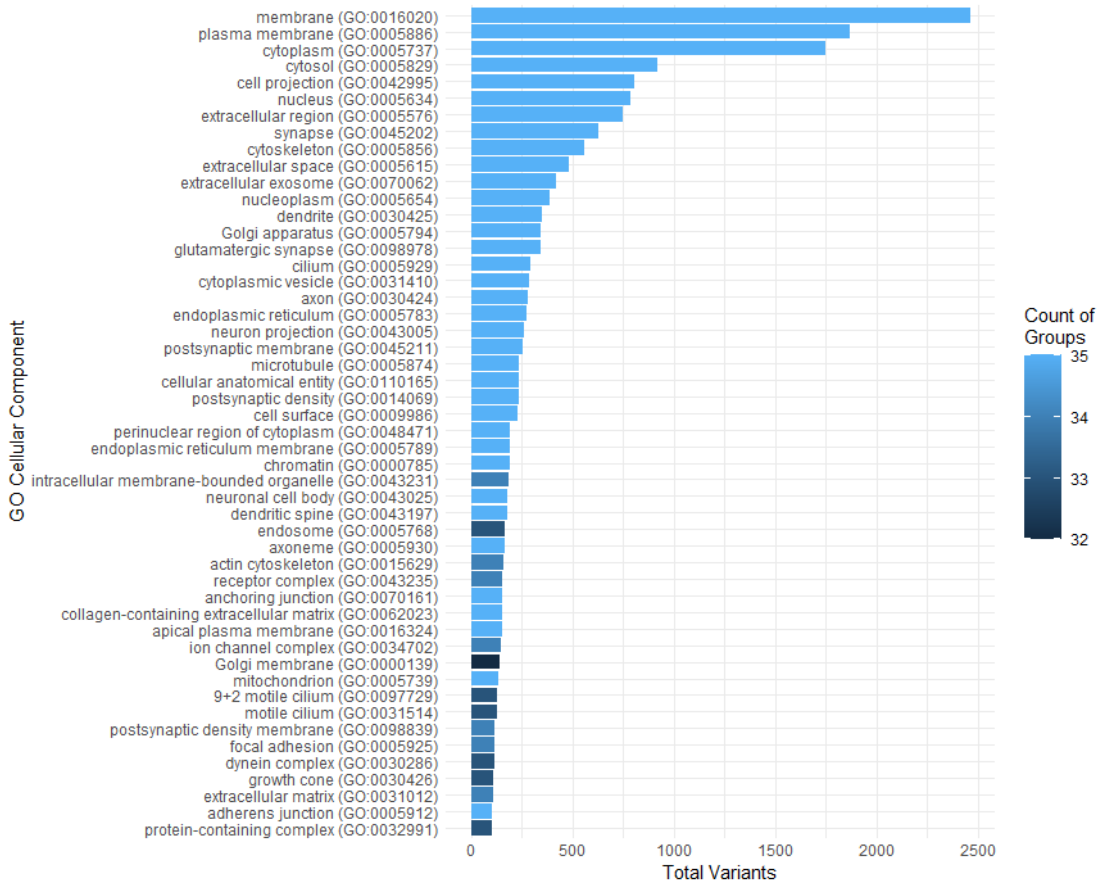

Figure 5.18 Top 50 Cellular Components by count of variants

480 RDVs in 104 genes were in the GO CC term 'extracellular space' affecting approximately 13% of cases. These genes were analysed using STRING for enrichment in Reactome pathways. Breaking down and remodelling the ECM is an important process for neuronal plasticity, axon guidance, and synapse formation and maintenance. The Reactome pathway with the lowest false discovery rate (FDR = 2.05 x 10^{-5}) was for 'extracellular matrix (ECM) organisation' (HSA-1474244), including several collagenases and other structural proteins involved in the structure and organisation of the ECM (COL5A3, COL10A1, COL5A2, COL2A1, COL24A1, COL25A1, NCAN and HAPLN1), serine proteases involved in the remodelling of the ECM (ADAMTS3, MMP17, KLK7 and ELANE) and alpha-actinin 1 (ACTN1). The collagenases are involved in the trimerization (pathways 4 and 5 in Table 5.9) of the collagen fibres in the ECM. Also involved in axon guidance were 3 genes (SLIT1, SLIT2 and NELL2) with the strongest enrichment score in 'regulation of commissural axon pathfinding by SLIT and ROBO'. The genes are also strongly enriched in the WNT signalling pathway and the associated pathways, 'Class B/2 (Secretin family receptors)', and 'WNT ligand biosynthesis and trafficking'.

Table 5.9 Reactome enrichment analysis of 104 genes with RDVs involved in GO Extracellular Spaces (CC)
Results of a gene enrichment analysis using the Reactome database via STRING. 104 genes (with 480 RDVs across all 35 SCZ groups) associating with the 'extracellular space' in the above GO Cellular Component analysis were analysed.

Reactome pathway	observed genes	background genes	strength	FDR
Extracellular matrix organization	13	300	0.92	2.05E-05
NCAN, COL5A3, HAPLN1, ADAMTS3, COL10A1, MMP17, COL24A1, COL5A2, COL2A1, ACTN1, COL25A1, ELANE, KLK7				
WNT ligand biogenesis/trafficking	6	26	1.64	2.05E-05
WNT16, WNT10A, WNT2, WNT1, WNT10B, WNT7B				
Class B/2 (Secretin family receptors)	8	94	1.21	4.53E-05
WNT16, WNT10A, WNT2, CRH, WNT1, WNT10B, WNT7B, PTHLH				
Collagen biosynthesis and modifying enzymes	7	67	1.3	6.47E-05
COL5A3, ADAMTS3, COL10A1, COL24A1, COL5A2, COL2A1, COL25A1				
Collagen chain trimerization	6	44	1.42	0.0001
COL5A3, COL10A1, COL24A1, COL5A2, COL2A1, COL25A1				

Reactome pathway	observed genes	background genes	strength	FDR
Degradation of the extracellular matrix	8	140	1.04	0.00032
COL5A3, COL10A1, MMP17, COL5A2, COL2A1, COL25A1, ELANE, KLK7				
MET activates PTK2 signalling	5	30	1.5	0.00032
HGF, COL5A3, COL24A1, COL5A2, COL2A1				
Collagen degradation	6	64	1.25	0.00043
COL5A3, COL10A1, COL5A2, COL2A1, COL25A1, ELANE				
Signalling by MET	6	79	1.16	0.0011
HGF, COL5A3, COL24A1, COL5A2, COL2A1, UBA52				
GPCR ligand binding	12	459	0.7	0.0012
CX3CL1, WNT16, WNT10A, WNT2, CRH, WNT1, WNT10B, WNT7B, CORT, NPY, PTHLH, KNG1				
Signal Transduction	30	2540	0.35	0.0021
CX3CL1, HGF, WNT16, VGF, WNT10A, COL5A3, NRP1, WNT2, SFRP2, CRH, WNT1, WNT10B, GREM2, FGF8, FLRT2, WNT7B, LINGO1, COL24A1, NRG3, COL5A2, CORT, COL2A1, ACTN1, YWHAZ, NPY, UBA52, FGF16, IL5RA, PTHLH, KNG1				
Assembly of collagen fibrils and other multimeric structures	5	61	1.2	0.0034
COL5A3, COL10A1, COL24A1, COL5A2, COL2A1				
Non-integrin membrane-ECM interactions	5	59	1.21	0.0034
COL5A3, COL10A1, COL5A2, COL2A1, ACTN1				
Signalling by Receptor Tyrosine Kinases	12	521	0.64	0.0034
HGF, VGF, COL5A3, NRP1, FGF8, FLRT2, COL24A1, NRG3, COL5A2, COL2A1, UBA52, FGF16				
Signalling by WNT	9	298	0.76	0.0043
WNT16, WNT10A, WNT2, SFRP2, WNT1, WNT10B, WNT7B, YWHAZ, UBA52				
Regulation of commissural axon pathfinding by SLIT and ROBO	3	10	1.76	0.0048
SLIT1, NELL2, SLIT2				
ECM proteoglycans	5	76	1.1	0.0078
NCAN, COL5A3, HAPLN1, COL5A2, COL2A1				
NCAM1 interactions	4	42	1.26	0.0109
NCAN, COL5A3, COL5A2, COL2A1				
Syndecan interactions	3	27	1.33	0.0497
COL5A3, COL5A2, ACTN1				

5.4 Discussion

The main aim of this chapter was to apply the filtering and analysing methods developed in Chapter 4 to further groups of 100 patients and controls sampled from the Swedish Exomes project. The objectives were to prioritize rare, likely-damaging variants unique to SCZ cases in the study and identify common pathways between the groups affected by these variants. Overall, 4398 ultra-rare damaging variants were

identified exclusive to all 3500 SCZ patients. Key GO terms identified were very general terms, for example 'protein binding', 'ion transport', 'signal transduction' and 'axon guidance'. However, more specific terms narrowed down the findings to implicate potentially key pathways, for example, axon guidance and synapse-related activities; metal, calcium, and zinc ion binding; DNA binding and transcription factor activity; GPCR receptor activity and calcium signalling; cell adhesion; extracellular matrix maintenance; and RNA polymerase-related gene expression. It is important to highlight that while many of the terms were neuron-specific, this is to be expected given that the CO/HP-genes were originally analysed from differential expression between four brain regions in Chapter 3. The CO/HP genes within these broad GO terms were related various pathways including synaptic pathway (neurotransmission, scaffolding, RHO GTPase activity, and PSD structure), GPCR-related signal transduction, the complement system, chromatin remodelling, gene expression regulation, WNT and calcium signalling and RHO GTPase activity.

In this chapter, the functional analysis was conducted on the entire dataset, with variants assigned to individual groups of 100. This approach was chosen to ensure a focused assessment of the biological pathways in each defined group and an inter-group comparison, rather than aiming for an analysis across all samples. Although running a functional analysis for all variants would enhance the detection of specific pathways in the whole cohort due to increased statistical power, the methodology was designed to balance thoroughness with granularity. By examining variants within specific groups of cases and controls, this study aimed to identify common pathways with variants affecting different member genes in affected individuals. Future work might benefit from exploring how the findings differ from an aggregated approach to further validate and expand upon the findings.

5.4.1 Filtering and Analysing Variants in CO/HP-genes

Variants from both SCZ and control exomes were restricted to the CO/HP-genes and any variants found in gnomAD with an allele count greater than 5 were removed. Overall, there were 26,976 unique SNVs and indels with a total allele count of 29,199 among the 8,300 individuals in the 48 CTL groups and 35 SCZ groups. Once filtered to ultra-rare, likely-damaging variants exclusive to SCZ cases, there were 4,398 variants in 3500 SCZ cases.

It is challenging to compare this with other studies, because of the focus on variants in specific genes. For an idea, Genovese et al. (2016) filtered out variants from ExAC and found 244,246 coding-sequence and splice-site ultra-rare variants (URVs) that were exclusive to the full Swedish cohort of 12,332 individuals. This is approximately 20 variants per individual in 21,000 genes. In our set of 1,076 brain region-specific genes, there were approximately 3 URVs per individual. It may not be appropriate to deduce that this is equivalent to 60 URVs per individual in 21,000 genes, because it would be unlikely that all genes would be affected equally by URVs. By filtering variants from gnomAD rather than ExAC, there should be fewer URVs in this thesis overall compared to their study. If the true gene set for SCZ-related genes contained 3000 genes, then perhaps the figure in Genovese et al. (2016) is closer to 20 URVs per individual in 3000 genes (and very few variants were found in other genes), and ours would be greater than 3, approximate 3-20 URVs in 3000 genes depending on the relevance of the missing genes to SCZ.

Filtering out variants exclusive to SCZ cases also removed all the low-quality variant calls from the data set. A bulk of indels identified at the ends of exons and missense variants with low Fisher Strand Bias scores were filtered out. Fisher Strand Bias indicates is a sequencing error determined by different reads produced from the forward and reverse strands from the same DNA fragment. Many of these were shared between cases and controls, but more importantly were not in gnomAD and could be mistaken for true ultra-rare variants. False splice variants at exon boundaries appeared to be another common technical error in Exome sequencing. This may be exacerbated by smaller exon flanking regions that do not allow sufficient read depth of the exon boundaries.

A patient-centred analysis, performed after the work in this study was completed, revealed that 1,066 of the 3,500 cases (approximately 3.3 cases per group of 100) did not possess any RDVs in the 1076 CO/HP genes. This suggests two additional observations. Firstly, that the full genetic heterogeneity of SCZ pathophysiology was not fully captured by the CO/HP-genes. And secondly, that the RDVs per gene in the 2,444 affected cases were greater than reported. Among the remaining 2444 cases, each SCZ case harboured between one and 15 RDVs in the CO/HP-genes. This finding suggests that while a substantial proportion of patients may have rare variants

contributing to their condition, there might be alternative mechanisms or pathways involved in SCZ development for the remaining patients. Further research is necessary to fully characterise the variants among cases.

5.4.2 Identifying Consistent Biological Pathways Enriched for Rare Variants

GO Term variant aggregation analysis revealed 80 terms with RDVs in all 35 SCZ groups.

In the Molecular Function category, the most frequent term was 'protein binding.' Although it is a broad term, this finding emphasizes the potential importance of protein-protein interactions in SCZ pathology. Other significant themes included various ion binding activities (metal ion, calcium ion, and zinc ion binding), nucleotide and ATP binding, and kinase activity (including protein kinase, serine/threonine kinase, and serine kinase activities). Additionally, terms related to DNA binding, transcription factor activity, and G protein-coupled receptor activity were identified. These shared Molecular Function terms suggest that disruptions in protein interactions, ion regulation, enzymatic activities, and gene expression may be crucial underlying factors in SCZ pathophysiology.

In the Biological Process category, several terms were related to signal transduction, highlighting the potential importance of cellular signalling pathways in SCZ pathology. Additional themes included nervous system development and cell differentiation, indicating that disruptions in these processes may contribute to SCZ development. Furthermore, terms associated with G protein-coupled receptor signalling, phosphorylation, and cell adhesion were identified. Other shared terms involved various aspects of ion transport, intracellular signal transduction, proteolysis, and axon guidance. Notably, several terms were related to the regulation of transcription by RNA polymerase II, emphasizing the potential impact of gene expression dysregulation in SCZ.

In the Cellular Component category, various terms related to cellular structures and compartments were identified. Key themes included the plasma membrane and various organelles, such as the nucleus, Golgi apparatus, endoplasmic reticulum, and mitochondria. Additionally, terms associated with the cytoskeleton, synapses, and neuron projections were present, highlighting the potential importance of neuronal structure and connectivity in SCZ. The identification of extracellular components, such

as the extracellular matrix and adherens junctions, indicated that cell-cell interactions and communication may contribute to SCZ pathology, particularly through their roles in axon guidance and synapse maintenance.

The identification of synapse-related GO terms is consistent with previous research on SCZ pathology. Numerous studies have highlighted the importance of synaptic formation, maintenance, and function in the context of SCZ (Breitmeyer et al., 2023; Coelewij & Curtis, 2018; Genovese et al., 2016; Howes & Onwordi, 2023; Owen et al., 2023; Singh et al., 2022; Wu et al., 2022). The SCHEMA study (Singh et al., 2022) also found significant enrichment of synaptic-related processes and cellular components, such as postsynaptic density, regulation of ion transmembrane transport, and synaptic transmission. A review of genomic finding in SCZ (Owen et al. 2023) indicated that synaptic-related processes are among the various biological pathways implicated in SCZ. These common findings across different studies and analytical approaches further support the hypothesis that synaptic dysfunction plays a critical role in the development of schizophrenia. However, this study also implies that it may not be a fundamental role in all cases.

In addition, the involvement in SCZ of GPCR receptors and downstream signalling components, such as calcium signalling, IP3, and adenylyl cyclase has been reviewed in Boczek et al. (2021). The roles of netrins, semaphorins, ephrins and slits is reviewed in Casas et al. (2022) who point out that these genes are also involved in angiogenesis and the formation and maintenance of the blood brain barrier as well as axon guidance and growth cones. Ephrin action in the front cortex in relation to SCZ was also further characterised in a single-cell and spatial transcriptomics study (Huuki-Myers et al., 2024). The interaction of EPHA5 and EFNA5 was located to layers 5/6 and layer 6 excitatory neurons in the DLPFC. This thesis identified RDVs in EPHA5 in 0.514% of patients, but also identified RDVs in other members of the ephrin family (EFNB3, EFNA3, and EPHA10), which may lead to similar effects in different patients.

The presence of diverse affected terms across Molecular Function, Biological Processes, and Cellular Component highlights ‘pathway heterogeneity’ in SCZ groups, indicating that multiple molecular pathways may contribute to schizophrenia. This finding underscores the complexity of the disorder and the need for further

investigation into the specific roles these pathways play in SCZ development and progression.

The analysis of the pathways could be extended by analysing patterns of co-occurrence within individual cases in which more than one RDV was identified.

RVs could not be detected using WES in promotor or enhancer regions – but finding RDVs in transcription factors – offers and alternative method through which gene expression can be altered instead of requiring cis-bound variants. Future studies identifying rare variants in whole genomes would be more comprehensive than can be achieved through whole exomes.

5.4.3 Overlapping GWAS loci and rare damaging variants (RDVs)

An analysis of the overlap between the RDV positions and the 624 risk loci, of which 287 showed a genome-wide significant association with SCZ, from the latest SCZ PGC3 GWAS study (Trubetskoy et al., 2022) revealed a modest overlap of 7% with all loci and approximately, 5% with genome-wide significant loci. On the hypothesis that the effect of ultra-rare damaging variants may have an impact further away from common SNPs (Dickson et al., 2010), a buffer 100 kilobases was added to all GWAS loci. This increased the overlap to approximately 15% for all loci and to 9.5% for genome-wide significant loci.

The two loci with the greatest number of RDVs were ADGRV1 with 38 RDVs, 10 of which were within the loci and 28 nearby; and ACE which itself was one of the key findings in the PGC3 GWAS study (Trubetskoy et al., 2022). Notably, CSMD1 was not identified because the GWAS loci overlapped an intron within the gene. This locus is likely to have been significant due to ultra-rare variants throughout the CSMD1 gene in the large cohort of patients, rather than due to any influence by the common tag SNPs used to determine the association. Genovese et al. (2016) tested for overlap in the PGC2 GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) by testing for enrichment of variants within the 108 GWAS loci. They found a nominally significant enrichment (OR = 1.37; P = 0.027), which like this thesis also hints at some convergence (Genovese et al., 2016). The nominal signification also adds evidence that the loci are incomplete and that further studies are needed to identify the full extent of locus heterogeneity in SCZ.

5.4.4 Potassium channels and shared associations with other disorders

Many of the genes and biological processes identified in this chapter share an involvement in other diseases such as potassium channels in autism (Cheng, P. et al., 2021) and epilepsy (Gao et al., 2022). These channels represent the largest family of genes identified in this study. Potassium channels are also involved in heart function and could explain the comorbidity of SCZ and heart problems in a subset of patients (Garrett et al., 2022).

5.4.5 Conclusion and Further Work

One way to compare whether GO terms are closer between the groups of SCZ cases versus those of controls, would be to create semantic similarity matrices between all lists of GO terms associated with the genes with RDVs in each group. A semantic similarity matrix would be a measure of the distance between the GO terms in each group and the average similarity between SCZ groups vs between SCZ and controls could be statistically tested.

While this study revealed a high number of rare damaging variants (RDVs) in the CO/HP genes for both SCZ cases and controls, further characterization of these differences was beyond the scope of the current analysis. Future work should focus on investigating these discrepancies, as understanding the distinctions between patient and control groups may provide additional insights into the pathophysiology of schizophrenia and the specific roles of these genes in the disorder's development.

6 Discussion

6.1 Introduction

Exome studies in schizophrenia (SCZ) are confronted with the intricate task of discerning pathogenic variants amidst the collective landscape of ultra-rare variants present across individuals. This challenge is compounded by SCZ's polygenic nature, where multiple genetic factors contribute to the disorder's manifestation. Discriminating between causal and benign variants remains challenging, particularly given the limited understanding of genes implicated in SCZ pathogenesis and the complexities of shared genetic involvement in SCZ and related psychiatric disorders.

Central to these endeavours are gene sets, which encapsulate genes anticipated to manifest in the SCZ phenotype, derived from sources such as GWAS, microarray studies, curated databases, and multi-tissue gene expression projects like GTEx and Human Brain Map. Although genome-wide significant loci from GWAS provide robust starting points, their scope has traditionally been limited, requiring larger cohorts to reveal modest increases in significant loci. Moreover, GWAS fail to identify genes less commonly associated with SCZ that might contribute to SCZ pathology. For instance, early GWAS identified 108 significant loci associated with SCZ (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). A more recent study by Trubetskoy et al. (2022) significantly increased this number to 287 significant loci, highlighting one perspective of the evolving landscape of SCZ genetics research. This expansion underscores the necessity for broader gene sets, including not only known loci associated with SCZ but also rarer genes potentially involved in SCZ pathogenesis, as well as connected genes identified through co-expression network analysis (Radulescu et al., 2020). However, care must be taken to prevent overexpansion; for example, including irrelevant genes within GWAS significant loci can dilute the relevance of the gene set to SCZ pathogenesis.

One approach has been to focus on genes highly expressed in cells affected in SCZ. As SCZ is a neurological disease, gene expression profiles of the brain are the most relevant. To assess which genes are likely to be affected by variants, a reasonable strategy is to focus on genes that function in healthy, non-diseased brains. Gene expression in diseased tissue is already perturbed by the disease and may include

genes compensating for the loss of genes knocked out by damaging variants, or genes upregulated to cope with an ongoing diseased cell state and may miss genes with lower expression due to damaging variants.

Brain cells are less accessible for study compared to other body cells. Studying gene expression in brain cells is currently restricted to post-mortem studies, but because RNA and proteins degrade soon after death, donor tissues must be processed rapidly and preserved for microarray or exome studies. Genes derived from microarray and RNA-Seq expression data for the hippocampal formation and DLPFC have been used to test for variant enrichment (Genovese et al., 2016; Purcell et al., 2014; Singh et al., 2022).

Prior to the availability of comprehensive gene expression resources like those provided by the GTEx project, many studies investigating neurological and psychiatric disorders were limited to gene sets derived from expression within broad anatomical regions such as 'brain' or the 'frontal cortex'. While valuable for regional-specificity, the gene sets lacked disease specificity. This study aimed to refine the gene set beyond a focus solely on regional-specificity by leveraging the rich data from GTEx to refine the gene set by focussing on genes highly expressed in brain regions consistently implicated by SCZ (frontal cortex and hippocampus) and excluding genes with high expression in control regions (cerebellum and substantia nigra). This approach aimed to generate a more targeted gene set potentially more relevant to the pathophysiology of SCZ in these implicated regions.

6.2 Increased exclusivity of genes to the frontal cortex and hippocampal gene sets compared to studies

The present thesis adds a valuable resource for studies investigating neurological and psychiatric disorders that affect the frontal cortex and hippocampus by establishing a novel gene set focussed on these key brain regions. Unlike existing sets defined by broad anatomical regions or disease categories, the approach in this thesis leverages interregional RNA-Seq differential expression analysis to identify genes with expression patterns more restricted and exclusive to the frontal cortex and hippocampus. This study presents a novel method of employing two tissues that are evidenced by many studies to be the main tissues involved in SCZ as the target tissues,

and two brain tissues with less supporting evidence for involvement in SCZ as the control or contrast tissues. Furthermore, RNA-Seq datasets for all four regions were sampled from each donor to control for interregional expression differences that might stem from variation at eQTLs. This refined gene set, enriched for genes with region-specific, high expression in the frontal cortex and hippocampus, offers a valuable tool for future studies investigating the functional and biological underpinnings of any disorder affecting these critical brain regions.

Additionally, all RNA-Sequencing samples from GTEx were re-aligned and analysed against the latest Human Reference Genome and the latest gene annotation models, ensuring the most accurate and up-to-date analysis of the data. This meticulous approach resulted in improved detection and quantification of differentially expressed genes across the four brain regions, ultimately leading to more reliable downstream analysis and interpretation of the identified gene sets.

6.2.1 Discovering disease-relevant gene sets

One approach to identifying region-specific gene sets is to compare each gene's expression level in the target tissue to all other tissues, calculate a correlation (Fagerberg et al., 2014), and then take the top x% genes based on these statistics. This approach will find a 'fingerprint' region-specific expression profile in the context of all tissues tested. However, this assumes that expression levels are equally independent in all cells and that all highly expressed genes in relevant tissues can be treated equally. If the level of regulation of expression is variable between cells, the values for expression in each tissue cannot be normally distributed. Expression levels can also be skewed by outliers, which can be extreme if any sample had been mislabelled and is not detected in the study, for example, if a Cortex dataset mislabelled as a Cerebellar Hemisphere dataset (identified in Section 3.2.5), the mean and standard deviation of expression levels of each gene in the Cerebellar Hemisphere will be skewed by larger outlier values normally found in the Cortex. Genes with inflated averages could be included in a Cerebellar Hemisphere gene set, and in subsequent enrichment analyses, genes may show a false positive enrichment in this brain region. This emphasises the importance of re-analysing raw data where possible, rather than reliance on summary statistics, and the effectiveness of tissue-to-tissue PCA analysis to identify mislabelled samples.

For finding a gene set more relevant to specific neural diseases that affect a particular tissue with comorbidity with diseases seen in other tissues, the statistical comparison with all tissues may eliminate important genes. For example, if the expression of potassium channels is higher in muscle, heart, aorta, cerebellum, hypothalamus, and spinal cord, than in the frontal cortex, then these genes risk not being in the top x% of genes in the frontal cortex. With the possibility that there are many different secondary symptoms in SCZ patients caused by different pleiotropic effects of variants in multiple causal genes, a more targeted approach to tissue selection is required.

In this thesis, a novel approach was presented to identify disease-relevant gene sets for SCZ by tailoring the selection of target and contrast tissues. To minimise the exclusion of genes with pleiotropic effects due to potential comorbidity with other disorders, two brain tissues with less supporting evidence for involvement in SCZ were chosen as the control or contrast tissues. By implementing this tailored approach, the study aimed to uncover gene sets contributing to SCZ pathogenesis or pathology while accounting for the complex interplay between different brain regions and the potential for genes to have multiple effects across various tissues.

While both cortical and cerebellar hemisphere tissues are categorized as 'brain' in GTEx, their gene expression profiles show minimal overlap. An RNA-Seq analysis comparing these regions across 83 donors revealed that nearly half of all genes were differentially expressed between them (as detailed in Section 3.2.5). This stark contrast in gene expression can be attributed to the distinct neuronal composition of each region. In the frontal cortex, neurons comprise excitatory glutamatergic pyramidal and spiny stellate cells, alongside inhibitory GABAergic stellate, and basket interneurons. Pyramidal cells extend axonal processes to both subcortical regions, and contralateral and ipsilateral cortical regions, with cortical cells typically arranged in columns, sharing functional properties and connectivity patterns. While this columnar organization is common across many areas, its extent and prominence can vary. Conversely, the neurons in the cerebellar hemispheres are primarily Purkinje cells, characterized by expansive tree-like dendritic arborizations, along with millions of granule neurons, the most abundant cell type in the human brain. Together, these neurons collectively form floret-like structures, with additional contributions from interneurons such as Golgi cells. These cell types lack direct counterparts in the cortex.

Given the nearly exclusive gene expression patterns observed in each region, it is less likely that genes implicated in the pathogenesis of schizophrenia (SCZ) in the cortical regions would play a significant role in the cerebellar hemispheres. However, it's important to acknowledge that this may not apply universally to all SCZ cases. There may still be instances where the causal gene originates in cerebellar structures, indirectly impacting interregional signalling and contributing to a similar phenotype.

6.2.2 The selection of control regions

The cerebellar hemisphere and substantia nigra were chosen based on the lack of evidence of their involvement in SCZ from different brain imaging modalities and post-mortem studies, these two brain regions were among those least included in brain imaging studies. It is possible that genes with higher expression in the excluded regions are still play a role in a subset of patients with SCZ. This may be limited to one or two cell types in those regions with connections back to the frontal cortex or hippocampus. An example of this is the known SCZ risk gene, SETD1A (Singh et al., 2016). The two differential expression analyses between GTEx Frontal Cortex and Cerebellar Hemisphere samples in this study (Section 3.2.4 and 3.2.5.3), demonstrated that SETD1A is more highly expressed in the Cerebellar Hemisphere than the Frontal Cortex. This could be because high levels of gene expression in the cerebellum can cope with some loss of function in one copy of SETD1A. But in regions of the brain where SETD1A is less expressed or be strongly regulated at lower levels, then loss-of-function variants might effectively knock out the gene. While this implies that the method requires further refinement, this thesis did find supporting evidence. for other proteins, such as SMYD2 and DYDC2, that are also involved in methylation of histone 3 lysine 4 residues (Kummeling et al., 2020) discussed below in Section 6.4.3.

This also highlights a blind spot for studies like this that are looking for highly-expressed genes that will be affected by loss-of-function rare variants. Some rare variants may also lead to a gain-of-function in genes that are not normally expressed in the frontal cortex and hippocampus. This study was not designed to detect such cases. By focussing on highly-expressed genes, it would not have been possible to identify those genes with higher expression in patients with SCZ, but lower expression in controls. One approach to identify such genes would be to compare genes with lower expression in key brain regions to RNA-Seq differential expression analyses between cases and

controls, where up-regulated genes may hint towards a rare gain-of-function variants but can also be related to longitudinal alterations in the gene expression regulation caused by the disease state itself (Porcu et al., 2021).

6.2.3 Focus on four brain regions

As far as is known, no study has attempted to contrast two most-affected regions with two least-affected (control) regions to remove less likely gene candidates. Previous Exome studies have concurred with the selection of Frontal Cortex and Hippocampus as key brain regions in SCZ. However, this still potentially misses genes expressed exclusively in other brain regions containing unique cell types – such as the anterior cingulate cortex, amygdala, putamen, or caudate nucleus – that might also be important in some cases of SCZ. Performing differential expression analyses between all brain tissues in GTEx, followed by assessing variants against each of the comparisons, could be a valuable next step for a more comprehensive analysis, enabling the determination of which brain regions are most likely involved per case. GTEx tissue sites in the brain were selected for their involvement in dementias and neurodegenerative disorders, rather than for developmental or ongoing mental diseases. A wider range of tissue sites might produce better contrasts, especially if different tissues could be re-selected for sampling as least-affected sites. For example, the subiculum, multiple sample sites from around the prefrontal dorsolateral cortex, rather than the frontal pole.

6.2.4 Use of bulk tissue RNA-Seq

This approach averaged gene expression across all cell types within each tissue sampled from GTEx. As highlighted above the cortical regions contain a number of individual types of neuron and glial cells, as well as connective tissues, blood vessels and supporting cells. The set of highly expressed genes in bulk tissue may pick up genes that are highly expressed in each of the cell types. Single-cell RNA-Seq offers a more refined approach for identifying genes specifically enriched in relevant neuronal populations, such as glutamatergic neurons or parvalbumin interneurons, which may play a more crucial role in SCZ, as previously suggested in an enrichment of common variants from GWAS in single cell gene sets derived from mice (Skene et al., 2018). More recent studies published, since this thesis was well underway, have been leveraging the advances in single-cell sequencing the rapid increase in brain single-cell

data available for researchers. Studies comparing bulk RNA tissue samples, such as bulk hippocampus and single cell sequencing of the granule cell layer of the dentate gyrus in the hippocampal formation, demonstrate that a lot of important gene expression signals are lost in bulk RNA (Jaffe, Andrew E. et al., 2020).

Furthermore, recent efforts (Glinos et al., 2022) to analyse whole RNA transcripts from GTEx bulk RNA – currently limited to non-brain samples – using long-read sequencing will allow the study of transcripts and their differential expression between samples, rather than relying on gene level counts which hides patterns of alternate splicing between different samples. As covered in section 3.1.5.2 short-read sequencing without indexed reads is very challenging to deconvolute, especially with using a transcript modal biased towards general transcriptomes.

6.2.5 Differential Region Expression finds more region-specific transcripts

An extension of this thesis would be to analyse differences in alternate splicing between the most- and least-affected brain regions. Studies have shown that alternate splicing of RNA is most widely employed in the brain and liver (Yeo et al., 2004). RNA-Seq differential transcript expression (Glaus et al., 2012) or exon usage (Anders, Simon et al., 2012; Reyes et al., 2013) between the 4 areas would identify genes in which exons are expressed in one tissue but not in the control tissues. The transcriptome in the brain is incomplete (Zhang et al., 2020), complicating differential transcript expression analyses that rely on transcript annotation models. At the extreme level, base-level differential expression analysis, or differential region expression analysis, is computationally expensive but can detect differences in expression agnostic to annotation models that could infer a more complex splicing landscape (Collado-Torres, Leonardo et al., 2017). On top of the list of genes included in this study, these lower-level differential analyses could identify those genes which might not show higher gene expression overall in the most-affected regions, but in which some RNA transcripts are more exclusive to those tissues. The identified differential exons or regions within those genes can be examined for a burden of rare damaging variants.

6.2.6 Future work

The increased specificity for genes with higher expression in SCZ-relevant tissues moves closer to finding a suitable list of genes that more closely envelope those involved in SCZ. However, as knowledge advances, the focus needs to shift to genes

that are important to individual cell types that are important in SCZ. Previous work has shown some promise mapping GWAS genes to single cell RNA-Seq data in mice (Skene et al., 2018), and as discussed in 3.4.5, Skene's software, EWCE, can be used with custom datasets that can be built from data obtained from the Human Cell Atlas, for example. Rich data sets from regions across the brain can be compiled and tested for enrichment with risk genes, GWAS risk genes. This is a computationally intensive process, however the costs of these can be kept to a minimum by designing an efficient pipeline on an elastic cloud server. Gene set enrichment analyses in individual layers of the human cortex based on single-cell transcriptomics are now possible using *spatialLIBD* in R (Pardo et al., 2022) developed after the studies in this thesis was completed. Further analysis of the CO/HP-genes in this and future adaptations in *spatialLIBD* for the hippocampal formation could be performed.

6.3 Rare damaging variants exclusive to a cohort of schizophrenia cases

Ultra-rare variants exclusive to SCZ patients from the Swedish Exome cohort were identified by filtering genes specific to the frontal cortex and hippocampus. A key contribution of this work is the use of the most recent gnomAD (r2.1.1) for filtering common variants, as opposed to ExAC used in previous studies. This updated filtering method, although also employed by SCHEMA in their more recent publication (Singh et al., 2022), ensures a more accurate and up-to-date exclusion of common variants in the identification of ultra-rare, potentially damaging variants exclusive to SCZ patients. The remaining variants were analysed using Ensembl VEP, with a focus on those predicted to be pathogenic or deleterious by at least one of six *in silico* prediction tools or those with adverse effects on transcription or translation, such as stop gain/lost, frameshift, or splicing variants.

The resulting list of exclusive, potentially damaging variants provides a foundation for exploring rare variants among patients that could have a disruptive effect in the frontal cortex and hippocampus. This tailored method lays the groundwork for further exploration and analysis, with the aim of advancing our understanding of SCZ's underlying mechanisms and improving patient outcomes.

An important note here is that at the time of writing, gnomAD has released r4.1 with data from 400,000 genomes from the UK BioBank (Koenig et al., 2024) and includes a

set of common MNVs identified in the cohort (Wang, Q. et al., 2020). Future work would have to reanalyse the Swedish Exomes cohort filtering out common variants in gnomAD v4 to remove false positive ultra-rare variants.

6.3.1 SCZ is oligogenic for rare damaging variants in highly expressed genes

This study diverged from previous studies by restricting the search for rare damaging variants (RDVs) to the gene set identified in the differential expression analysis. Rather than test for enrichment of all rare variants in the gene set, RDVs were isolated to the list of highly expressed genes. A total of 4,398 RDVs were identified in the cohort of 3,500 cases. Under a third of the patients (N=1056) with SCZ did not have any RDVs in all 1,074 genes. Just over one third of the patients (N=1210) have a single RDV and the remainder (N=1234) between 2 and 15 RDVs. Two cases had 12 and 15 RDVs: one case with 12 RDVs in 11 genes (ADRA1B, B4GALNT3, CLSPN, DNAH11, GAS2L2, GRM5, KCNIP3, KIF21B, NRP1, SGSM2, STOML3), and one with 15 RDVs in 15 genes (BCL11B, CRYGA, CYP26A1, DGAT2, DLGAP4, DNAAF3, KIAA1522, LHX6, LRTM2, MAPK15, PTPRN2, RHOBTB2, SLCO4C1, TCHH, ZNF831) of which two frameshifts in PTPRN1 and MAPK15 were homozygous. In cases with more than 6 RDVs there were a total of 80 RDVs in 78 genes, demonstrating that among those with the most RDVs no common genes were affected.

Further analysis is required to characterise the distinctions between the groups of genes affected in cases compared to controls. This initial exploration did not involve examining individual cases; however, it sets the stage for future research to investigate gene-gene interactions and common pathway disruptions.

6.3.2 Some important genes are not represented by this gene set

The high number of patients without RDVs in highly expressed genes specific to affected tissues suggested that the list might not have captured all genes relevant to SCZ. For instance, the known SCZ risk gene, SETD1A (Singh et al., 2016) was missing, and there was incomplete coverage of all subunits for each of the signalling receptor complexes.

To illustrate this, the expression levels of the four glutamate AMPAR subunits (GRIA1-4; shown in Figure 6.1) in the four brain regions were examined. While GRIA3, which was included, had higher expression in both the frontal cortex and hippocampus

samples, GRIA4 showed highest expression in the cerebellar hemispheres. However, GRIA1 and GRIA2's expression patterns were inconclusive. Upon inspection, GRIA2 could have been included given its much higher expression in the frontal cortex compared to other regions.

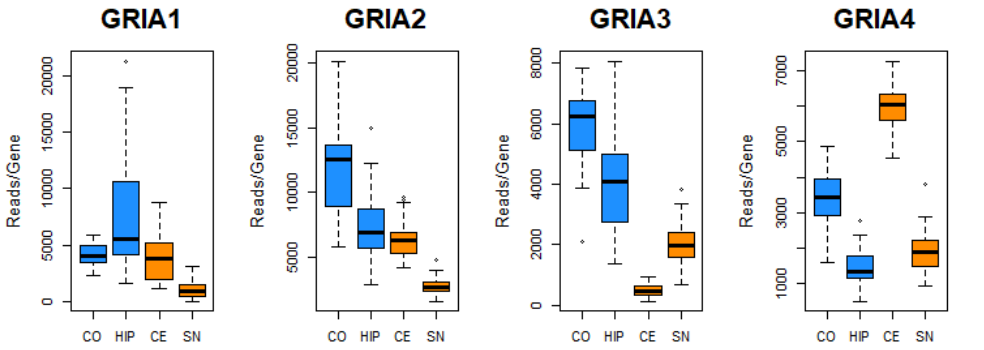

Figure 6.1 Regional expression of AMPAR subunits
Independent boxplots of reads per AMPAR gene per sample for four brain regions, CO, HP, CE, and SN. GRIA3 is a CO- and HP-gene. GRIA1, GRIA2 and GRIA4 are not in CO- or HP-genes.

Discrepancies in regional expression patterns of GRIA1 and GRIA2 compared to previous studies (Shen et al., 2020) were attributed to several experimental distinctions, primarily revolving around sample size and cell type specificity. Shen et al. (2020) used a limited sample size of only 6 healthy donor microarray samples from the Human Brain Map (HBM) at Allen Institute (Hawrylycz, Michael J. et al., 2012). In contrast, the analysis in this thesis used data from 25 healthy donor RNA-Sequencing samples, substantially enhancing statistical power and the ability to detect consistent gene signals across samples. The use of single-cell and microarray data in Shen et al. (2020) minimised variability in gene expression signals from other cell types, offering higher specificity despite the small sample size. In this thesis, bulk RNA-Seq samples from GTEx were used, combining signals from diverse cell types within the analysed tissues. While this approach introduced variability in gene expression profiles, the larger sample sizes helped mitigate these discrepancies. The interplay between sample sizes and cell type specificity highlights the importance of considering these factors when interpreting regional expression data. Further studies will be needed to

optimise the threshold for differential expression, but also adjusting how genes from different contrasts are assimilated.

One further possibility for missed genes, is that the GTEx region does not overlap well with brain regions involved in SCZ. This is highlighted by Jaffe et al. (2020) in which they discover missing signals (e.g., eQTLs, GRM3 and CACNA1C gene expression) from bulk Hippocampus, that they detect in single-cell sequencing of the dentate gyrus, which is located next to CA1 in the hippocampus. Frontal Cortex corresponds best with the frontal pole, the most anterior part of the frontal lobe. It is associated with various cognitive functions, such as social cognition, prospective memory, and high-level decision-making. But the DLPFC, considered the key area in the frontal cortex in SCZ, is located on the dorsolateral aspect of the frontal lobe and is involved in various higher cognitive functions, such as working memory, executive function, and cognitive control. Future single cell studies will be needed to understand whether there is heterogeneity of neuronal subtypes across the cortical regions that might underlie differential gene or transcript expression between these regions.

As hinted at in Section 6.2.2 above, these gaps may also be due to gain-of-function variants, or some patients may have structural variants or CNVs not detectable by short read exome sequencing, or variants in cis-regulatory sequences. This is a possibility given the high number of transcription factors identified in Section 5.3.7.2. As well as variants in the transcription factors themselves, variants within the target DNA binding site might have similar outcomes.

6.3.3 Combining scores from appropriate *in silico* prediction tools

The identification of potentially damaging variants can be significantly improved by using a combination of *in silico* prediction tools. Relying on a single tool may not be sufficient, as each tool has its strengths and limitations. By selecting multiple tools and employing a structured selection process, researchers can enhance their ability to identify and analyse rare damaging variants accurately.

In this study as in previous studies (Genovese et al., 2016; Purcell et al., 2014), a variety of software tools were used, including SIFT (Kumar et al., 2009), PolyPhen-2 (HumVar) (Adzhubei, Ivan A. et al., 2010), Ada-score and RF-score from dbscSNV (Jian et al., 2014), and REVEL (Ioannidis et al., 2016). These tools employ different algorithms and

criteria for evaluating the potential impact of genetic variants on protein function. Combining the results from these prediction methods enabled a more comprehensive and reliable assessment of the variants in our patient cohort.

Using a larger range of software has several advantages over relying on a single tool. This approach reduces the likelihood of overlooking potentially damaging variants due to the limitations or biases of a particular algorithm. It also enables researchers to cross-validate and compare the predictions generated by different tools, leading to increased confidence in the identified variants. Other tools to consider include Combined Annotation-Dependent Depletion (CADD) (Rentzsch et al., 2019) and the recent AlphaMissense (Cheng, J. et al., 2023). CADD itself is an integrative model which brings together multiple genomic annotations and scores into a single metric to predict the deleteriousness of genetic variants. AlphaMissense leverages deep learning and protein structures from AlphaFold to assess whether amino acid changes might have a damaging effect on protein structure. Machine learning-based methods have a slightly imperfect approximation in their prediction of pathogenicity because the data available for training on targets is incomplete and can be inaccurate (Rentzsch et al., 2019). Selecting a wide range of tools, and scoring variants based on the number of tools that predict pathogenicity might help to mitigate these problems.

A further use of AlphaMissense and AlphaFold (version 3 at the time of writing) would be to assess the aggregation of RDVs on the 3D model of each protein and their relevant isoforms. AlphaMissense already supplies a value to predict the extent of the perturbation in the secondary and tertiary forms of proteins. The hypothesis would be that RDVs distant from each other in the DNA template might be observed to aggregate to a common domain in the 3D isoform of the protein. These hotspots might cause obstructions. for example, of ligand-binding sites, or affect natural conformational changes when a protein is activated, or inhibit proper DNA or RNA binding, molecule recognition, or disrupt heavy metal binding.

6.4 Highlighted potential role of rare variants in common pathways

A novel approach of randomly sampling cases and controls in the cohort into smaller subgroups and aggregating RDV-containing genes by Gene Ontologies (biological processes, molecular functions, and cellular compartments) tested whether there were common GO terms consistently affected by RDVs between these groups. As well as locus and allelic heterogeneity, the findings suggest that there is also 'pathway heterogeneity' between cases. The results reinforce the importance of genes in synaptic formation, function, and maintenance in a subset of patients, but also suggests that variants in either upstream and downstream processes of signalling, such as gene expression and trafficking, downstream calcium signalling and adenylate cyclase inhibition in the postsynaptic membrane may be solely important in another subset of patients. Variants that disrupt axon outgrowth, guidance, or intercellular adhesion can affect the synapse strength or number, which may be equally important in pathogenesis, without requiring variants in synaptic transmission. Finally, this study also adds to growing evidence of involvement of intracellular signalling processes like WNT signalling (Vallee, 2022), RHO GTPase signalling (Sekiguchi et al., 2020), calcium signalling (Boczek et al., 2021), and pathways that involve adenylyl cyclase inhibition (Chen, J. et al., 2022) in the pathology of SCZ.

6.4.1 Signal transduction is the most consistently affected pathway

Signal transduction is a fundamental process underlying efficient neural communication. It enables the conversion of extracellular signals, such as neurotransmitters released by presynaptic neurons, into intracellular signals that modulate the activity of postsynaptic neurons. This precise translation is critical for all neural processes, from basic sensory integration to complex cognitive functions. Dysregulation of glutamatergic or GABAergic signalling has been linked to the pathophysiology of schizophrenia (Luvsannyam et al., 2022). Uncovering the intricate mechanisms of signal transduction at excitatory synapses holds promise for a deeper understanding of this complex disorder. These findings may contribute to the development of targeted therapeutic interventions. While it is a broad category, the results in this thesis do not point exclusively to glutamatergic or GABAergic synapses, but also implicate serotonin, adrenergic, cholinergic signalling, underscoring the

complexity of multiple pathways and their potential relevance to effective personalised drug treatments.

6.4.2 The complement system pathway

One well-studied pathway is the reduction of dendritic spines through alterations in C4 in SCZ patients suggests that this is important in the risk of developing SCZ (Bohlson & Tenner, 2023; Woo et al., 2020) as it drastically reduces the number of synapses that can be formed. CSMD1 and SUSD4 are two genes involved in the regulation of the complement system (Bohlson & Tenner, 2023; Gialeli et al., 2018) and both are represented in the results, with RDVs in CSMD1 affecting 41 cases (approximately 1.2% of cases), the most common gene affected in cases, and SUSD4 in another 5 cases. Additionally, there are other cases with RDVs in HOMER1, another gene of interest in the context of dendritic spines growth in SCZ and other neuropsychiatric disorders. Homer1 is a postsynaptic density-localized scaffolding protein that interacts with Shank3 to regulate metabotropic glutamate receptor activity and promote dendritic spine growth (Yoon et al., 2021). This highlights the effectiveness of aggregating variants into pathways to identify common pathways.

6.4.3 Epigenetic and expression regulatory pathways affect almost 10% of cases

As discussed above in Section 6.2.2, SETD1A (Singh et al., 2016) was missing from the CO/HP-genes. It is involved in epigenetic signals and histone modifications that could contribute to the risk of developing SCZ. The SCHEMA consortium identified variants involved in transcriptional regulation and chromatin modification including SETD1A as well as SP4, RB1CC1 and XPO7 (Singh et al., 2022). However, this thesis identified RDVs in related genes, including DYDC2, PAX7, SMYD2, GAD1, that are also associated with methylation at lysine-4 residues on histone-3, enhancer regions marked by these epigenetic markers, or genes involved in chromatin remodelling (including CHD5, MTA3, HIPK4, PRDM8). This strengthens the evidence for H3K4me involvement in SCZ and epigenetic mechanisms in general (Akbarian, 2014; Gusev et al., 2019). Furthermore, 333 RDVs were found in 93 genes involved in transcriptional regulation or transcription complexes (including PAX7, CUX2, SATB1-2, TBR1, EMX1). These 333 RDVs are found in 315 patients representing almost 10% of all cases. This strengthens the importance of transcriptional regulation in SCZ, either additional to synapse or other pathway involvement or not. This also could suggest that while certain genes

may show genome-wide significance, these genes only account for a small number of patients. The search for variants must consider all members of the pathways affected by those genes – for example SETD1B, as well as SETD1A; or G-protein receptor regulators, RGS20 and RGS4, identified in this study but RGS3 identified in a family with early-onset schizophrenia (Kanwal et al., 2022).

6.4.4 Combinations of affected pathways may lead to different neuropsychiatric diseases

Variants in genes affecting similar pathways have been implicated in multiple neuropsychiatric diseases, including autism, SCZ, bipolar disease, and epilepsy. Many genes implicated in SCZ was also found to be associated with bipolar disease. Genes with RDVs in this study also include 9 RDVs in two sodium channels (SCN1A, SCN3B) and 97 RDVs in 26 potassium channels (KCNA4, KCNA6, KCNB1, KCNB2, KCNC2, KCNF1, KCNG2, KCNH3, KCNH4, KCNH5, KCNIP2, KCNIP3, KCNJ1, KCNJ4, KCNMA1, KCNN1, KCNQ3, KCNQ5, KCNS1, KCNS2, KCNV1, KCTD1, KCTD13, KCTD16, KCTD17, KCTD4), some of which have been implicated in both epilepsy (Gao et al., 2022) and autism (Cheng, P. et al., 2021). As with all previously mentioned genes, the 102 cases affected by RDVs in these genes are only affected by one RDV in any of them, except in a few cases with two RDVs. This suggests that an RDV in any of these genes can be disruptive for signalling overall, but it is not possible to determine whether these are risk variants for SCZ or a comorbid, underlying risk for epilepsy (Adachi & Ito, 2022). Further still, many of these channels are also implicated in cardiovascular disease which could be pleiotropic effects of variants in a subset of patients (Bernardi et al., 2020).

6.4.5 Future work

Several avenues for further investigation arise from the findings in this thesis. First, conducting a more extensive analysis of case and controls could provide additional insights into the relationship between rare genetic variants and SCZ pathology. By mapping out genes in pathways using a large network diagram and highlighting the genes affected in patients, patterns may be revealed in how they relate in the larger picture. Additionally, examining the affected amino acid sequences in the final three-dimensional protein structures could uncover whether seemingly disconnected RDVs may, in fact, be neighbours in final protein complex, potentially explaining why similar phenotypes arise from different RDVs.

Another promising direction for future research involves comparing groups of variants and their affected pathways with the individual patient phenotypes or symptom clusters. This approach would require a large-scale project with closer patient contact and detailed medical histories to determine whether identified variants are causal to any secondary symptoms. Subgroup analyses could focus on patients with RDVs in similar gene families, such as potassium and sodium channels, or those with RDVs in synaptic genes related to glutamatergic or GABAergic transmission, as well as pre- or post-synaptic genes.

Lastly, it would be valuable to compile a comprehensive list of secondary symptoms potentially caused by pleiotropic effects of the original RDVs. This information could inform diagnosis processes, enabling physicians to request targeted gene sequencing based on a patient's specific symptom profile. For instance, a patient presenting with both cardiovascular issues and long-term psychosis, prioritising the analysis of sodium and potassium channels along with other genes in their respective pathways, could help focus the search for RDVs and facilitate a more accurate diagnosis and tailored treatment strategies.

6.5 Conclusion

The work described in this thesis adds evidence to the oligogenic nature of SCZ, the idea that specific combinations of rare genetic variants can determine the manifestation and diverse range of symptoms seen in neuropsychiatric disorders, blurring the boundaries between these diseases. The findings in this study offer a fresh perspective on the complex and immense genetic landscape of schizophrenia. With trillions of neurons and even more connections between them, even subtle disruptions in these networks can have profound effects on perception, thought, and behaviour. By highlighting the impact of ultra-rare damaging variants on neural pathways, this thesis demonstrates the potential for subtle genetic disruptions to significantly contribute to disease manifestation and symptomatology.

This research also offers a unique contribution to the field by showcasing the potential that studying interregional differences in gene expression in healthy individuals has to serve as models for understanding the dysfunction observed in neuropsychiatric disorders. This approach and the findings provide valuable insights into the complex

aetiology of SCZ, emphasising the need for more personalised medicine in targeting specific pathways in subsets of patients for more effective and precise treatment strategies.

6.5.1 Personalised medicine is required for effective drug treatment

In the context of SCZ treatment, the genetic architecture of the disorder must be considered. The findings in this thesis offer a unique perspective that despite SCZ being highly polygenic across the patient population, it exhibits 'oligogenicity' at the individual level, wherein a limited number of genetic variants significantly contribute to the disease. This genetic complexity also underscores the importance of personalized medicine in targeting specific biological pathways with the most appropriate pharmacological interventions for each patient.

A tailored treatment approach has the potential to improve therapeutic outcomes, minimize adverse effects, and ultimately enhance the quality of life for individuals living with schizophrenia. As our understanding of the disorder's intricacies continues to advance, precision medicine emerges as a promising strategy to address the challenges posed by this heterogeneous condition.

7 References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68–74. https://doi.org/10.1038/nature15393 [doi]

Adachi, N., & Ito, M. (2022). Epilepsy in patients with schizophrenia: Pathophysiology and basic treatments. *Epilepsy & Behavior : E&B, 127*, 108520. https://doi.org/10.1016/j.yebeh.2021.108520

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork P FAU - drashov, A., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248 [doi]

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, r., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Akbarian, S. (2014). Epigenetic mechanisms in schizophrenia. *Dialogues in Clinical Neuroscience, 16*(3), 405–417. https://doi.org/10.31887/DCNS.2014.16.3/sakbarian

Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Juettemann, T., Keenan, S., Laird, M. R., . . . Flicek, P. (2017). Ensembl 2017. *Nucleic Acids Research, 45*(D1), D635–D642. https://doi.org/10.1093/nar/gkw1104 [doi]

Allen Institute for Brain Science. (2010). Allen Human Brain Atlas.

Amit, I., Ardlie, K., Arzuaga, F., Awandare, G., Bader, G., Bernier, A., Carninci, P., Donnelly, S., Eils, R., Forrest, A. R. R., Greely, H. T., Guigo, R., Hacohen, N., Haniffa, M., Kirby, E. S., Knoppers, B. M., Kriegstein, A., Lein, E. S., Linnarsson, S., . . . Zhuang, X. (2024). The commitment of the human cell atlas to humanity. *Nature Communications, 15*(1), 10019–x. https://doi.org/10.1038/s41467-024-54306-x

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England), 31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638 [doi]

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research, 22*(10), 2008–2017. https://doi.org/10.1101/gr.133744.111

Andrews, S. (2015). FastQC: A Quality Control Tool for High Throughput Sequence Data [computer software]

Bahner, F., & Meyer-Lindenberg, A. (2017). Hippocampal-prefrontal connectivity as a translational phenotype for schizophrenia. *European Neuropsychopharmacology*

: The Journal of the European College of Neuropsychopharmacology, 27(2), 93–106. https://doi.org/10.1016/j.euroneuro.2016.12.007

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods, 14*(2), 135–139. https://doi.org/10.1038/nmeth.4106 [doi]

Batiuk, M. Y., Tyler, T., Dragicevic, K., Mei, S., Rydbirk, R., Petukhov, V., Deviatiiarov, R., Sedmak, D., Frank, E., Feher, V., Habek, N., Hu, Q., Igolkina, A., Roszik, L., Pfisterer, U., Garcia-Gonzalez, D., Petanjek, Z., Adorjan, I., Kharchenko, P. V., & Khodosevich, K. (2022). Upper cortical layer-driven network impairment in schizophrenia. *Science Advances, 8*(41), eabn8367. https://doi.org/10.1126/sciadv.abn8367

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 (Oxford, England), 32*(10), 1493–1501. https://doi.org/10.1093/bioinformatics/btw018

Bernardi, J., Aromolaran, K. A., & Aromolaran, A. S. (2020). Neurological Disorders and Risk of Arrhythmia. *International Journal of Molecular Sciences, 22*(1), 188. doi: 10.3390/ijms22010188. https://doi.org/10.3390/ijms22010188

Bernardin, F., Schwitzer, T., Angioi-Duprez, K., Giersch, A., Jansen, C., Schwan, R., & Laprevote, V. (2020). Retinal ganglion cells dysfunctions in schizophrenia patients with or without visual hallucinations. *Schizophrenia Research, 219*, 47–55. https://doi.org/10.1016/j.schres.2019.07.007

Besteher, B., Gaser, C., Spalthoff, R., & Nenadic, I. (2017). Associations between urban upbringing and cortical thickness and gyrification. *Journal of Psychiatric Research, 95*, 114–120. https://doi.org/S0022-3956(17)30528-9 [pii]

Birnbaum, R., & Weinberger, D. R. (2017). Genetic insights into the neurodevelopmental origins of schizophrenia. *Nature Reviews.Neuroscience, 18*(12), 727–740. https://doi.org/10.1038/nrn.2017.125

Blighe, K., Rana, S., & Lewis, M. (2019). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. *R Package Version, 1.6.0* https://bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html

Boczek, T., Mackiewicz, J., Sobolczyk, M., Wawrzyniak, J., Lisek, M., Ferenc, B., Guo, F., & Zylinska, L. (2021). The Role of G Protein-Coupled Receptors (GPCRs) and Calcium Signaling in Schizophrenia. Focus on GPCRs Activated by Neurotransmitters and Chemokines. *Cells, 10*(5), 1228. doi: 10.3390/cells10051228. https://doi.org/10.3390/cells10051228

Bohlson, S. S., & Tenner, A. J. (2023). Complement in the Brain: Contributions to Neuroprotection, Neuronal Plasticity, and Neuroinflammation. *Annual Review of Immunology, 41*, 431–452. https://doi.org/10.1146/annurev-immunol-101921-035639

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England), 30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 [doi]

Borkowska, M., Millar, J. K., & Price, D. J. (2016). Altered Disrupted-in-Schizophrenia-1 Function Affects the Development of Cortical Parvalbumin Interneurons by an Indirect Mechanism. *PloS One, 11*(5), e0156082. https://doi.org/10.1371/journal.pone.0156082

BrainSeq: A Human Brain Genomics Consortium. (2015). BrainSeq: Neurogenomics to Drive Novel Target Discovery for Neuropsychiatric Disorders. *Neuron, 88*(6), 1078–1083. https://doi.org/https://doi.org/10.1016/j.neuron.2015.10.047

BrainSeq: A Human Brain Genomics Consortium. (2018, 31 Ma). *LieberInstitute/BrainSeq_Phase1: code from brainseq phase 1 paper.* GitHub. Retrieved 25 July 2019, from https://github.com/LieberInstitute/BrainSeq_Phase1/

BrainSeq: A Human Brain Genomics Consortium. (2019, 21 Au). *BrainSeq Phase1 DLPFC - syn12299750 - Wiki.* Synapse.org. Retrieved 25 July 2019, from https://www.synapse.org/#!Synapse:syn12299750/wiki/551288

Breitmeyer, R., Vogel, S., Heider, J., Hartmann, S., Wust, R., Keller, A., Binner, A., Fitzgerald, J. C., Fallgatter, A. J., & Volkmer, H. (2023). Regulation of synaptic connectivity in schizophrenia spectrum by mutual neuron-microglia interaction. *Communications Biology, 6*(1), 472–9. https://doi.org/10.1038/s42003-023-04852-9

Broad Institute. (2020, 25 Ma). *GitHub: broadinstitute/gtex-pipeline.* https://github.com/broadinstitute/gtex-pipeline/tree/gtex_v8/rnaseq

Broekema, R. V., Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biology, 10*(1), 190221. https://doi.org/10.1098/rsob.190221

Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews.Genetics, 21*(10), 630–644. https://doi.org/10.1038/s41576-020-0258-4 [doi]

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z [doi]

Cantor, R. M. (2019). 8 - Analysis of Genetic Linkage. In R. E. Pyeritz, B. R. Korf & W. W. Grody (Eds.), *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics (Seventh Edition)* (pp. 227–236). Academic Press. https://doi.org/10.1016/B978-0-12-812537-3.00008-1

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., Compton, C. C., DeLuca, D. S., Peter-Demchok, J., Gelfand, E. T., Guan, P., Korzeniewski, G. E., Lockhart, N. C., Rabiner, C. A., Rao, A. K., Robinson, K. L., Roche, N. V., Sawyer, S. J., Segre, A. V., . . . GTEx Consortium. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking, 13*(5), 311–319. https://doi.org/10.1089/bio.2015.0032 [doi]

Carithers, L. J., & Moore, H. M. (2015). The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking, 13*(5), 307–308. https://doi.org/10.1089/bio.2015.29031.hmm [doi]

Carlen, M. (2017). What constitutes the prefrontal cortex? *Science (New York, N.Y.), 358*(6362), 478–482. https://doi.org/10.1126/science.aan8868 [doi]

Casas, B. S., Arancibia-Altamirano, D., Acevedo-La Rosa, F., Garrido-Jara, D., Maksaev, V., Perez-Monje, D., & Palma, V. (2022). It takes two to tango: Widening our understanding of the onset of schizophrenia from a neuro-angiogenic perspective. *Frontiers in Cell and Developmental Biology, 10*, 946706. https://doi.org/10.3389/fcell.2022.946706

Castel, S. E., Aguet, F., Mohammadi, P., GTEx Consortium, Ardlie, K. G., & Lappalainen, T. (2020). A vast resource of allelic expression data spanning human tissues. *Genome Biology, 21*(1), 234–z. https://doi.org/10.1186/s13059-020-02122-z [doi]

Chang, X., Lima, L. A., Liu, Y., Li, J., Li, Q., Sleiman, P. M. A., & Hakonarson, H. (2018). Common and Rare Genetic Risk Factors Converge in Protein Interaction Networks Underlying Schizophrenia. *Frontiers in Genetics, 9*, 434. https://doi.org/10.3389/fgene.2018.00434 [doi]

Chang, X., Liu, Y., Hahn, C. G., Gur, R. E., Sleiman, P. M. A., & Hakonarson, H. (2017). RNA-seq analysis of amygdala tissue reveals characteristic expression profiles in schizophrenia. *Translational Psychiatry, 7*(8), e1203. https://doi.org/10.1038/tp.2017.154 [doi]

Chang, X., Collin, G., Xi, Y., Cui, L., Scholtens, L. H., Sommer, I. E., Wang, H., Yin, H., Kahn, R. S., & van den Heuvel, M. P. (2017). Resting-state functional connectivity in medication-naive schizophrenia patients with and without auditory verbal hallucinations: A preliminary report. *Schizophrenia Research, 188*, 75–81. https://doi.org/10.1016/j.schres.2017.01.024

Chen, H., & Boutros, P. C. (2018). *Generate High-Resolution Venn and Euler Plots - Package 'VennDiagram'*. https://cran.rproject.org/web/packages/VennDiagram/VennDiagram.pdf

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics, 12*, 35–35. https://doi.org/10.1186/1471-2105-12-35 [doi]

Chen, J., Ding, Q., An, L., & Wang, H. (2022). Ca(2+)-stimulated adenylyl cyclases as therapeutic targets for psychiatric and neurodevelopmental disorders. *Frontiers in Pharmacology, 13*, 949384. https://doi.org/10.3389/fphar.2022.949384

Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulyte, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Z. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science (New York, N.Y.), 381*(6664), eadg7492. https://doi.org/10.1126/science.adg7492

Cheng, P., Qiu, Z., & Du, Y. (2021). Potassium channels and autism spectrum disorder: An overview. *International Journal of Developmental Neuroscience : The Official*

Journal of the International Society for Developmental Neuroscience, 81(6), 479–491. https://doi.org/10.1002/jdn.10123

Cheslack-Postava, K., & Brown, A. S. (2022). Prenatal infection and schizophrenia: A decade of further progress. *Schizophrenia Research, 247*, 7–15. https://doi.org/10.1016/j.schres.2021.05.014

Chi, Y., Qi, R., Zhou, Y., Tong, H., Jin, H., Turck, C. W., Chen, W., & Wang, G. (2023). scBrainMap: a landscape for cell types and associated genetic markers in the brain. *Database : The Journal of Biological Databases and Curation, 2023*, 10.1093/database/baad035. https://doi.org/10.1093/database/baad035

Choi, J., Kim, S., Kim, J., Son, H., Yoo, S., Kim, C., Park, Y. J., Moon, S., Cha, B., Jeon, M. C., Park, K., Yun, J. M., Cho, B., Kim, N., Kim, C., Kwon, N., Park, Y. J., Matsuda, F., Momozawa, Y., . . . Im, S. (2023). A whole-genome reference panel of 14,393 individuals for East Asian populations accelerates discovery of rare functional variants. *Science Advances, 9*(32), eadg6319. https://doi.org/10.1126/sciadv.adg6319

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly, 6*(2), 80–92. https://doi.org/10.4161/fly.19695 [doi]

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews.Genetics, 11*(6), 415–425. https://doi.org/10.1038/nrg2779 [doi]

Coelewij, L., & Curtis, D. (2018). Mini-review: Update on the genetics of schizophrenia. *Annals of Human Genetics, 82*(5), 239–243. https://doi.org/10.1111/ahg.12259 [doi]

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*(2), 203–216.e8. https://doi.org/S0896-6273(19)30438-6 [pii]

Collado-Torres, L., Nellore, A., Frazee, A. C., Wilks, C., Love, M. I., Langmead, B., Irizarry, R. A., Leek, J. T., & Jaffe, A. E. (2017). Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Research, 45*(2), e9. https://doi.org/10.1093/nar/gkw852

Corvin, A., Craddock, N., & Sullivan, P. F. (2010). Genome-wide association studies: a primer. *Psychological Medicine, 40*(7), 1063–1077. https://doi.org/10.1017/S0033291709991723

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin

Fioretto, L., . . . Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research, 50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Curtis, D., Coelewij, L., Liu, S., Humphrey, J., & Mott, R. (2018). Weighted Burden Analysis of Exome-Sequenced Case-Control Sample Implicates Synaptic Genes in Schizophrenia Aetiology. *Behavior Genetics, 48*(3), 198–208. https://doi.org/10.1007/s10519-018-9893-3

Danecek, P., & McCarthy, S. A. (2017). BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics (Oxford, England), 33*(13), 2037–2039. https://doi.org/10.1093/bioinformatics/btx100 [doi]

de Hollander, G., Keuken, M. C., & Forstmann, B. U. (2015). The subcortical cocktail problem; mixed signals from the subthalamic nucleus and substantia nigra. *PloS One, 10*(3), e0120572. https://doi.org/10.1371/journal.pone.0120572

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., & Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology, 8*(1), e1000294. https://doi.org/10.1371/journal.pbio.1000294

Ding, S., Royall, J. J., Sunkin, S. M., Ng, L., Facer, B. A. C., Lesnar, P., Guillozet-Bongaarts, A., McMurray, B., Szafer, A., Dolbeare, T. A., Stevens, A., Tirrell, L., Benner, T., Caldejon, S., Dalley, R. A., Dee, N., Lau, C., Nyhus, J., Reding, M., . . . Lein, E. S. (2016). Comprehensive cellular-resolution atlas of the adult human brain. *The Journal of Comparative Neurology, 524*(16), 3127–3481. https://doi.org/10.1002/cne.24080

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 (Oxford, England), 29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635 [doi]

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics, 51*, 11.14.1–11.14.19. https://doi.org/10.1002/0471250953.bi1114s51 [doi]

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Durkie, M., Cassidy, E., Berry, I., Owens, M., Turnbull, C., Scott, R. H., Taylor, R. W., Deans, Z. C., Ellard, S., Baple, E. L., & Mcmullan, D. J. (2024). *ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2024.* .ACGS. https://www.acgs.uk.com/media/12533/uk-practice-guidelines-for-variant-classification-v12-2024.pdf

Edenberg, H. J. (2007). The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Research & Health : The Journal of the National Institute on Alcohol Abuse and Alcoholism, 30*(1), 5–13.

Ellard, S., Morgan, S., Wynn, S. L., Walker, S., Parrish, A., Mein, R., Juett, A., Ahn, J. W., Berry, I., Cassidy, E., Durkie, M., Fish, L., Hall, R., Howard, E., Rankin, J., Wright, C. F., Deans, Z. C., Scott, R. H., Hill, S. L., . . . Association for Clinical Genomic Science

Rare Disease Position Statement Working Group. (2024). Rare disease genomic testing in the UK and Ireland: promoting timely and equitable access. *Journal of Medical Genetics, 61*(12), 1103–1112. https://doi.org/10.1136/jmg-2024-110228

EMBL-EBI. (n.d., n.d.). *Assembly Converter - Homo sapiens - Ensembl Genome browser 97.* Retrieved 2019, from https://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G. H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., . . . Stefansson, K. (2008). Genetics of gene expression and its effect on disease. *Nature, 452*(7186), 423–428. https://doi.org/10.1038/nature06758 [doi]

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57–74. https://doi.org/10.1038/nature11247

Ermis Akyuz, E., & Bell, S. M. (2022). The Diverse Role of CUB and Sushi Multiple Domains 1 (CSMD1) in Human Diseases. *Genes, 13*(12), 2332. doi: 10.3390/genes13122332. https://doi.org/10.3390/genes13122332

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fagerberg, L., Hallstrom, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjostedt, E., Lundberg, E., Szigyarto, C. A., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., . . . Uhlen, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics : MCP, 13*(2), 397–406. https://doi.org/10.1074/mcp.M113.035600

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P., Lareau, C., Shoresh, N., Genovese, G., Saunders, A., Macosko, E., Pollack, S., Brainstorm Consortium, Perry, J. R. B., Buenrostro, J. D., Bernstein, B. E., Raychaudhuri, S., . . . Price, A. L. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics, 50*(4), 621–629. https://doi.org/10.1038/s41588-018-0081-4

Frank, R. A. W., McRae, A. F., Pocklington, A. J., van de Lagemaat, L. N., Navarro, P., Croning, M. D. R., Komiyama, N. H., Bradley, S. J., Challiss, R. A. J., Armstrong, J. D., Finn, R. D., Malloy, M. P., MacLean, A. W., Harris, S. E., Starr, J. M., Bhaskar, S. S., Howard, E. K., Hunt, S. E., Coffey, A. J., . . . Grant, S. G. N. (2011). Clustered coding variants in the glutamate receptor complexes of individuals with schizophrenia and bipolar disorder. *PloS One, 6*(4), e19011. https://doi.org/10.1371/journal.pone.0019011

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., . .

. Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research, 47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955 [doi]

Frankle, W. G., Lerma, J., & Laruelle, M. (2003). The synaptic hypothesis of schizophrenia. *Neuron, 39*(2), 205–216.

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., . . . O'Donovan, M. C. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature, 506*(7487), 179–184. https://doi.org/10.1038/nature12929

Ganzola, R., Maziade, M., & Duchesne, S. (2014). Hippocampus and amygdala volumes in children and young adults at high-risk of schizophrenia: research synthesis. *Schizophrenia Research, 156*(1), 76–86. https://doi.org/10.1016/j.schres.2014.03.030

Gao, K., Lin, Z., Wen, S., & Jiang, Y. (2022). Potassium channels and epilepsy. *Acta Neurologica Scandinavica, 146*(6), 699–707. https://doi.org/10.1111/ane.13695

Garrett, L., Trumbach, D., Spielmann, N., Wurst, W., Fuchs, H., Gailus-Durner, V., Hrabe de Angelis, M., & Holter, S. M. (2022). A rationale for considering heart/brain axis control in neuropsychiatric disease. *Mammalian Genome : Official Journal of the International Mammalian Genome Society,* https://doi.org/10.1007/s00335-022-09974-9

Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research, 49*(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113 [doi]

Genovese, G., Fromer, M., Stahl, E. A., Ruderfer, D. M., Chambert, K., Landen, M., Moran, J. L., Purcell, S. M., Sklar, P., Sullivan, P. F., Hultman, C. M., & McCarroll, S. A. (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature Neuroscience, 19*(11), 1433–1441. https://doi.org/10.1038/nn.4402 [doi]

Gialeli, C., Gungor, B., & Blom, A. M. (2018). Novel potential inhibitors of complement system and their roles in complement regulation and beyond. *Molecular Immunology, 102*, 73–83.

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., . . . D'Eustachio, P. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Research, 50*(D1), D687–D692. https://doi.org/10.1093/nar/gkab1028

Giordano, G. M., Caporusso, E., Pezzella, P., & Galderisi, S. (2022). Updated perspectives on the clinical significance of negative symptoms in patients with schizophrenia. *Expert Review of Neurotherapeutics, 22*(7), 541–555. https://doi.org/10.1080/14737175.2022.2092402

Glaus, P., Honkela, A., & Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics (Oxford, England), 28*(13), 1721–1728. https://doi.org/10.1093/bioinformatics/bts260

Glausier, J. R., & Lewis, D. A. (2018). Mapping pathologic circuitry in schizophrenia. *Handbook of Clinical Neurology, 150*, 389–417. https://doi.org/B978-0-444-63639-3.00025-6 [pii]

Glinos, D. A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K. L., Garimella, K., Bowers, T., Costello, M., Ardlie, K., Jian, R., Tucker, N. R., Ellinor, P. T., Harrington, E. D., Tang, H., Snyder, M., . . . Cummings, B. B. (2022). Transcriptome variation in human tissues revealed by long-read sequencing. *Nature, 608*(7922), 353–359. https://doi.org/10.1038/s41586-022-05035-y

Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., Stoeckert, C. J., Hogenesch, J. B., & Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics (Oxford, England), 27*(18), 2518–2528. https://doi.org/10.1093/bioinformatics/btr427 [doi]

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics, 45*(6), 580–585. https://doi.org/10.1038/ng.2653 [doi]

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.), 348*(6235), 648–660. https://doi.org/10.1126/science.1262110 [doi]

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.), 369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776 [doi]

GTEx Symposium. (20 May 2015). The GTEx Symposium: all things considered - biospecimens, omics data and ethical issues / National Cancer Institute, National Institute for Mental Health, National Human Genome Research Institute, NIH Common Fund. Paper presented at the https://videocast.nih.gov/Summary.asp?file=19024&bhcp=1

Gusev, F. E., Reshetov, D. A., Mitchell, A. C., Andreeva, T. V., Dincer, A., Grigorenko, A. P., Fedonin, G., Halene, T., Aliseychik, M., Filippova, E., Weng, Z., Akbarian, S., & Rogaev, E. I. (2019). Chromatin profiling of cortical neurons identifies individual epigenetic signatures in schizophrenia. *Translational Psychiatry, 9*(1), 256–1. https://doi.org/10.1038/s41398-019-0596-1

Haber, S. N. (2016). Corticostriatal circuitry. *Dialogues in Clinical Neuroscience, 18*(1), 7–21. https://doi.org/10.31887/DCNS.2016.18.1/shaber [doi]

Haddad, L., Schafer, A., Streit, F., Lederbogen, F., Grimm, O., Wust, S., Deuschle, M., Kirsch, P., Tost, H., & Meyer-Lindenberg, A. (2015). Brain structure correlates of urban upbringing, an environmental risk factor for schizophrenia. *Schizophrenia Bulletin, 41*(1), 115–122. https://doi.org/10.1093/schbul/sbu072 [doi]

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C. M. P., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Brain volumes in schizophrenia: a meta-analysis in over 18 000

subjects. *Schizophrenia Bulletin, 39*(5), 1129–1138. https://doi.org/10.1093/schbul/sbs118 [doi]

Hashimoto, N., Ito, Y. M., Okada, N., Yamamori, H., Yasuda, Y., Fujimoto, M., Kudo, N., Takemura, A., Son, S., Narita, H., Yamamoto, M., Tha, K. K., Katsuki, A., Ohi, K., Yamashita, F., Koike, S., Takahashi, T., Nemoto, K., Fukunaga, M., . . . COCORO. (2017). The effect of duration of illness and antipsychotics on subcortical volumes in schizophrenia: Analysis of 778 subjects. *NeuroImage.Clinical, 17*, 563–569. https://doi.org/S2213-1582(17)30284-X [pii]

Hatzimanolis, A., McGrath, J. A., Wang, R., Li, T., Wong, P. C., Nestadt, G., Wolyniec, P. S., Valle, D., Pulver, A. E., & Avramopoulos, D. (2013). Multiple variants aggregate in the neuregulin signaling pathway in a subset of schizophrenia patients. *Translational Psychiatry, 3*, e264. https://doi.org/10.1038/tp.2013.33 [doi]

Hawrylycz, M. J., Lein, E. S., Abajian, C., Beckmann, C. F., Bernard, A., Bertagnolli, D., Boe, A. F., Cartagena, P. M., MALLAR CHAKRAVARTY, M., CHAPIN, M., CHONG, J., DALLEY, R. A., GUILLOZET-BONGAARTS, A. L., DAVID DALY, B., CHINH DANG, DATTA, S., DEE, N., DOLBEARE, T. A., FABER, V., . . . Riley, Z. L. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature, 489*(7416), 391–399. https://doi.org/10.1038/nature11405

Hawrylycz, M., Martone, M. E., Ascoli, G. A., Bjaalie, J. G., Dong, H., Ghosh, S. S., Gillis, J., Hertzano, R., Haynor, D. R., Hof, P. R., Kim, Y., Lein, E., Liu, Y., Miller, J. A., Mitra, P. P., Mukamel, E., Ng, L., Osumi-Sutherland, D., Peng, H., . . . Zingg, B. (2023). A guide to the BRAIN Initiative Cell Census Network data ecosystem. *PLoS Biology, 21*(6), e3002133. https://doi.org/10.1371/journal.pbio.3002133

Hibar, D. P., Adams, H. H. H., Jahanshad, N., Chauhan, G., Stein, J. L., Hofer, E., Renteria, M. E., Bis, J. C., Arias-Vasquez, A., Ikram, M. K., Desrivieres, S., Vernooij, M. W., Abramovic, L., Alhusaini, S., Amin, N., Andersson, M., Arfanakis, K., Aribisala, B. S., Armstrong, N. J., . . . Ikram, M. A. (2017). Novel genetic loci associated with hippocampal volume. *Nature Communications, 8*, 13624. https://doi.org/10.1038/ncomms13624 [doi]

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivieres, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., Aribisala, B. S., Armstrong, N. J., Bernard, M., Bohlken, M. M., Boks, M. P., Bralten, J., Brown, A. A., Chakravarty, M. M., Chen, Q., . . . Medland, S. E. (2015). Common genetic variants influence human subcortical brain structures. *Nature, 520*(7546), 224–229. https://doi.org/10.1038/nature14101 [doi]

Howes, O. D., & Onwordi, E. C. (2023). The synaptic hypothesis of schizophrenia version III: a master mechanism. *Molecular Psychiatry, 28*(5), 1843–1856. https://doi.org/10.1038/s41380-023-02043-w

Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M., & Haussler, D. (2006). The UCSC Known Genes. *Bioinformatics (Oxford, England), 22*(9), 1036–1046. https://doi.org/10.1093/bioinformatics/btl048

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, P., Xi, Y., Lu, Z., Chen, Y., Li, X., Li, W., Zhu, X., Cui, L., Tan, Q., Liu, W., Li, C., Miao, D., & Yin, H. (2015). Decreased bilateral thalamic gray matter volume in first-episode schizophrenia with prominent hallucinatory symptoms: A volumetric MRI study. *Scientific Reports, 5*, 14505. https://doi.org/10.1038/srep14505

Hurlemann, R., Jessen, F., Wagner, M., Frommann, I., Ruhrmann, S., Brockhaus, A., Picker, H., Scheef, L., Block, W., Schild, H. H., Moller-Hartmann, W., Krug, B., Falkai, P., Klosterkotter, J., & Maier, W. (2008). Interrelated neuropsychological and anatomical evidence of hippocampal pathology in the at-risk mental state. *Psychological Medicine, 38*(6), 843–851. https://doi.org/10.1017/S0033291708003279

Huuki-Myers, L. A., Spangler, A., Eagles, N. J., Montgomery, K. D., Kwon, S. H., Guo, B., Grant-Peters, M., Divecha, H. R., Tippani, M., Sriworarat, C., Nguyen, A. B., Ravichandran, P., Tran, M. N., Seyedian, A., PsychENCODE Consortiumdagger, Hyde, T. M., Kleinman, J. E., Battle, A., Page, S. C., . . . PsychENCODE Consortium. (2024). A data-driven single-cell and spatial transcriptomic map of the human prefrontal cortex. *Science (New York, N.Y.), 384*(6698), eadh1938. https://doi.org/10.1126/science.adh1938

IBM. (2023). *Aspera Connect.* Retrieved 2019, from http://downloads.asperasoft.com/connect2/

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., . . . Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the

Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics, 99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Iwata, N., Suzuki, T., Ikeda, M., Kitajima, T., Yamanouchi, Y., Inada, T., & Ozaki, N. (2004). No association with the neuregulin 1 haplotype to Japanese schizophrenia. *Molecular Psychiatry, 9*(2), 126–127. https://doi.org/10.1038/sj.mp.4001456

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., . . . Weinberger, D. R. (2018). Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nature Neuroscience, 21*(8), 1117–1125. https://doi.org/10.1038/s41593-018-0197-y

Jaffe, A. E., Hoeppner, 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. *Nature Neuroscience, 23*(4), 510–519. https://doi.org/10.1038/s41593-020-0604-z

Janak, P. H., & Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature, 517*(7534), 284–292. https://doi.org/10.1038/nature14188

Jauhar, S., Johnstone, M., & McKenna, P. J. (2022). Schizophrenia. *Lancet (London, England), 399*(10323), 473–486. https://doi.org/10.1016/S0140-6736(21)01730-X

Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research, 42*(22), 13534–13544. https://doi.org/10.1093/nar/gku1206

Kadakia, A., Catillon, M., Fan, Q., Williams, G. R., Marden, J. R., Anderson, A., Kirson, N., & Dembek, C. (2022). The Economic Burden of Schizophrenia in the United States. *The Journal of Clinical Psychiatry, 83*(6), 22m14458. doi: 10.4088/JCP.22m14458. https://doi.org/10.4088/JCP.22m14458

Kanwal, A., Pardo, J. V., & Naz, S. (2022). RGS3 and IL1RAPL1 missense variants implicate defective neurotransmission in early-onset inherited schizophrenias. *Journal of Psychiatry & Neuroscience : JPN, 47*(6), E379–E390. https://doi.org/10.1503/jpn.220070

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7 [doi]

Kerns, J. G., Cohen, J. D., MacDonald, A. W. 3., Johnson, M. K., Stenger, V. A., Aizenstein, H., & Carter, C. S. (2005). Decreased conflict- and error-related activity in the anterior cingulate cortex in subjects with schizophrenia. *The American Journal of Psychiatry, 162*(10), 1833–1839. https://doi.org/10.1176/appi.ajp.162.10.1833

Khan, F. F., Melton, P. E., McCarthy, N. S., Morar, B., Blangero, J., Moses, E. K., & Jablensky, A. (2018). Whole genome sequencing of 91 multiplex schizophrenia families reveals increased burden of rare, exonic copy number variation in schizophrenia probands and genetic heterogeneity. *Schizophrenia Research, 197*, 337–345. https://doi.org/10.1016/j.schres.2018.02.034

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods, 12*(4), 357–360. https://doi.org/10.1038/nmeth.3317 [doi]

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology, 37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4 [doi]

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*(4), R36–r36. https://doi.org/10.1186/gb-2013-14-4-r36 [doi]

Koenig, Z., Yohannes, M. T., Nkambule, L. L., Zhao, X., Goodrich, J. K., Kim, H. A., Wilson, M. W., Tiao, G., Hao, S. P., Sahakian, N., Chao, K. R., Walker, M. A., Lyu, Y., gnomAD Project Consortium, Rehm, H., Neale, B. M., Talkowski, M. E., Daly, M. J., Brand, H., . . . Martin, A. R. (2024). A harmonized public resource of deeply sequenced diverse human genomes. *Genome Research,* https://doi.org/10.1101/gr.278378.123

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research, 44*(W1), 90. https://doi.org/10.1093/nar/gkw377

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols, 4*(7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Kummeling, J., Stremmelaar, D. E., Raun, N., Reijnders, M. R. F., Willemsen, M. H., Ruiterkamp-Versteeg, M., Schepens, M., Man, C. C. O., Gilissen, C., Cho, M. T., McWalter, K., Sinnema, M., Wheless, J. W., Simon, M. E. H., Genetti, C. A., Casey, A. M., Terhal, P. A., van der Smagt, J. J., van Gassen, K. L. I., . . . Kleefstra, T. (2020). Characterization of SETD1A haploinsufficiency in humans and Drosophila defines a novel neurodevelopmental syndrome. *Molecular Psychiatry,* https://doi.org/10.1038/s41380-020-0725-5 [doi]

Lander, E. S. (1996). The new genomics: global views of biology. *Science (New York, N.Y.), 274*(5287), 536–539. https://doi.org/10.1126/science.274.5287.536

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357–359. https://doi.org/10.1038/nmeth.1923 [doi]

Lariviere, S., Paquola, C., Park, B., Royer, J., Wang, Y., Benkarim, O., Vos de Wael, R., Valk, S. L., Thomopoulos, S. I., Kirschner, M., Lewis, L. B., Evans, A. C., Sisodiya, S. M., McDonald, C. R., Thompson, P. M., & Bernhardt, B. C. (2021). The ENIGMA

Toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nature Methods, 18*(7), 698–700. https://doi.org/10.1038/s41592-021-01186-4 [doi]

Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., & Ritchie, M. E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research, 5*, 10.12688/f1000research.9005.3. eCollection 2016. https://doi.org/ISCB Comm J-1408 [pii]

Le Berre, A., Kamagata, K., Otsuka, Y., Andica, C., Hatano, T., Saccenti, L., Ogawa, T., Takeshige-Amano, H., Wada, A., Suzuki, M., Hagiwara, A., Irie, R., Hori, M., Oyama, G., Shimo, Y., Umemura, A., Hattori, N., & Aoki, S. (2019). Convolutional neural network-based segmentation can help in assessing the substantia nigra in neuromelanin MRI. *Neuroradiology, 61*(12), 1387–1395. https://doi.org/10.1007/s00234-019-02279-w

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research, 39*(Database issue), 19. https://doi.org/10.1093/nar/gkq1019 [doi]

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., . . . Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285–291. https://doi.org/10.1038/nature19057 [doi]

Lencz, T., Morgan, T. V., Athanasiou, M., Dain, B., Reed, C. R., Kane, J. M., Kucherlapati, R., & Malhotra, A. K. (2007). Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Molecular Psychiatry, 12*(6), 572–580. https://doi.org/10.1038/sj.mp.4001983

Lewis, C. M., Levinson, D. F., Wise, L. H., DeLisi, L. E., Straub, R. E., Hovatta, I., Williams, N. M., Schwab, S. G., Pulver, A. E., Faraone, S. V., Brzustowicz, L. M., Kaufmann, C. A., Garver, D. L., Gurling, H. M. D., Lindholm, E., Coon, H., Moises, H. W., Byerley, W., Shaw, S. H., . . . Helgason, T. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *American Journal of Human Genetics, 73*(1), 34–48. https://doi.org/10.1086/376549

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England), 30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656 [doi]

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research, 47*(8), e47. https://doi.org/10.1093/nar/gkz114 [doi]

Liao, Y., & Shi, W. (2020). Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. *NAR Genomics and Bioinformatics, 2*(3)https://doi.org/10.1093/nargab/lqaa068

Lin, K., Hu, T., Hsu, S., Tsai, H., & Cheng, M. (2023). Identification of rare missense mutations in the glutamate ionotropic receptor AMPA type subunit genes in

schizophrenia. *Psychiatric Genetics, 33*(1), 20–25. https://doi.org/10.1097/YPG.0000000000000328

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation, 37*(3), 235–241. https://doi.org/10.1002/humu.22932

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550–8. https://doi.org/s13059-014-0550-8 [pii]

Luvsannyam, E., Jain, M. S., Pormento, M. K. L., Siddiqui, H., Balagtas, A. R. A., Emuze, B. O., & Poprawski, T. (2022). Neurobiology of Schizophrenia: A Comprehensive Review. *Cureus, 14*(4), e23959. https://doi.org/10.7759/cureus.23959 [doi]

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fajarado, K. V. F., Maile, M. S., Ripke, S., Agartz, I., Albus, M., . . . CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics, 49*(1), 27–35. https://doi.org/10.1038/ng.3725

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal, 17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., Catallini, J. L. 2., Tran, M. N., Besich, Z., Tippani, M., Chew, J., Yin, Y., Kleinman, J. E., Hyde, T. M., Rao, N., Hicks, S. C., Martinowich, K., & Jaffe, A. E. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience, 24*(3), 425–436. https://doi.org/10.1038/s41593-020-00787-0 [doi]

McClellan, J. M., Susser, E., & King, M. (2007). Schizophrenia: a common disease caused by multiple rare alleles. *The British Journal of Psychiatry : The Journal of Mental Science, 190*, 194–199. https://doi.org/10.1192/bjp.bp.106.025585

McCutcheon, R. A., Keefe, R. S. E., & McGuire, P. K. (2023). Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. *Molecular Psychiatry, 28*(5), 1902–1918. https://doi.org/10.1038/s41380-023-01949-9

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology, 17*(1), 122–4. https://doi.org/10.1186/s13059-016-0974-4 [doi]

Meisner, O. C., Nair, A., & Chang, S. W. C. (2022). Amygdala connectivity and implications for social cognition and disorders. *Handbook of Clinical Neurology, 187*, 381–403. https://doi.org/10.1016/B978-0-12-823493-8.00017-1

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segre, A. V., Djebali, S., Niarchou, A., GTEx Consortium, Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., . . . Guigo, R. (2015). Human genomics. The human transcriptome

across tissues and individuals. *Science (New York, N.Y.), 348*(6235), 660–665. https://doi.org/10.1126/science.aaa0355

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L. P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research, 49*(D1), D394–D403. https://doi.org/10.1093/nar/gkaa1106 [doi]

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. D. (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols, 14*(3), 703–721. https://doi.org/10.1038/s41596-019-0128-8 [doi]

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. *Human Molecular Genetics, 9*(9), 1415–1423. https://doi.org/10.1093/hmg/9.9.1415

Mitchell, K. J., & Porteous, D. J. (2009). GWAS for psychiatric disease: is the framework built on a solid foundation? *Molecular Psycry, 14*(8), 740–741. https://doi.org/10.1038/mp.2009.17

Moberget, T., Doan, N. T., Alnaes, D., Kaufmann, T., Cordova-Palomera, A., Lagerberg, T. V., Diedrichsen, J., Schwarz, E., Zink, M., Eisenacher, S., Kirsch, P., Jonsson, E. G., Fatouros-Bergman, H., Flyckt, L., KaSP, Pergola, G., Quarto, T., Bertolino, A., Barch, D., . . . Westlye, L. T. (2018). Cerebellar volume and cerebellocerebral structural

covariance in schizophrenia: a multisite mega-analysis of 983 patients and 1349 healthy controls. *Molecular Psychiatry, 23*(6), 1512–1520. https://doi.org/10.1038/mp.2017.106 [doi]

Moberget, T., & Ivry, R. B. (2019). Prediction, Psychosis, and the Cerebellum. *Biological Psychiatry.Cognitive Neuroscience and Neuroimaging, 4*(9), 820–831. https://doi.org/10.1016/j.bpsc.2019.06.001

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods, 5*(7), 621–628. https://doi.org/10.1038/nmeth.1226 [doi]

Munoz-Estrada, J., Lora-Castellanos, A., Meza, I., Alarcon Elizalde, S., & Benitez-King, G. (2018). Primary cilia formation is diminished in schizophrenia and bipolar disorder: A possible marker for these psychiatric diseases. *Schizophrenia Research, 195*, 412–420. https://doi.org/10.1016/j.schres.2017.08.055

Myles-Worsley, M., Tiobech, J., Browning, S. R., Korn, J., Goodman, S., Gentile, K., Melhem, N., Byerley, W., Faraone, S. V., & Middleton, F. A. (2013). Deletion at the SLC1A1 glutamate transporter gene co-segregates with schizophrenia and bipolar schizoaffective disorder in a 5-generation family. *American Journal of Medical Genetics.Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics, 162B*(2), 87–95. https://doi.org/10.1002/ajmg.b.32125

Nakamura, T., & Takata, A. (2023). The molecular pathology of schizophrenia: an overview of existing knowledge and new directions for future research. *Molecular Psychiatry, 28*(5), 1868–1889. https://doi.org/10.1038/s41380-023-02005-2

NCBI. (2019). sra-toolkit [computer software]. NCBI:

NCBI. (2022, Jan). *dbGaP File Selector.* Retrieved Jun 2018, from https://www.ncbi.nlm.nih.gov/Traces/study/?acc=phs000473&ff=1&o=consents%3Aa&s=d1593410,d1593406,d1593399,d1593403,d1593412,d1593396,d1593404,d1593398,d1593395,d1593397,d1593401

NCBI. (n.d., n.d.). *SRA Run Selector.* https://www.ncbi.nlm.nih.gov/Traces/study/

Neale, B. M., & Sklar, P. (2015). Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but also suggests new directions for molecular interrogation. *Current Opinion in Neurobiology, 30*, 131–138. https://doi.org/10.1016/j.conb.2014.12.001 [doi]

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., . . . Lewis, C. M. (2009). Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Molecular Psychiatry, 14*(8), 774–785. https://doi.org/10.1038/mp.2008.135

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research, 11*(5), 863–874. https://doi.org/10.1101/gr.176601 [doi]

Nguyen, H. T., Bryois, J., Kim, A., Dobbyn, A., Huckins, L. M., Munoz-Manchado, A. B., Ruderfer, D. M., Genovese, G., Fromer, M., Xu, X., Pinto, D., Linnarsson, S., Verhage, M., Smit, A. B., Hjerling-Leffler, J., Buxbaum, J. D., Hultman, C., Sklar, P., Purcell, S. M., . . . Stahl, E. A. (2017). Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Medicine, 9*(1), 114–y. https://doi.org/10.1186/s13073-017-0497-y

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., Baker, J., Ferrer, J., Raies, A., Razuvayevskaya, O., Faulconbridge, A., Petsalaki, E., Mutowo, P., Machlitt-Northen, S., Peat, G., . . . McDonagh, E. M. (2021). Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research, 49*(D1), D1302–D1310. https://doi.org/10.1093/nar/gkaa1027 [doi]

Ohtani, T., Bouix, S., Hosokawa, T., Saito, Y., Eckbo, R., Ballinger, T., Rausch, A., Melonakos, E., & Kubicki, M. (2014). Abnormalities in white matter connections between orbitofrontal cortex and anterior cingulate cortex and their associations with negative symptoms in schizophrenia: a DTI study. *Schizophrenia Research, 157*(1-3), 190–197. https://doi.org/10.1016/j.schres.2014.05.016

Okada, N., Fukunaga, M., Yamashita, F., Koshiyama, D., Yamamori, H., Ohi, K., Yasuda, Y., Fujimoto, M., Watanabe, Y., Yahata, N., Nemoto, K., Hibar, D. P., van Erp, T. G. M., Fujino, H., Isobe, M., Isomura, S., Natsubori, T., Narita, H., Hashimoto, N., . . . Hashimoto, R. (2016). Abnormal asymmetries in subcortical brain volume in

schizophrenia. *Molecular Psychiatry, 21*(10), 1460–1466. https://doi.org/10.1038/mp.2015.209 [doi]

Olabi, B., Ellison-Wright, I., McIntosh, A. M., Wood, S. J., Bullmore, E., & Lawrie, S. M. (2011). Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. *Biological Psychiatry, 70*(1), 88–96. https://doi.org/10.1016/j.biopsych.2011.01.032 [doi]

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research, 44*(D1), 733. https://doi.org/10.1093/nar/gkv1189 [doi]

Oliva, M., Munoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gewirtz, A. D. H., Cotter, D. J., Parsana, P., Kasela, S., Balliu, B., Vinuela, A., Castel, S. E., Mohammadi, P., Aguet, F., Zou, Y., Khramtsova, E. A., Skol, A. D., Garrido-Martin, D., Reverter, F., Brown, A., . . . Stranger, B. E. (2020). The impact of sex on gene expression across human tissues. *Science (New York, N.Y.), 369*(6509), 10.1126/science.aba3066. https://doi.org/eaba3066 [pii]

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology, 11*(12), 220–220. Epub 2010 Dec 22. https://doi.org/10.1186/gb-2010-11-12-220 [doi]

Ott, J. (2004). Association of genetic loci: replication or not, that is the question. *Neurology, 63*(6), 955–958. https://doi.org/10.1212/wnl.63.6.955

Ott, J., Wang, J., & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews.Genetics, 16*(5), 275–284. https://doi.org/10.1038/nrg3908

Owen, M. J., Legge, S. E., Rees, E., Walters, J. T. R., & O'Donovan, M. C. (2023). Genomic findings in schizophrenia and their implications. *Molecular Psychiatry, 28*(9), 3638–3647. https://doi.org/10.1038/s41380-023-02293-8

Paquin, V., Lapierre, M., Veru, F., & King, S. (2021). Early Environmental Upheaval and the Risk for Schizophrenia. *Annual Review of Clinical Psychology, 17*, 285–311. https://doi.org/10.1146/annurev-clinpsy-081219-103805

Pardo, B., Spangler, A., Weber, L. M., Page, S. C., Hicks, S. C., Jaffe, A. E., Martinowich, K., Maynard, K. R., & Collado-Torres, L. (2022). spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genomics, 23*(1), 434–w. https://doi.org/10.1186/s12864-022-08601-w

Park, I. H., Park, H., Chun, J., Kim, E. Y., & Kim, J. (2008). Dysfunctional modulation of emotional interference in the medial prefrontal cortex in patients with schizophrenia. *Neuroscience Letters, 440*(2), 119–124. https://doi.org/10.1016/j.neulet.2008.05.094

Park, S. M., Jang, H. J., & Lee, J. H. (2019). Roles of Primary Cilia in the Developing Brain. *Frontiers in Cellular Neuroscience, 13*, 218. https://doi.org/10.3389/fncel.2019.00218

Parker, S. E., Bellingham, M. C., & Woodruff, T. M. (2022). Complement drives circuit modulation in the adult brain. *Progress in Neurobiology, 214*, 102282. https://doi.org/S0301-0082(22)00068-5 [pii]

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods, 14*(4), 417–419. https://doi.org/10.1038/nmeth.4197 [doi]

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology, 32*(5), 462–464. https://doi.org/10.1038/nbt.2862 [doi]

Pavlopoulos, G. A., Malliarakis, D., Papanikolaou, N., Theodosiou, T., Enright, A. J., & Iliopoulos, I. (2015). Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *GigaScience, 4*, 38–2. eCollection 2015. https://doi.org/10.1186/s13742-015-0077-2 [doi]

Pertea, M. (2012). The human transcriptome: an unfinished story. *Genes, 3*(3), 344–360. https://doi.org/10.3390/genes3030344 [doi]

Pillinger, T., Beck, K., Gobjila, C., Donocik, J. G., Jauhar, S., & Howes, O. D. (2017). Impaired Glucose Homeostasis in First-Episode Schizophrenia: A Systematic Review and Meta-analysis. *JAMA Psychiatry, 74*(3), 261–269. https://doi.org/10.1001/jamapsychiatry.2016.3803

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods, 14*(7), 687–690. https://doi.org/10.1038/nmeth.4324 [doi]

Pinero, J., Ramirez-Anguita, J. M., Sauch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research, 48*(D1), D845–D855. https://doi.org/10.1093/nar/gkz1021

Porcu, E., Sadler, M. C., Lepik, K., Auwerx, C., Wood, A. R., Weihs, A., Sleiman, M. S. B., Ribeiro, D. M., Bandinelli, S., Tanaka, T., Nauck, M., Volker, U., Delaneau, O., Metspalu, A., Teumer, A., Frayling, T., Santoni, F. A., Reymond, A., & Kutalik, Z. (2021). Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications, 12*(1), 5647–y. https://doi.org/10.1038/s41467-021-25805-y

Pruessner, M., Cullen, A. E., Aas, M., & Walker, E. F. (2017). The neural diathesis-stress model of schizophrenia revisited: An update on recent findings considering illness stage and neurobiological and methodological complexities. *Neuroscience and Biobehavioral Reviews, 73*, 191–218. https://doi.org/10.1016/j.neubiorev.2016.12.013

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S. E., Kahler, A., Duncan, L., Stahl, E., Genovese, G., Fernandez, E., Collins, M. O., Komiyama, N. H., Choudhary, J. S., Magnusson, P. K., Banks, E., . . . Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature, 506*(7487), 185–190. https://doi.org/10.1038/nature12975 [doi]

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England), 26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033 [doi]

Radulescu, E., Jaffe, A. E., Straub, R. E., Chen, Q., Shin, J. H., Hyde, T. M., Kleinman, J. E., & Weinberger, D. R. (2020). Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Molecular Psychiatry, 25*(4), 791–804. https://doi.org/10.1038/s41380-018-0304-1

Ragland, J. D., Layher, E., Hannula, D. E., Niendam, T. A., Lesh, T. A., Solomon, M., Carter, C. S., & Ranganath, C. (2016). Impact of schizophrenia on anterior and posterior hippocampus during memory for complex scenes. *NeuroImage.Clinical, 13*, 82–88. https://doi.org/10.1016/j.nicl.2016.11.017

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., & Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology, 14*(9), R95–r95. https://doi.org/gb-2013-14-9-r95 [pii]

Reble, E., Dineen, A., & Barr, C. L. (2018). The contribution of alternative splicing to genetic risk for psychiatric disorders. *Genes, Brain, and Behavior, 17*(3), e12430. https://doi.org/10.1111/gbb.12430 [doi]

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research, 47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Ressler, K. J. (2010). Amygdala activity, fear, and anxiety: modulation by stress. *Biological Psychiatry, 67*(12), 1117–1119. https://doi.org/10.1016/j.biopsych.2010.04.027

Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., & Huber, W. (2013). Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences of the United States of America, 110*(38), 15377–15382. https://doi.org/10.1073/pnas.1307202110 [doi]

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 Research, 43*(7), e47. https://doi.org/10.1093/nar/gkv007 [doi]

Ruggeri, A., Nerland, S., Morch-Johnsen, L., Jorgensen, K. N., Barth, C., Wortinger, L. A., Andreou, D., Andreassen, O. A., & Agartz, I. (2024). Hypothalamic Subunit Volumes in Schizophrenia and Bipolar Spectrum Disorders. *Schizophrenia Bulletin, 50*(3), 533–544. https://doi.org/10.1093/schbul/sbad176

Ruzicka, W. B., Mohammadi, S., Fullard, J. F., Davila-Velderrain, J., Subburaju, S., Tso, D. R., Hourihan, M., Jiang, S., Lee, H., Bendl, J., PsychENCODE Consortium section sign, Voloudakis, G., Haroutunian, V., Hoffman, G. E., Roussos, P., Kellis, M., & PsychENCODE Consortium. (2024). Single-cell multi-cohort dissection of the schizophrenia transcriptome. *Science (New York, N.Y.), 384*(6698), eadg5136. https://doi.org/10.1126/science.adg5136

Ryan, N. M., Lihm, J., Kramer, M., McCarthy, S., Morris, S. W., Arnau-Soler, A., Davies, G., Duff, B., Ghiban, E., Hayward, C., Deary, I. J., Blackwood, D. H. R., Lawrie, S. M., McIntosh, A. M., Evans, K. L., Porteous, D. J., McCombie, W. R., & Thomson, P. A. (2018). DNA sequence-level analyses reveal potential phenotypic modifiers in a large family with psychiatric disorders. *Molecular Psychiatry, 23*(12), 2254–2265. https://doi.org/10.1038/s41380-018-0087-4

Sah, P., Faber, E. S. L., Lopez De Armentia, M., & Power, J. (2003). The amygdaloid complex: anatomy and physiology. *Physiological Reviews, 83*(3), 803–834. https://doi.org/10.1152/physrev.00002.2003

Saha, A., & Battle, A. (2018). False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Research, 7*, 1860. https://doi.org/10.12688/f1000research.17145.2

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., GTEx Consortium, Engelhardt, B. E., & Battle, A. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research, 27*(11), 1843–1858. https://doi.org/10.1101/gr.216721.116 [doi]

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., Bani Asadi, N., Gerstein, M. B., Wong, W. H., Snyder, M. P., Schadt, E., & Lam, H. Y. K. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications, 8*(1), 59–4. https://doi.org/10.1038/s41467-017-00050-4 [doi]

Samson, G. D., Lahti, A. C., & Kraguljac, N. V. (2022). The neural substrates of neurological soft signs in schizophrenia: a systematic review. *Schizophrenia (Heidelberg, Germany), 8*(1), 42–9. https://doi.org/10.1038/s41537-022-00245-9 [doi]

Sartorius, L. J., Nagappan, G., Lipska, B. K., Lu, B., Sei, Y., Ren-Patterson, R., Li, Z., Weinberger, D. R., & Harrison, P. J. (2006). Alternative splicing of human metabotropic glutamate receptor 3. *Journal of Neurochemistry, 96*(4), 1139–1148. https://doi.org/10.1111/j.1471-4159.2005.03609.x

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics, 43*(10), 969–976. https://doi.org/10.1038/ng.940

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421–427. https://doi.org/10.1038/nature13595 [doi]

Schmid, M. W., & Grossniklaus, U. (2015). Rcount: simple and flexible RNA-Seq read counting. *Bioinformatics (Oxford, England), 31*(3), 436–437. https://doi.org/10.1093/bioinformatics/btu680 [doi]

Schulz, J., Zimmermann, J., Sorg, C., Menegaux, A., & Brandl, F. (2022). Magnetic resonance imaging of the dopamine system in schizophrenia - A scoping review. *Frontiers in Psychiatry, 13*, 925476. https://doi.org/10.3389/fpsyt.2022.925476

Sekiguchi, M., Sobue, A., Kushima, I., Wang, C., Arioka, Y., Kato, H., Kodama, A., Kubo, H., Ito, N., Sawahata, M., Hada, K., Ikeda, R., Shinno, M., Mizukoshi, C., Tsujimura,

K., Yoshimi, A., Ishizuka, K., Takasaki, Y., Kimura, H., . . . Ozaki, N. (2020). ARHGAP10, which encodes Rho GTPase-activating protein 10, is a novel gene for schizophrenia risk. *Translational Psychiatry, 10*(1), 247–z. https://doi.org/10.1038/s41398-020-00917-z [doi]

SEQC/MAQC-III Consortium. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology, 32*(9), 903–914. https://doi.org/10.1038/nbt.2957

Sequeira, A., Shen, K., Gottlieb, A., & Limon, A. (2019). Human brain transcriptome analysis finds region- and subject-specific expression signatures of GABA(A)R subunits. *Communications Biology, 2*, 153–7. eCollection 2019. https://doi.org/10.1038/s42003-019-0413-7

Shen, K., Zeppillo, T., & Limon, A. (2020). Regional transcriptome analysis of AMPA and GABA(A) receptor subunit expression generates E/I signatures of the human brain. *Scientific Reports, 10*(1), 11352–1. https://doi.org/10.1038/s41598-020-68165-1

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research, 50*(W1), W216–W221. https://doi.org/10.1093/nar/gkac194

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, (., The AMP-AD Consortium, Omberg, L.,

Peters, M. A., Logsdon, B. A., De Jager, P. L., . . . Mangravite, L. M. (2020). Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Scientific Data,* *7*(1), 340–8. https://doi.org/10.1038/s41597-020-00642-8 [doi]

Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., Pietilainen, O., Gerety, S. S., Ayub, M., Blyth, M., Cole, T., Collier, D., Coomber, E. L., Craddock, N., Daly, M. J., . . . Barrett, J. C. (2016). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nature Neuroscience,* *19*(4), 571–577. https://doi.org/10.1038/nn.4267 [doi]

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., . . . Daly, M. J. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature,* *604*(7906), 509–516. https://doi.org/10.1038/s41586-022-04556-w [doi]

Singh, T., Walters, J. T. R., Johnstone, M., Curtis, D., Suvisaari, J., Torniainen, M., Rees, E., Iyegbe, C., Blackwood, D., McIntosh, A. M., Kirov, G., Geschwind, D., Murray, R. M., Di Forti, M., Bramon, E., Gandal, M., Hultman, C. M., Sklar, P., INTERVAL Study, . . . Barrett, J. C. (2017). The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nature Genetics,* *49*(8), 1167–1173. https://doi.org/10.1038/ng.3903 [doi]

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., Munoz-Manchado, A. B., O'Donovan, M. C., Owen, M. J., Pardinas, 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. *Nature Genetics, 50*(6), 825–833. https://doi.org/10.1038/s41588-018-0129-5

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. *Frontiers in Neuroscience, 10*, 16. https://doi.org/10.3389/fnins.2016.00016

Slay, O. (2021). PhD-codebase [source code] [computer software]. GitHub:

Smucny, J., Dienel, S. J., Lewis, D. A., & Carter, C. S. (2022). Mechanisms underlying dorsolateral prefrontal cortex contributions to cognitive dysfunction in schizophrenia. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology, 47*(1), 292–308. https://doi.org/10.1038/s41386-021-01089-0

Solmi, M., Seitidis, G., Mavridis, D., Correll, C. U., Dragioti, E., Guimond, S., Tuominen, L., Dargel, A., Carvalho, A. F., Fornaro, M., Maes, M., Monaco, F., Song, M., Il Shin, J., & Cortese, S. (2023). Incidence, prevalence, and global burden of schizophrenia - data, with critical appraisal, from the Global Burden of Disease (GBD) 2019. *Molecular Psychiatry, 28*(12), 5319–5327. https://doi.org/10.1038/s41380-023-02138-4

SRA Toolkit Development Team. (n.d., n.d.). *NCBI SRA Toolkit.* https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Srinivasan, S., Kalinava, N., Aldana, R., Li, Z., van Hagen, S., Rodenburg, S. Y. A., Wind-Rotolo, M., Qian, X., Sasson, A. S., Tang, H., & Kirov, S. (2021). Misannotated Multi-Nucleotide Variants in Public Cancer Genomics Datasets Lead to Inaccurate Mutation Calls with Significant Implications. *Cancer Research, 81*(2), 282–288. https://doi.org/10.1158/0008-5472.CAN-20-2151

St Clair, D., Blackwood, D., Muir, W., Carothers, A., Walker, M., Spowart, G., Gosden, C., & Evans, H. J. (1990). Association within a family of a balanced autosomal translocation with major mental illness. *Lancet (London, England), 336*(8706), 13–16. https://doi.org/10.1016/0140-6736(90)91520-k

Steen, R. G., Hamer, R. M., & Lieberman, J. A. (2005). Measurement of brain metabolites by 1H magnetic resonance spectroscopy in patients with schizophrenia: a systematic review and meta-analysis. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology, 30*(11), 1949–1962. https://doi.org/10.1038/sj.npp.1300850

Steiner, J., Brisch, R., Schiltz, K., Dobrowolny, H., Mawrin, C., Krzyzanowska, M., Bernstein, H. G., Jankowski, Z., Braun, K., Schmitt, A., Bogerts, B., & Gos, T. (2016). GABAergic system impairment in the hippocampus and superior temporal gyrus of patients with paranoid schizophrenia: A post-mortem study. *Schizophrenia Research, 177*(1-3), 10–17. https://doi.org/S0920-9964(16)30074-3 [pii]

Stilo, S. A., & Murray, R. M. (2019). Non-Genetic Factors in Schizophrenia. *Current Psychiatry Reports, 21*(10), 100–3. https://doi.org/10.1007/s11920-019-1091-3

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry, 60*(12), 1187–1192. https://doi.org/10.1001/archpsyc.60.12.1187

Sun, B. B., Kurki, M. I., Foley, C. N., Mechakra, A., Chen, C. Y., Marshall, E., Wilk, J. B., Biogen Biobank Team, Chahine, M., Chevalier, P., Christe, G., FinnGen, Palotie, A., Daly, M. J., & Runz, H. (2022). Genetic associations of protein-coding variants in human disease. *Nature, 603*(7899), 95–102. https://doi.org/10.1038/s41586-022-04394-w [doi]

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research, 49*(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074 [doi]

Tandon, R., Gaebel, W., Barch, D. M., Bustillo, J., Gur, R. E., Heckers, S., Malaspina, D., Owen, M. J., Schultz, S., Tsuang, M., Van Os, J., & Carpenter, W. (2013). Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research, 150*(1), 3–10. https://doi.org/10.1016/j.schres.2013.05.028

Tandon, R., Nasrallah, H., Akbarian, S., Carpenter, W. T. J., DeLisi, L. E., Gaebel, W., Green, M. F., Gur, R. E., Heckers, S., Kane, J. M., Malaspina, D., Meyer-Lindenberg, A., Murray, R., Owen, M., Smoller, J. W., Yassin, W., & Keshavan, M. (2024). The

schizophrenia syndrome, circa 2024: What we know and how that informs its nature. *Schizophrenia Research, 264*, 1–28. https://doi.org/10.1016/j.schres.2023.11.015

Teare, M. D., & Santibanez Koref, M. F. (2014). Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. *Briefings in Functional Genomics, 13*(5), 378–383. https://doi.org/10.1093/bfgp/elu024

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., . . . NHLBI Exome Sequencing Project. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.), 337*(6090), 64–69. https://doi.org/10.1126/science.1219240

Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., Buckner, R. L., Buitelaar, J. K., Bulayeva, K. B., Cannon, D. M., Cohen, R. A., Conrod, P. J., Dale, A. M., Deary, I. J., Dennis, E. L., de Reus, M. A., Desrivieres, S., Dima, D., Donohoe, G., . . . ENIGMA Consortium. (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage, 145*(Pt B), 389–408. https://doi.org/S1053-8119(15)01081-2 [pii]

Thompson, P. M., Jahanshad, N., Ching, C. R. K., Salminen, L. E., Thomopoulos, S. I., Bright, J., Baune, B. T., Bertolin, S., Bralten, J., Bruin, W. B., Bulow, R., Chen, J., Chye, Y., Dannlowski, U., de Kovel, C. G. F., Donohoe, G., Eyler, L. T., Faraone, S. V., Favre, P., . . . ENIGMA Consortium. (2020). ENIGMA and global neuroscience:

A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Translational Psychiatry, 10*(1), 100–1. https://doi.org/10.1038/s41398-020-0705-1

Torre, D., Lachmann, A., & Ma'ayan, A. (2018). BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. *Cell Systems, 7*(5), 556–561.e3. https://doi.org/10.1016/j.cels.2018.10.007

Trubetskoy, V., Pardinas, 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., . . . Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature, 604*(7906), 502–508. https://doi.org/10.1038/s41586-022-04434-5 [doi]

UK10K Consortium, Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., . . . Soranzo, N. (2015). The UK10K project identifies rare variants in health and disease. *Nature, 526*(7571), 82–90. https://doi.org/10.1038/nature14962 [doi]

Vallee, A. (2022). Neuroinflammation in Schizophrenia: The Key Role of the WNT/beta-Catenin Pathway. *International Journal of Molecular Sciences, 23*(5), 2810. doi: 10.3390/ijms23052810. https://doi.org/10.3390/ijms23052810

Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I., Patro, R., & Robinson, M. D. (2019). RNA Sequencing Data: Hitchhiker's Guide to

Expression Analysis. *Annual Review of Biomedical Data Science, 2*(1), 139–173. https://doi.org/10.1146/annurev-biodatasci-072018-021255

van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., Agartz, I., Westlye, L. T., Haukvik, U. K., Dale, A. M., Melle, I., Hartberg, C. B., Gruber, O., Kraemer, B., Zilles, D., Donohoe, G., Kelly, S., McDonald, C., Morris, D. W., . . . Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry, 21*(4), 547–553. https://doi.org/10.1038/mp.2015.63 [doi]

van Erp, T. G. M., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., Pearlson, G. D., Yao, N., Fukunaga, M., Hashimoto, R., Okada, N., Yamamori, H., Bustillo, J. R., Clark, V. P., Agartz, I., Mueller, B. A., Cahn, W., de Zwarte, S. M. C., Hulshoff Pol, H. E., . . . Turner, J. A. (2018). Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. *Biological Psychiatry, 84*(9), 644–654. https://doi.org/S0006-3223(18)31517-8 [pii]

Varambally, S., Venkatasubramanian, G., & Gangadhar, B. N. (2012). Neurological soft signs in schizophrenia - The past, the present and the future. *Indian Journal of Psychiatry, 54*(1), 73–80. https://doi.org/10.4103/0019-5545.94653

Vien, T. N., Ta, M. C., Kimura, L. F., Onay, T., & DeCaen, P. G. (2023). Primary cilia TRP channel regulates hippocampal excitability. *Proceedings of the National Academy of Sciences of the United States of America, 120*(22), e2219686120. https://doi.org/10.1073/pnas.2219686120

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics, 90*(1), 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029 [doi]

Vita, A., Bressi, S., Perani, D., Invernizzi, G., Giobbio, G. M., Dieci, M., Garbarini, M., Del Sole, A., & Fazio, F. (1995). High-resolution SPECT study of regional cerebral blood flow in drug-free and drug-naive schizophrenic patients. *The American Journal of Psychiatry, 152*(6), 876–882. https://doi.org/10.1176/ajp.152.6.876

Wakeling, M. N., Laver, T. W., Colclough, K., Parish, A., Ellard, S., & Baple, E. L. (2019). Misannotation of multiple-nucleotide variants risks misdiagnosis. *Wellcome Open Research, 4*, 145. https://doi.org/10.12688/wellcomeopenres.15420.2 [doi]

Wallwork, R. S., Fortgang, R., Hashimoto, R., Weinberger, D. R., & Dickinson, D. (2012). Searching for a consensus five-factor model of the Positive and Negative Syndrome Scale for schizophrenia. *Schizophrenia Research, 137*(1-3), 246–250. https://doi.org/10.1016/j.schres.2012.01.031

Wang, A. M., Pradhan, S., Coughlin, J. M., Trivedi, A., DuBois, S. L., Crawford, J. L., Sedlak, T. W., Nucifora, F. C. J., Nestadt, G., Nucifora, L. G., Schretlen, D. J., Sawa, A., & Barker, P. B. (2019). Assessing Brain Metabolism With 7-T Proton Magnetic Resonance Spectroscopy in Patients With First-Episode Psychosis. *JAMA Psychiatry, 76*(3), 314–323. https://doi.org/10.1001/jamapsychiatry.2018.3637

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human

tissue transcriptomes. *Nature,* *456*(7221), 470–476. https://doi.org/10.1038/nature07509 [doi]

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164. https://doi.org/10.1093/nar/gkq603 [doi]

Wang, M. H., Cordell, H. J., & Van Steen, K. (2019). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology, 55*, 53–60. https://doi.org/10.1016/j.semcancer.2018.04.008

Wang, Q., Pierce-Hoffman, E., Cummings, B. B., Alfoldi, J., Francioli, L. C., Gauthier, L. D., Hill, A. J., O'Donnell-Luria, A. H., Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Karczewski, K. J., & MacArthur, D. G. (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nature Communications, 11*(1), 2539–5. https://doi.org/10.1038/s41467-019-12438-5 [doi]

Wang, S., Rhijn, J. V., Akkouh, I., Kogo, N., Maas, N., Bleeck, A., Ortiz, I. S., Lewerissa, E., Wu, K. M., Schoenmaker, C., Djurovic, S., van Bokhoven, H., Kleefstra, T., Nadif Kasri, N., & Schubert, D. (2022). Loss-of-function variants in the schizophrenia risk gene SETD1A alter neuronal network activity in human neurons through the cAMP/PKA pathway. *Cell Reports, 39*(5), 110790. https://doi.org/S2211-1247(22)00557-5 [pii]

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews.Genetics, 10*(1), 57–63. https://doi.org/10.1038/nrg2484 [doi]

Weinberger, D. R., & Laruelle, M. (2001). Neurochemical and neuropharmacological imaging in schizophrenia. *Neuropharmacology-the Fifth Generation of Progress,*

Wen, L., & Tang, F. (2022). Recent advances in single-cell sequencing technologies. *Precision Clinical Medicine, 5*(1), pbac002. https://doi.org/10.1093/pcmedi/pbac002

Whiffin, N., Roberts, A. M., Minikel, E., Zappala, Z., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K. J., Harrison, S. M., Thomson, K. L., Sage, H., Ing, A. Y., Barton, P. J. R., Funke, B., Cook, S. A., MacArthur, D. G., & Ware, J. S. (2019). Using High-Resolution Variant Frequencies Empowers Clinical Genome Interpretation and Enables Investigation of Genetic Architecture. *American Journal of Human Genetics, 104*(1), 187–190. https://doi.org/S0002-9297(18)30420-8 [pii]

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Ing, A. Y., Barton, P. J. R., Funke, B., Cook, S. A., MacArthur, D., & Ware, J. S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine : Official Journal of the American College of Medical Genetics, 19*(10), 1151–1158. https://doi.org/10.1038/gim.2017.26

Wilkinson, M. E., Charenton, C., & Nagai, K. (2020). RNA Splicing by the Spliceosome. *Annual Review of Biochemistry, 89*, 359–388. https://doi.org/10.1146/annurev-biochem-091719-064225 [doi]

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics, 17*, 103–2. https://doi.org/10.1186/s12859-016-0956-2 [doi]

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2017). Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics, 18*(1), 38–z. https://doi.org/10.1186/s12859-016-1457-z [doi]

Witthaus, H., Kaufmann, C., Bohner, G., Ozgurdal, S., Gudlowski, Y., Gallinat, J., Ruhrmann, S., Brune, M., Heinz, A., Klingebiel, R., & Juckel, G. (2009). Gray matter abnormalities in subjects at ultra-high risk for schizophrenia and first-episode schizophrenic patients compared to healthy controls. *Psychiatry Research, 173*(3), 163–169. https://doi.org/10.1016/j.pscychresns.2008.08.002

Witthaus, H., Mendes, U., Brune, M., Ozgurdal, S., Bohner, G., Gudlowski, Y., Kalus, P., Andreasen, N., Heinz, A., Klingebiel, R., & Juckel, G. (2010). Hippocampal subdivision and amygdalar volumes in patients in an at-risk mental state for schizophrenia. *Journal of Psychiatry & Neuroscience : JPN, 35*(1), 33–40. https://doi.org/10.1503/jpn.090013

Woo, J. J., Pouget, J. G., Zai, C. C., & Kennedy, J. L. (2020). The complement system in schizophrenia: where are we now and what's next? *Molecular Psychiatry, 25*(1), 114–130. https://doi.org/10.1038/s41380-019-0479-0

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S. A., Baekvad-

Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschon, H. N., Bybjerg-Grauholm, J., . . . Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics, 50*(5), 668–681. https://doi.org/10.1038/s41588-018-0090-3 [doi]

Wright, C., Calhoun, V. D., Ehrlich, S., Wang, L., Turner, J. A., & Bizzozero, N. I. (2015). Meta gene set enrichment analyses link miR-137-regulated pathways with schizophrenia risk. *Frontiers in Genetics, 6*, 147. https://doi.org/10.3389/fgene.2015.00147 [doi]

Wright, I. C., Rabe-Hesketh, S., Woodruff, P. W., David, A. S., Murray, R. M., & Bullmore, E. T. (2000). Meta-analysis of regional brain volumes in schizophrenia. *The American Journal of Psychiatry, 157*(1), 16–25. https://doi.org/10.1176/ajp.157.1.16

Wu, X. L., Yan, Q. J., & Zhu, F. (2022). Abnormal synaptic plasticity and impaired cognition in schizophrenia. *World Journal of Psychiatry, 12*(4), 541–557. https://doi.org/10.5498/wjp.v12.i4.541 [doi]

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols, 1*(3), e90. https://doi.org/10.1002/cpz1.90

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J. A., & Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics, 44*(12), 1365–1369. https://doi.org/10.1038/ng.2446

Yeo, G., Holste, D., Kreiman, G., & Burge, C. B. (2004). Variation in alternative splicing across human tissues. *Genome Biology, 5*(10), R74–r74. Epub 2004 Sep 13. https://doi.org/gb-2004-5-10-r74 [pii]

Yoon, S., Piguel, N. H., Khalatyan, N., Dionisio, L. E., Savas, J. N., & Penzes, P. (2021). Homer1 promotes dendritic spine growth through ankyrin-G and its loss reshapes the synaptic proteome. *Molecular Psychiatry, 26*(6), 1775–1789. https://doi.org/10.1038/s41380-020-00991-1

Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., & Linnarsson, S. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (New York, N.Y.), 347*(6226), 1138–1142. https://doi.org/10.1126/science.aaa1934

Zhang, D., Guelfi, S., Garcia-Ruiz, S., Costa, B., Reynolds, R. H., D'Sa, K., Liu, W., Courtin, T., Peterson, A., Jaffe, A. E., Hardy, J., Botia, J. A., Collado-Torres, L., & Ryten, M. (2020). Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances, 6*(24), eaay8299. doi: 10.1126/sciadv.aay8299. Print 2020 Jun. https://doi.org/10.1126/sciadv.aay8299

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Gorczak, K., Klamecka, K., Figlerowicz, M., & Siatkowski, I. (2015). The Impact of Normalization Methods on RNA-Seq Data Analysis. *BioMed Research International, 2015*, 621690. https://doi.org/10.1155/2015/621690 [doi]

8 Appendix

8.1 Hardy Scale classification

Definitions from dbGaP variable accession: phv00169092.v8.p2:

0 Ventilator case – on a ventilator immediately before death

1 Violent and fast death – deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 min.

2 Fast death of natural causes – sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hr (for example, sudden death from a myocardial infarction).

3 Intermediate death – death after a terminal phase of 1-24 hours (not classifiable as 2 or 4) – patients who were ill, but death was unexpected.

4 Slow death – death was after a long illness, with a terminal phase longer than 1 day (for example, cancer or chronic pulmonary disease); deaths are not unexpected.

8.2 Differential Expression Analysis Phenotype data – 100 samples

RNA-Seq data from 4 brain regions from 25 donors were used in the main Differential Expression Analysis in Section 3.2.1.4.

- Frontal Cortex (cortex)
- Cerebellar Hemisphere (cereb)
- Substantia nigra (sn)
- Hippocampus (hippo):

sample_id	subject_id	ctrl	Roi	sex	age	RIN	HardyScale
SRR1363810	2	0	cereb	f	40-49	6.8	1
SRR1317344	2	1	cortex	f	40-49	6.8	1
SRR1459655	2	0	sn	f	40-49	6.8	1
SRR1490524	2	1	hippo	f	40-49	8.5	1
SRR1340420	3	0	cereb	f	50-59	7.9	2
SRR1354446	3	1	cortex	f	50-59	7.9	2
SRR1420493	3	0	sn	f	50-59	7.9	2
SRR1340463	3	1	hippo	f	50-59	7.2	2
SRR1387072	7	0	cereb	f	60-69	7.1	3

sample_id	subject_id	ctrl	Roi	sex	age	RIN	HardyScale
SRR1437538	7	1	cortex	f	60-69	7.1	3
SRR1479779	7	0	sn	f	60-69	7.1	3
SRR1348339	7	1	hippo	f	60-69	6.9	3
SRR1478294	8	0	cereb	f	50-59	7.5	2
SRR1487685	8	1	cortex	f	50-59	7.5	2
SRR1335688	8	0	sn	f	50-59	7.5	2
SRR1489971	8	1	hippo	f	50-59	7.3	2
SRR600957	15	0	cereb	f	60-69	8.9	4
SRR602314	15	1	cortex	f	60-69	8.9	4
SRR603397	15	0	sn	f	60-69	8.9	4
SRR1362888	15	1	hippo	f	60-69	6.7	4
SRR608718	17	0	cereb	f	40-49	6.3	2
SRR602704	17	1	cortex	f	40-49	6.3	2
SRR604074	17	0	sn	f	40-49	6.3	2
SRR1473979	17	1	hippo	f	40-49	8.3	2
SRR661290	18	0	cereb	f	30-39	6.2	4
SRR661973	18	1	cortex	f	30-39	6.2	4
SRR660638	18	0	sn	f	30-39	6.2	4
SRR1380680	18	1	hippo	f	30-39	6.2	4
SRR1387875	22	0	cereb	f	40-49	7	1
SRR1328895	22	1	cortex	f	40-49	7	1
SRR1455171	22	0	sn	f	40-49	7	1
SRR1398411	22	1	hippo	f	40-49	8.6	1
SRR1394677	32	0	cereb	m	50-59	6.4	2
SRR1348360	32	1	cortex	m	50-59	6.4	2
SRR1410956	32	0	sn	m	50-59	6.4	2
SRR1349013	32	1	hippo	m	50-59	6	2
SRR1388323	36	0	cereb	m	60-69	7.7	4
SRR1421177	36	1	cortex	m	60-69	7.7	4
SRR1359673	36	0	sn	m	60-69	7.7	4
SRR1323196	36	1	hippo	m	60-69	6.9	4
SRR1327975	40	0	cereb	m	50-59	6.2	2
SRR1474777	40	1	cortex	m	50-59	6.2	2
SRR1458180	40	0	sn	m	50-59	6.2	2
SRR1358501	40	1	hippo	m	50-59	6.9	2
SRR1499442	41	0	cereb	m	60-69	8.5	2
SRR1318163	41	1	cortex	m	60-69	8.5	2
SRR1430570	41	0	sn	m	60-69	8.5	2
SRR1440082	41	1	hippo	m	60-69	8	2
SRR1445328	44	0	cereb	m	40-49	7.1	2
SRR1414897	44	1	cortex	m	40-49	7.1	2
SRR1415337	44	0	sn	m	40-49	7.1	2
SRR817758	44	1	hippo	m	40-49	6.3	2
SRR1406917	47	0	cereb	m	50-59	6.4	2

sample_id	subject_id	ctrl	Roi	sex	age	RIN	HardyScale
SRR1408208	47	1	cortex	m	50-59	6.4	2
SRR1413959	47	0	sn	m	50-59	6.4	2
SRR821690	47	1	hippo	m	50-59	6.8	2
SRR1488751	49	0	cereb	m	50-59	6.5	2
SRR1440986	49	1	cortex	m	50-59	6.5	2
SRR1360868	49	0	sn	m	50-59	6.5	2
SRR598768	49	1	hippo	m	50-59	7.1	2
SRR1431366	51	0	cereb	m	60-69	7.5	2
SRR1418837	51	1	cortex	m	60-69	7.5	2
SRR1338923	51	0	sn	m	60-69	7.5	2
SRR602169	51	1	hippo	m	60-69	6.5	2
SRR615599	61	0	cereb	m	40-49	7.8	4
SRR613627	61	1	cortex	m	40-49	7.8	4
SRR615467	61	0	sn	m	40-49	7.8	4
SRR660733	61	1	hippo	m	40-49	6.6	4
SRR599225	62	0	cereb	m	30-39	7	0
SRR601500	62	1	cortex	m	30-39	7	0
SRR664048	62	0	sn	m	30-39	7	0
SRR659673	62	1	hippo	m	30-39	7.3	0
SRR664819	64	0	cereb	m	50-59	6.1	4
SRR657754	64	1	cortex	m	50-59	6.1	4
SRR655111	64	0	sn	m	50-59	6.1	4
SRR662007	64	1	hippo	m	50-59	6.1	4
SRR655684	66	0	cereb	m	20-29	8.3	2
SRR662534	66	1	cortex	m	20-29	8.3	2
SRR663693	66	0	sn	m	20-29	8.3	2
SRR660969	66	1	hippo	m	20-29	8.3	2
SRR660812	67	0	cereb	m	50-59	8	2
SRR662233	67	1	cortex	m	50-59	8	2
SRR657973	67	0	sn	m	50-59	8	2
SRR660103	67	1	hippo	m	50-59	6.4	2
SRR662799	68	0	cereb	m	60-69	6.3	3
SRR657127	68	1	cortex	m	60-69	6.3	3
SRR661925	68	0	sn	m	60-69	6.3	3
SRR5125536	68	1	hippo	m	60-69	7.1	3
SRR1080927	73	0	cereb	m	60-69	7.8	4
SRR1085495	73	1	cortex	m	60-69	7.8	4
SRR5125533	73	0	sn	m	60-69	7.8	4
SRR1475969	73	1	hippo	m	60-69	7.8	4
SRR1458422	77	0	cereb	m	60-69	7	2
SRR1389979	77	1	cortex	m	60-69	7	2
SRR1499823	77	0	sn	m	60-69	7	2
SRR1487295	77	1	hippo	m	60-69	8	2
SRR1429307	80	0	cereb	m	60-69	8.1	2

sample_id	subject_id	ctrl	Roi	sex	age	RIN	HardyScale
SRR1447693	80	1	cortex	m	60-69	8.1	2
SRR1420625	80	0	sn	m	60-69	8.1	2
SRR1472323	80	1	hippo	m	60-69	8.1	2

8.3 GO enrichment analyses for CO- and HP-genes

8.3.1 GO Biological Process enrichment

In Chapter 3 (Methods 3.2.6.9), the differentially-expressed gene lists, CO-genes and HP-genes, were analysed using EnrichR in R for enrichment in GO Biological Processes. Table 8.1 displays all terms from both analyses. These terms were reduced, clustered separately, and visualised using functional annotation clustering in DAVID in Section 3.3.9.

Table 8.1 GO Biological Process terms enriched for CO/HP Genes
GO Biological Process (2021) terms enriched for genes from either the CO-genes or HP-genes set were calculated using enrichR in R. Only significant biological processes with an adjusted P value < 0.05 were included. Hits: number of genes in the set/number of genes associated with the term. The rows were placed in descending order by the maximum 'combined score' between the two lists (CO or HP) (enrichr's combined score for P value and odds ratio). Row colours: light green CO-only; blue HP-only; light-grey enriched with genes from both CO- and HP-gene sets.

GO Biological Process 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
adenosine 5'-(hexahydrogen pentaphosphate) catabolic process (GO:1901911)	866.0	4/5		
adenosine 5'-(hexahydrogen pentaphosphate) metabolic process (GO:1901910)	866.0	4/5		
diadenosine hexaphosphate catabolic process (GO:1901909)	866.0	4/5		
diadenosine hexaphosphate metabolic process (GO:1901908)	866.0	4/5		
diadenosine pentaphosphate catabolic process (GO:1901907)	866.0	4/5		
diadenosine pentaphosphate metabolic process (GO:1901906)	866.0	4/5		
diphosphoinositol polyphosphate metabolic process (GO:0071543)	866.0	4/5		
regulation of cilium movement (GO:0003352)			465.3	5/10

GO Biological Process 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
diadenosine polyphosphate catabolic process (GO:0015961)	389.9	4/6		
regulation of neurotransmitter receptor activity (GO:0099601)	179.0	17/58	386.4	16/58
regulation of epinephrine secretion (GO:0014060)			289.5	3/6
cerebral cortex cell migration (GO:0021795)			289.5	3/6
positive regulation of heart rate by epinephrine-norepinephrine (GO:0001996)			289.5	3/6
adenylate cyclase-inhibiting serotonin receptor signalling pathway (GO:0007198)	238.0	4/7		
positive regulation of behaviour (GO:0048520)			226.0	4/10
regulation of postsynaptic neurotransmitter receptor activity (GO:0098962)	96.3	4/10	226.0	4/10
anterograde trans-synaptic signalling (GO:0098916)	212.8	49/244	140.3	30/244
cellular response to macrophage colony-stimulating factor stimulus (GO:0036006)	211.2	3/5		
response to macrophage colony-stimulating factor (GO:0036005)	211.2	3/5		
G protein-coupled serotonin receptor signalling pathway (GO:0098664)	202.5	5/10		
regulation of cilium beat frequency (GO:0003356)			202.3	3/7
synaptic transmission, GABAergic (GO:0051932)	159.1	5/11	184.7	4/11
chemical synaptic transmission (GO:0007268)	183.7	55/306	105.1	32/306
potassium ion transport (GO:0006813)	55.5	20/122	170.5	20/122
G protein-coupled receptor signalling pathway, coupled to cyclic nucleotide second messenger (GO:0007187)	166.3	18/65	132.7	12/65
positive regulation of synaptic transmission (GO:0050806)	72.1	15/73	165.4	14/73
regulation of cation channel activity (GO:2001257)	97.6	19/89	165.3	16/89
adenylate cyclase-inhibiting G protein-coupled acetylcholine receptor signalling pathway (GO:0007197)			152.0	3/8
potassium ion transmembrane transport (GO:0071805)	47.6	21/139	151.6	21/139
adenylate cyclase-activating adrenergic receptor signalling pathway (GO:0071880)	104.0	6/17	147.3	5/17
G protein-coupled glutamate receptor signalling pathway (GO:0007216)	138.4	6/15		
regulation of synaptic transmission, glutamatergic (GO:0051966)	128.3	13/46		

GO Biological Process 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
phenol-containing compound biosynthetic process (GO:0046189)	123.2	4/9		
regulation of AMPA receptor activity (GO:2000311)	45.2	6/25	116.9	6/25
modulation of chemical synaptic transmission (GO:0050804)	113.8	23/109	59.4	13/109
adrenergic receptor signalling pathway (GO:0071875)	72.8	6/20	107.4	5/20
long-term synaptic potentiation (GO:0060291)	105.2	7/21	97.9	5/21
neuropeptide signalling pathway (GO:0007218)	100.2	15/63		
axoneme assembly (GO:0035082)			100.1	7/34
glutamate receptor signalling pathway (GO:0007215)	91.5	10/37		
adenylate cyclase-inhibiting G protein-coupled receptor signalling pathway (GO:0007193)	89.9	14/60	67.9	9/60
cilium movement (GO:0003341)			89.8	9/52
regulation of neuron migration (GO:2001222)	84.5	8/28	56.8	5/28
cation transport (GO:0006812)	18.7	19/174	81.9	20/174
positive regulation of heart rate (GO:0010460)			78.7	4/17
axonemal dynein complex assembly (GO:0070286)			73.0	6/32
regulation of NMDA receptor activity (GO:2000310)	42.3	7/32	73.0	6/32
regulation of cAMP-mediated signalling (GO:0043949)	71.9	7/25		
positive regulation of filopodium assembly (GO:0051491)			65.3	5/26
regulation of neuron projection development (GO:0010975)	64.2	26/165		
neuron differentiation (GO:0030182)	64.1	27/174	59.1	18/174
generation of neurons (GO:0048699)	63.4	30/202	50.3	19/202
regulation of postsynaptic membrane potential (GO:0060078)	34.8	7/35	61.5	6/35
adenylate cyclase-modulating G protein-coupled receptor signalling pathway (GO:0007188)	56.3	25/165	30.2	14/165
regulation of neuronal synaptic plasticity (GO:0048168)	52.2	7/29	53.1	5/29
inorganic cation transmembrane transport (GO:0098662)	21.5	28/274	53.0	24/274
negative regulation of peptide hormone secretion (GO:0090278)	52.7	5/18		
nervous system development (GO:0007399)	34.3	46/447	52.3	34/447
lamellipodium organization (GO:0097581)	52.2	7/29		
phospholipase C-activating G protein-coupled receptor signalling pathway (GO:0007200)	36.0	13/81	50.5	10/81

GO Biological Process 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
negative regulation of inflammatory response to antigenic stimulus (GO:0002862)	50.2	21/136	28.9	12/136
regulation of vasoconstriction (GO:0019229)			49.8	5/30
regulation of inflammatory response to antigenic stimulus (GO:0002861)	49.4	21/137	28.5	12/137
purine ribonucleotide catabolic process (GO:0009154)	49.3	6/24		
chloride transmembrane transport (GO:1902476)	34.2	9/50	47.5	7/50
negative regulation of neuron projection development (GO:0010977)	45.6	11/58	35.4	7/58
regulation of dendritic spine morphogenesis (GO:0061001)	42.3	7/32		
lamellipodium assembly (GO:0030032)	41.6	6/26		
lipid modification (GO:0030258)			41.4	6/43
cardiac ventricle morphogenesis (GO:0003208)			39.6	6/44
adenylate cyclase-activating G protein-coupled receptor signalling pathway (GO:0007189)	23.1	15/118	39.5	12/118
regulation of insulin secretion (GO:0050796)	39.1	16/104		
axon development (GO:0061564)	38.2	13/79		
negative regulation of cell projection organization (GO:0031345)	35.7	9/49		
neuron migration (GO:0001764)	34.2	9/50		
regulation of axonogenesis (GO:0050770)	32.7	9/51		
axonogenesis (GO:0007409)	31.5	28/240		
cellular response to calcium ion (GO:0071277)	30.7	11/69		
phospholipid metabolic process (GO:0006644)			30.0	8/76
negative regulation of immune response (GO:0050777)	29.8	22/178		
response to calcium ion (GO:0051592)	29.5	15/107		
negative regulation of inflammatory response (GO:0050728)	29.3	25/212		
cell morphogenesis involved in neuron differentiation (GO:0048667)	24.5	11/76		
peptidyl-serine phosphorylation (GO:0018105)			21.1	12/156
axon guidance (GO:0007411)	20.8	22/203		
synapse organization (GO:0050808)	19.5	15/126		
neuron projection morphogenesis (GO:0048812)	18.4	16/140		
central nervous system development (GO:0007417)	14.7	25/268		

8.3.2 GO Cellular Component enrichment

In Chapter 3 (Methods 3.2.6.9), the differentially-expressed gene lists, CO-genes and HP-genes, were analysed using EnrichR in R for enrichment in GO Cellular Components.

Table 8.2 displays all terms from both analyses. These terms were reduced, clustered separately, and visualised using functional annotation clustering from DAVID Bioinformatics in Section 3.3.9.

Table 8.2 GO Cellular Component terms enriched for CO- and HP-Genes
GO Cellular Component (2021) terms enriched for genes from either the CO-genes or HP-genes set were calculated using enrichR in R. Only significant cellular components with an adjusted P value < 0.05 were included. Hits: number of genes in the set/number of genes associated with the term. The rows were placed in descending order by the maximum 'combined score' between the two lists (CO or HP) (Enrichr's combined score for P value and odds ratio). Row colours: light green CO-only; blue HP-only; light-grey enriched with genes from both CO- and HP-gene sets.

GO Cellular Component 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
dendrite membrane (GO:0032590)	374.3	13/29	282.2	9/29
voltage-gated potassium channel complex (GO:0008076)	152.4	19/73	244.5	16/73
potassium channel complex (GO:0034705)	104.5	18/80	204.1	16/80
neuron projection (GO:0043005)	185.2	84/556	101.5	48/556
GABA-A receptor complex (GO:1902711)	47.0	5/19	118.6	5/19
dendrite (GO:0030425)	106.7	43/270	101.2	29/270
glutamatergic synapse (GO:0098978)	81.8	15/69		
integral component of postsynaptic specialization membrane (GO:0099060)	78.0	5/15		
integral component of postsynaptic density membrane (GO:0099061)	77.7	4/11		
intrinsic component of postsynaptic density membrane (GO:0099146)	77.7	4/11		
postsynaptic density membrane (GO:0098839)	72.8	6/20	58.0	4/20
postsynaptic specialization membrane (GO:0099634)	65.6	6/21	52.9	4/21
integral component of plasma membrane (GO:0005887)	62.6	131/1454	58.0	83/1454
AMPA glutamate receptor complex (GO:0032281)	42.2	5/20	58.0	4/20
GABAergic synapse (GO:0098982)	52.7	5/18		
postsynaptic density (GO:0014069)	48.5	21/138	28.0	12/138
ionotropic glutamate receptor complex (GO:0008328)	26.7	6/32	44.1	5/32
synaptic membrane (GO:0097060)	41.6	6/26		
asymmetric synapse (GO:0032279)	37.9	19/133	30.4	12/133
anchored component of plasma membrane (GO:0046658)			36.3	6/46

GO Cellular Component 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
cation channel complex (GO:0034703)	19.9	10/73	32.6	8/73
calcium channel complex (GO:0034704)	30.1	8/45		
voltage-gated calcium channel complex (GO:0005891)	28.6	6/31		
axon (GO:0030424)	17.2	21/204	18.2	14/204

8.3.3 GO Molecular Function enrichment

In Chapter 3 (Methods 3.2.6.9), the differentially-expressed gene lists, CO-genes and HP-genes, were analysed using EnrichR in R for enrichment in GO Molecular Functions. Table 8.3 displays all terms from both analyses. These terms were reduced, clustered separately, and visualised using functional annotation clustering in DAVID in Section 3.3.9.

Table 8.3 GO Molecular Function enriched for CO/HP Genes
GO Molecular Function (2021) terms enriched for genes from either the CO-genes or HP-genes set were calculated using enrichR in R. Only significant molecular functions with an adjusted P value < 0.05 were included. Hits: number of genes in the set/number of genes associated with the term. The rows were placed in descending order by the maximum 'combined score' between the two lists (CO or HP) (Enrichr's combined score for P value and odds ratio). Row colours: light green CO-only; blue HP-only; light-grey enriched with genes from both CO- and HP-gene sets.

GO Molecular Function 2021 Terms	**CO Score**	**CO Hits**	**HP Score**	**HP Hits**
bis(5'-adenosyl)-hexaphosphatase activity (GO:0034431)	866.0	4/5		
bis(5'-adenosyl)-pentaphosphatase activity (GO:0034432)	866.0	4/5		
diphosphoinositol-polyphosphate diphosphatase activity (GO:0008486)	866.0	4/5		
endopolyphosphatase activity (GO:0000298)	866.0	4/5		
alpha-adrenergic receptor activity (GO:0004936)			289.5	3/6
benzodiazepine receptor activity (GO:0008503)	202.5	5/10	226.0	4/10
neuropeptide Y receptor activity (GO:0004983)	211.2	3/5		
G protein-coupled acetylcholine receptor activity (GO:0016907)	87.6	3/7	202.3	3/7
G protein-coupled neurotransmitter receptor activity (GO:0099528)	87.6	3/7	202.3	3/7
voltage-gated potassium channel activity (GO:0005249)	98.9	18/82	194.4	16/82

GO Molecular Function 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
voltage-gated cation channel activity (GO:0022843)	193.9	25/97	115.2	15/97
extracellular ligand-gated ion channel activity (GO:0005230)	159.1	5/11	184.7	4/11
inhibitory extracellular ligand-gated ion channel activity (GO:0005237)	162.8	6/14	114.0	4/14
serotonin receptor activity (GO:0099589)	158.9	11/33	106.0	7/33
G protein-coupled serotonin receptor activity (GO:0004993)	158.9	11/33	106.0	7/33
phosphatidate phosphatase activity (GO:0008195)			131.8	4/13
GABA-gated chloride ion channel activity (GO:0022851)	107.3	5/13	131.8	4/13
G protein-coupled opioid receptor activity (GO:0004985)	127.5	3/6		
calmodulin-dependent protein kinase activity (GO:0004683)			126.5	6/24
adenylate cyclase inhibiting G protein-coupled glutamate receptor activity (GO:0001640)	123.2	4/9		
G protein-coupled glutamate receptor activity (GO:0098988)	123.2	4/9		
GABA-A receptor activity (GO:0004890)	47.0	5/19	118.6	5/19
lipid phosphatase activity (GO:0042577)			114.0	4/14
G protein-coupled amine receptor activity (GO:0008227)	109.6	11/39	76.9	7/39
potassium channel activity (GO:0005267)	78.1	18/91	106.9	14/91
delayed rectifier potassium channel activity (GO:0005251)	59.2	8/33	106.0	7/33
neurotransmitter receptor activity involved in regulation of postsynaptic membrane potential (GO:0099529)	38.1	5/21	97.9	5/21
G protein-coupled receptor activity (GO:0004930)	94.8	43/283	22.2	19/283
ligand-gated anion channel activity (GO:0099095)	91.7	6/18	70.6	4/18
GABA receptor activity (GO:0016917)	34.5	5/22	89.6	5/22
[heparan sulfate]-glucosamine 3-sulfotransferase 1 activity (GO:0008467)	87.6	3/7		
glutamate receptor binding (GO:0035254)	59.6	5/17	78.7	4/17
G protein-coupled photoreceptor activity (GO:0008020)	78.0	5/15		
heparan sulfate sulfotransferase activity (GO:0034483)	78.0	5/15		
nucleotide diphosphatase activity (GO:0004551)	78.0	5/15		
cGMP binding (GO:0030553)	77.7	4/11		

GO Molecular Function 2021 Terms	CO Score	CO Hits	HP Score	HP Hits
calcium-activated potassium channel activity (GO:0015269)			68.8	3/12
neuropeptide hormone activity (GO:0005184)	41.6	6/26	65.3	5/26
transmitter-gated ion channel activity (GO:0022824)	37.1	7/34	65.0	6/34
cAMP binding (GO:0030552)	64.3	4/12		
m7G(5')pppN diphosphatase activity (GO:0050072)	64.3	4/12		
transmembrane-ephrin receptor activity (GO:0005005)	59.6	5/17		
ephrin receptor activity (GO:0005003)	59.6	5/17		
G protein-coupled peptide receptor activity (GO:0008528)	55.3	15/82		
hormone activity (GO:0005179)	16.9	10/78	54.5	10/78
cyclic nucleotide binding (GO:0030551)	54.0	6/23		
hormone receptor binding (GO:0051427)			52.9	4/21
transmitter-gated ion channel activity involved in regulation of postsynaptic membrane potential (GO:1904315)			49.8	5/30
neuropeptide receptor activity (GO:0008188)	48.6	7/30		
acetylcholine receptor activity (GO:0015464)			48.5	4/22
peptide hormone receptor binding (GO:0051428)	46.3	4/14		
voltage-gated calcium channel activity (GO:0005245)	46.2	8/37		
neuropeptide binding (GO:0042923)	38.1	5/21		
potassium channel regulator activity (GO:0015459)			30.8	6/50
receptor ligand activity (GO:0048018)	17.4	29/307	21.5	20/307
serine-type peptidase activity (GO:0008236)	15.7	14/125		
protein serine/threonine kinase activity (GO:0004674)	10.2	28/344		

8.3.4 KEGG Pathways enrichment

In Chapter 3 (Methods 3.2.6.9), the differentially-expressed gene lists, CO-genes and HP-genes, were analysed using EnrichR in R for enrichment in KEGG 2021 Human Pathways. Table 8.4 displays all terms from both analyses. These terms were reduced, clustered separately, and visualised using functional annotation clustering in DAVID in Section 3.3.9.

Table 8.4 KEGG 2021 Human pathways enriched for CO/HP-genes

KEGG 2021 Human pathways enriched for genes from either the CO-genes or HP-genes sets were calculated using enrichR in R. Only significant pathways with an adjusted P value < 0.05 were included. Hits: number of genes in the set/number of genes associated with the pathway. The rows were placed in descending order by the maximum 'combined score' between the two lists (CO or HP) (Enrichr's combined score for P value and odds ratio). Row colours: light green CO-only; blue HP-only; light-grey enriched with genes from both CO- and HP-gene sets.

KEGG 2021 Human Pathways	CO Score	CO Hits	HP Score	HP Hits
Neuroactive ligand-receptor interaction	179.9	59/341	169.8	40/341
Calcium signalling pathway	70.0	35/240	46.7	21/240
Salivary secretion	50.9	16/93	27.8	9/93
Axon guidance	32.8	23/182		
Renin secretion	52.1	13/69		
Melanogenesis	33.9	15/101		
Amphetamine addiction	40.5	12/69		
cAMP signalling pathway	24.0	24/216		
Oxytocin signalling pathway	26.1	19/154		
Circadian entrainment	29.8	14/97		
Taste transduction	31.3	13/86		
Glutamatergic synapse	25.2	15/114		
Inflammatory mediator regulation of TRP channels	22.8	13/98		
Gap junction	22.8	12/88		
Wnt signalling pathway	17.6	18/166		
Cholinergic synapse	20.4	14/113		
Serotonergic synapse	20.4	14/113		
cGMP-PKG signalling pathway	17.3	18/167		
Rap1 signalling pathway	15.8	21/210		
Cushing syndrome	17.3	17/155		
Long-term potentiation	24.3	10/67		
Dopaminergic synapse	17.2	15/132		
Adrenergic signalling in cardiomyocytes	15.3	16/150		
Aldosterone synthesis and secretion	17.5	12/98		
Insulin secretion	18.2	11/86		
Nicotine addiction	26.0	7/40	47.6	6/40
Gastric acid secretion	18.0	10/76		
Glycosaminoglycan biosynthesis	20.8	8/53		
Morphine addiction	15.8	11/91	39.7	10/91
Gastric cancer	12.3	15/149		
Vascular smooth muscle contraction			23.2	11/133

8.3.5 Reactome 2022 Pathways enrichment

In Chapter 3 (Methods 3.2.6.9), the differentially-expressed gene lists, CO-genes and HP-genes, were analysed using EnrichR in R for enrichment in Reactome 2022 Pathways. Table 8.5 displays all terms from both analyses. These terms were reduced, clustered separately, and visualised using functional annotation clustering in DAVID in Section 3.3.9.

Table 8.5 Reactome 2022 pathways enriched for CO/HP-genes

Reactome 2022 pathways enriched for genes from either the CO-genes or HP-genes sets were calculated using enrichR in R. Only significant pathways with an adjusted P value < 0.05 were included. Hits: number of genes in the set/number of genes associated with the pathway. The rows were placed in descending order by the maximum 'combined score' between the two lists (CO or HP) (Enrichr's combined score for P value and odds ratio). Row colours: light green CO-only; blue HP-only; light-grey enriched with genes from both CO- and HP-gene sets.

Reactome 2022 Terms	CO Score	CO Hits	HP Score	HP Hits
Muscarinic Acetylcholine Receptors R-HSA-390648	211.2	3/5	471.1	3/5
Serotonin Receptors R-HSA-390666	303.3	6/11		
Voltage Gated Potassium Channels R-HSA-1296072	235.8	15/43	299.4	12/43
Adrenoceptors R-HSA-390696	123.2	4/9	285.6	4/9
Amine Ligand-Binding Receptors R-HSA-375280	278.8	15/40	267.6	11/40
Lysosphingolipid And LPA Receptors R-HSA-419408			218.1	5/14
Neuronal System R-HSA-112316	176.2	64/386	129.0	40/386
Signalling By GPCR R-HSA-372790	147.0	92/689	78.6	52/689
GPCR Ligand Binding R-HSA-500792	134.3	67/458	94.2	41/458
GPCR Downstream Signalling R-HSA-388396	127.2	82/619	47.6	42/619
Potassium Channels R-HSA-1296071	84.4	20/102	126.0	16/102
Class A/1 (Rhodopsin-like Receptors) R-HSA-373076	104.1	49/327	81.5	31/327
Neurotransmitter Receptors and Postsynaptic Signal Transmission R-HSA-112314	50.3	26/182	86.5	21/182
Glutamate Binding, Activation of AMPA Receptors and Synaptic Plasticity R-HSA-399721	45.3	7/31	77.5	6/31
Transmission Across Chemical Synapses R-HSA-112315	65.8	35/246	51.3	22/246

Reactome 2022 Terms	CO Score	CO Hits	HP Score	HP Hits
Nitric Oxide Stimulates Guanylate Cyclase R-HSA-392154	65.6	6/21		
G Alpha (I) Signalling Events R-HSA-418594	56.7	40/312		
Regulation Of Insulin Secretion R-HSA-422356	51.3	14/77		
Glucagon-like Peptide-1 (GLP1) Regulates Insulin Secretion R-HSA-381676	50.2	9/42		
ADORA2B Mediated Anti-Inflammatory Cytokine Production R-HSA-9660821	46.8	20/131		
G Alpha (Q) Signalling Events R-HSA-416476	44.0	28/212	25.7	16/212
G Alpha (S) Signalling Events R-HSA-418555	44.0	22/153		
Phase 0 - Rapid Depolarisation R-HSA-5576892	42.3	7/32		
Collagen Chain Trimerization R-HSA-8948216			39.6	6/44
Anti-inflammatory Response Favouring Leishmania Infection R-HSA-9662851	36.3	22/165		
GABA Receptor Activation R-HSA-977443			34.2	7/59
Class B/2 (Secretin Family Receptors) R-HSA-373080	33.9	14/92	28.4	9/92
Protein-protein Interactions at Synapses R-HSA-6794362	29.6	13/88		
PLC Beta Mediated Events R-HSA-112043	27.3	8/47		
Neurexins And Neuroligins R-HSA-6794361	25.4	9/57		
Integration Of Energy Metabolism R-HSA-163685	24.6	14/105		
Peptide Ligand-Binding Receptors R-HSA-375276	22.9	22/196		
Opioid Signalling R-HSA-111885	22.8	12/88		
Signal Transduction R-HSA-162582	19.2	163/2465	11.9	93/2465
Leishmania Infection R-HSA-9658195	13.7	23/247		

8.4 Deleterious SIFT and Damaging PolyPhen2 Rare Variants in SCZ000

The following variants have an existing SNP id that was annotated by VEP. Other variants may not have been annotated by VEP, or have been recorded in since:

- MAST4 (5:66460390) - rs17221458
- CSMD1 (8:2800096) - rs11984691

Table 8.6 Rare variants in SCZ000 - deleterious by SIFT or damaging by PolyPhen2
Chromosome positions and REF/ALT for each rare variant from group SCZ000 is given along with the amino acid change caused by the genomic variant. Allele Count (AC) is 1 but is measured against the full Swedish Exomes dataset. gnomAD AC is less than 5. SIFT < 0.05 and PolyPhen-2 score > 0.9.

Chrom	Position	REF	ALT	Gene	AC	gnomAD AC	Amino Position	Amino Change
1	21050974	G	A	SH2D5	1	3	98	P/L
1	49100226	G	A	AGBL4	1	3	30	P/L
1	70504560	G	T	LRRC7	1	1	980	R/L
1	82416118	A	T	LPHN2	1	1	482	T/S
1	82445626	G	C	LPHN2	1	4	1028	S/T
1	86289241	A	C	COL24A1	1	1	1256	L/R
1	167962487	A	C	DCAF6	1	3	238	T/P
1	200978429	C	T	KIF21B	1	5	77	E/K
2	46142	G	T	FAM110C	1	1	82	P/T
2	80646620	C	T	CTNNA2	1	1	74	T/I
2	85097573	C	A	TRABD2A	1	2	149	A/S
2	136409346	A	G	R3HDM1	1	1	556	Q/R
2	162273577	G	A	TBR1	1	1	219	R/Q
2	175201773	C	A	SP9	1	0	127	A/S
3	25636116	T	A	RARB	1	1	366	M/K
3	37725445	A	G	ITGA9	1	1	669	Y/C
3	122632834	G	A	SEMA5B	1	1	668	T/I
3	186461548	C	T	KNG1	1	3	420	P/S
4	30723792	G	A	PCDH7	1	0	250	D/N
4	66213872	C	T	EPHA5	1	3	853	R/Q
4	72897646	G	A	NPFFR2	1	1	10	A/T
4	79387402	A	G	FRAS1	1	2	2357	H/R
5	43277243	A	G	NIM1K	1	1	126	E/G
5	66460390	G	T	MAST4	1		1601	V/F
5	101576430	C	T	SLCO4C1	1	3	623	R/Q
5	140870378	A	G	PCDHGC5	1	2	524	Y/C
5	175110699	C	A	HRH2	1	1	155	H/N
6	76072542	A	C	FILIP1	1	1	123	V/G
6	107419781	C	T	BEND3	1	1	72	V/M
6	117710872	A	G	ROS1	1	1	476	I/T
7	16737785	A	G	BZW2	1	1	361	Q/R
7	21744094	G	A	DNAH11	1	1	2113	V/M
7	39247183	G	A	POU6F2	1	5	159	G/R
7	83675732	G	C	SEMA3A	1	2	192	A/G
7	132192587	G	A	PLXNA4	1	3	289	A/V
7	142563248	A	G	EPHB6	1	2	322	Y/C
8	2800096	C	G	CSMD1	1	4	2881	S/T
8	2876132	G	C	CSMD1	1	2	2050	N/K
8	3081374	T	C	CSMD1	1	1	935	N/S
8	3266995	A	G	CSMD1	1	3	46	I/T
8	35544131	C	T	UNC5D	1	1	274	R/W
8	51449304	C	T	SNTG1	1	1	206	L/F

Chrom	Position	REF	ALT	Gene	AC	gnomAD AC	Amino Position	Amino Change
9	6589288	A	G	GLDC	1	1	496	L/P
9	130474994	C	T	C9orf117	1	3	31	R/C
9	135102369	G	A	NTNG2	1	1	331	G/S
9	136507387	C	T	DBH	1	3	119	A/V
10	35896842	T	A	GJD4	1	2	134	L/H
10	99338034	G	A	ANKRD2	1	3	103	R/Q
10	104176267	G	T	PSD	1	4	177	P/T
11	13410443	G	A	BTBD10	1	2	455	L/F
11	33572692	G	A	KIAA1549L	1	1	912	G/E
11	40136996	G	A	LRRC4C	1	2	283	P/S
11	46393142	C	T	DGKZ	1	3	227	R/W
11	66049755	C	A	CNIH2	1	3	36	T/N
11	85422166	G	A	SYTL2	1	1	608	S/F
11	88323804	C	T	GRM5	1	1	552	C/Y
11	92616544	G	C	FAT3	1	2	4308	D/H
12	11546748	T	A	PRB2	1		88	K/N
12	111749919	G	C	CUX2	1	1	639	R/T
13	43360902	C	T	FAM216B	1	1	35	L/F
13	84454327	G	C	SLITRK1	1	4	439	S/C
15	41805874	C	T	LTK	1	1	12	G/R
15	41805877	A	C	LTK	1	1	11	F/V
15	52011668	G	A	SCG3	1	1	451	G/D
16	84050784	G	A	SLC38A8	1	2	305	S/F
17	4541950	A	T	ALOX15	1	2	191	L/Q
17	59820424	G	C	BRIP1	1	3	777	R/G
17	61499197	A	T	TANC2	1	1	1962	T/S
17	80332631	T	C	UTS2R	1	5	144	M/T
18	28736086	G	A	DSC1	1	2	131	R/C
19	1235865	G	A	C19orf26	1	4	47	S/L
19	3907836	C	T	ATCAY	1	3	155	R/W
19	47918175	G	A	MEIS3	1	3	194	P/S
20	19679317	G	A	SLC24A3	1	1	571	G/E
20	43588929	C	T	TOMM34	1	1	16	G/S
21	33340634	C	T	HUNK	1	2	316	A/V
22	30125126	T	C	CABP7	1	1	150	I/T
22	34046516	T	C	LARGE	1	3	82	Q/R
22	39239250	T	C	NPTXR	1	0	205	D/G
X	99596982	G	A	PCDH19	2	4	876	R/W

8.5 Genes with Rare Damaging Variants in 3500 Schizophrenia Cases

8.5.1 Genes with RDVs in SCZ cases only

ABRACL, AK5, ARX, ASCL1, BEND3, C16orf74, C16orf87, C9orf16, CBLN2, COG8, CXCL14, DLX5, DNAH6, E2F3, EFNB3, EGFL6, FAM19A2, FAM216B, FATE1, FGF16,

FOXG1, GPR21, HAPLN1, HPCA, HRH2, IGFL4, IL5RA, IQCF3, KIAA1671, LIPC, LMO3, LRRC4C, LYNX1, LYPD6B, MAS1, NNAT, NPY, NRSN1, NUDT11, OAF, OTOS, PCP4L1, PDP1, PLK2, PPP1R1A, PPP3R1, PRM3, PTHLH, RAP2B, RASD1, RDM1, RHEBL1, RMI2, SCGB3A1, SERPINI1, SERTM1, SLC22A10, SOWAHA, TEX29, TMEM183BP, TMEM213, UBE2E2, UBP1, UTS2R, VPS36, YWHAZ

8.5.2 Genes in which SCZ cases have over twice the number of RDVs

ABR, AGBL4, B3GNT5, BTBD10, C1QTNF9B, CA7, CABP1, CACNG8, CAMK2A, CDRT15L2, CHRM4, CHRNA1, CNIH2, CREG2, CTNNA2, CYP46A1, DAW1, DCLK3, DRC1, EPHX4, FAM155A, FOXP1, GABRA2, GABRA3, GFRA4, GPR150, GRM2, HTR2A, IGFBP2, KCNJ4, KL, KMO, LCN15, LRRC40, LRRC6, NRG3, OLFM1, PRSS16, RFX6, RGS20, RGS4, ROPN1L, RTN4RL2, SATB1, SATB2, SH2D5, SLC1A2, THEG, TOMM34, WNT16

8.5.3 Genes in which SCZ cases have more RDV but less than double

ADAD2, ARF3, ASIC5, ATP5S, BZW2, C2orf50, C4orf50, C6orf222, CACNG6, CCDC74A, CDH9, CDHR4, CDK5, CHRDL1, CLPSL1, CPNE7, CRYGA, CSRNP3, CYP4Z1, DDX54, DGAT2, DNAAF3, DPF1, DRGX, FAIM2, FAM163B, FAM49A, FGF12, FKBP11, GALNTL5, GDA, GDF7, GIT1, GRASP, GYPB, HCN1, HS6ST1, HTR4, IL1RL2, KCNV1, KCTD17, KIAA1045, LINGO1, LPAR3, LPAR4, LRRTM4, MLIP, MMD, NCALD, NIPAL2, NPAS1, NPY2R, PALM, PCDH8, PDE4C, PIK3R1, PNMA5, PRDM8, PRM2, PTGFR, RBP4, RGR, RHOBTB2, SCG3, SEMA3A, SH2D1B, SLC22A9, SLC24A3, SLC25A23, SMYD2, SNTG1, SORCS1, STS, THY1, TMBIM1, TNNT2, TOLLIP, TSHZ3, TSPAN16, TUBA4B, UPP2, WASF1, WNT10B, WNT2

8.5.4 Genes have at least one RDV in one case but more in controls

ABCA12, ABCG8, ABLIM2, ACPL2, ACRV1, ACTR3B, ADAMTS8, ADORA2B, ADRA1B, ADTRP, AHSG, ALG1L2, ALOX15, ANKRD24, ANO3, AP1S1, ARHGAP28, ARHGAP32, ATCAY, ATL1, ATP12A, ATRNL1, AZU1, BAI2, BAIAP3, BASP1, BCAS4, BCL11B, BEND5, BHMT, BICC1, C19orf26, C19orf60, C1orf194, C1orf216, C22orf15, C9orf117, CABP7, CACNA1E, CACNA2D1, CACNB1, CAPN11, CCDC108, CCDC3, CCDC37, CCDC85A, CCNA1, CDH12, CDH4, CDK5R2, CDX1, CHD5, CHL1, CHRM1, CHRM3, CHSY3, CIDEA, CIT, CITED1, CLEC2L, CLSTN2, CNGB1, CNGB3, CNTN5, CNTNAP5, COL5A3, CPNE4, CPO, CRHR2, CRYGN, CSMD1, CUX2, CX3CL1, CYP4A22, DAGLA, DBH, DCDC1, DCSTAMP, DGKB, DGKZ, DIO2, DLX3, DNAH11, DNAH3, DNAL4, DSC1, DUSP19, DUSP2, EFNA3, ELANE, ELAVL4, EMX1, EMX2, EPHA5, EPHB6, ESRP1, FAM171B,

FAM19A1, FAM237A, FAM65B, FAM81A, FAR2, FBXL2, FBXO34, FHL2, FHOD3, FIP1L1, FMN1, FNDC1, FRAS1, FSTL4, GABRG3, GAD1, GAS2L2, GJD2, GJD4, GLRB, GNA14, GPR6, GPR63, GREM2, GRM5, GRM7, GUCY1A3, GUCY1B3, HDGFL1, HIPK4, HLF, HS3ST2, HS6ST3, HSPB3, HTR3A, HUNK, IDS, IL12RB2, ILDR2, INSM2, ISLR2, ITGA9, ITPKA, JAKMIP1, KATNB1, KCNA4, KCNH4, KCNH5, KCNIP2, KCNIP3, KCNMA1, KCNQ5, KCTD13, KCTD16, KHDRBS3, KIAA1549, KNG1, KREMEN2, KRT84, LAMP5, LARGE, LHX6, LIMD2, LINGO2, LMO4, LMO7, LNX1, LPA, LPHN2, LPPR5, LRFN5, LRIT1, LTK, LY6D, LY6H, MAEL, MAST4, MATK, MCF2L2, MCHR1, MEPE, MET, METTL7B, MOXD1, MS4A8, MSLNL, MUC15, MYO5B, NAGPA, NAT8L, NDFIP2, NEK2, NGB, NIM1K, NR2E1, NRP1, NTNG2, NUAK1, NUDT4, OPRD1, OR2AG1, OR2T34, OR2T6, OSTN, OVOL2, OXGR1, P2RX2, PAX1, PCSK2, PDE1B, PDE2A, PGM2L1, PHACTR1, PKP2, PLEKHA1, PLEKHS1, PLPPR2, PLXNA4, PNPLA5, POSTN, POU6F2, PPFIA2, PPP1R16B, PPP4R4, PRICKLE2, PRSS12, PRTN3, PSD, PTGES3L-AARSD1, PTPRN2, PTPRT, PTRH1, RAPGEF2, RAPGEF4, RBBP8NL, RFPL2, RFTN1, RIMS3, RS1, RTBDN, RTKN2, SCN3B, SDR16C5, SEC16B, SFRP2, SHANK2, SHC3, SHD, SHROOM2, SLC22A25, SLCO5A1, SLIT1, SLITRK1, SNX32, SPAG6, ST6GAL2, STOML3, STRA6, STYK1, SYNPO, SYTL2, SYTL5, TBR1, TCHH, TEKT4, TESC, THRB, TMEM150C, TMEM155, TMPRSS13, TNFAIP8L3, TNFRSF11A, TNNI3K, TRABD2A, TRIM71, TRIML2, TRUB1, TSPAN13, TTC40, TUSC5, UNC5D, VGF, VSTM2A, VWA3A, WDR64, WDR65, WDR86, WNT1, YIF1A, ZFR2, ZNF184