Single-cell RNA-sequencing in epilepsy; discovery of cell-types, pathways, and drug targets

Maya Abouzeid

Submitted in part fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Imperial College London

Faculty of Medicine

Department of Brain Sciences

January 2024

Abstract

Epilepsy is a complex neurological disorder, affecting millions of people worldwide. It is characterised by spontaneous seizures caused by hyperexcitability of neurones in the brain. There is a large unmet need for new epilepsy therapeutic strategies, with one in three patients experiencing uncontrolled seizures despite medication. This thesis explores the cellular and molecular landscape of epilepsy at a single-cell level using single-nucleus RNA sequencing, offering greater resolution to the diverse cell-types involved in epilepsy. Using tissue from three mouse models of epilepsy; Pilocarpine induced status epilepticus, Self-Sustaining status epilepticus, Focal Cortical Dysplasia, and two human epilepsies; Hippocampal Sclerosis, and Focal Cortical Dysplasia, this work aimed to identify converging signatures across multiple forms of epilepsy and brain regions.

Using an optimised bioinformatic pipeline, this work identified differentially expressed genes (DEGs) in epilepsy, providing insights into the transcriptional and epigenetic changes associated with the disease state. Functional enrichment analyses revealed pathways implicated in neuronal excitability and neuroinflammation, emphasizing the role of glial cells as well as neurons in epilepsy pathology.

These analyses were then integrated with epilepsy GWAS and rare-variant analyses, to associate epilepsy risk with these transcriptional changes. A convergent signature was identified in astrocytes across both monogenic and polygenic epilepsies in both mouse and human forms of epilepsy. Perturbation datasets were integrated with the DEGs to predict potential regulators of epilepsy that could be targeted for treatment. With a focus on drug repurposing, these targets were intersected with interaction databases to identify candidate drugs for the treatment of epilepsy, focusing on disease modification. We highlight predicted

targets with pre-approved molecules, some which have been previously shown to be effective in epilepsy. This work enhances our understanding of epilepsy at a single-cell level, highlights the heterogeneity of the disease, identifies associated pathways, and points to potential therapeutic targets.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Statement of Originality

I, Maya Abouzeid, hereby declare that the work presented in this thesis, except where specific reference is made to the work of others, is original and conducted by me. All sources of information are appropriately referenced and where information has been derived from other collaboration with others, this has been acknowledged and cited. Where the thesis is based on work done by myself jointly with others, I have made clear what was done by others and what I have contributed myself. None of this work has been published before submission.

Publications not directly related to this thesis:

Feleke, R., Jazayeri, D., **Abouzeid, M**., Powell, K.L., Srivastava, P.K., O'Brien, T.J., Jones, N.C. and Johnson, M.R., 2022. Integrative genomics reveals pathogenic mediator of valproateinduced neurodevelopmental disability. Brain, 145(11), pp.3832-3842.

Haglund A, Zuber V, Yang Y, **Abouzeid M**, Feleke R, Ko JH, et al. Single-cell Mendelian randomisation identifies cell-type specific genetic effects on human brain disease and behaviour. bioRxiv. 2022 Nov 29;2022.11.28.517913.

Acknowledgements

I am deeply grateful for all the support and encouragement I have received throughout my PhD and to everyone who has been part of this journey.

Firstly, I would like to thank my supervisors, Professor Michael Johnson and Dr Prashant Srivastava for their guidance, and support throughout this journey. Their mentorship has been invaluable to my growth as a researcher.

I would also like to thank the members of the Johnson Lab past and present; Rahel Feleke, Alexander Haglund, Joeong Hun, Liisi Laniste, Yifei Yang and Daniel Clode for their support and guidance but mostly for their friendships. I am fortunate to have shared this experience with such kind and talented individuals who have made the most challenging moments more bearable and the successes even more rewarding.

I would also like to thank the team at UCB Pharma, whose collaboration and support were instrumental in making this research possible.

Finally, to my family and my friends, thank you for your love, patience, and unwavering belief in me, even when I doubted myself. You have been my greatest source of strength and comfort and I am grateful to have had you by my side in this chapter of my life.

Table of Contents

List of Figures **12**
List of Tables 15
Abbreviations 17
1 Introduction 19

1.1 Epilepsy **19**
1.1.1 Clinical symptoms and diagnosis 19
1.1.2 Pathogenesis and aetiology 22
1.1.3 Genetics of Epilepsy 23
1.1.4 Treatment and Drugs 26
1.1.5 Subtypes of Epilepsy 28
1.1.6 Modelling Epilepsy 30

1.2 Transcriptomics in epilepsy **31**
1.2.1 Single-cell RNA-sequencing 32
1.2.2 Cell-types in epilepsy 35

1.3 Aims **36**
1.3.1 Thesis outline 38
1.3.2 Limitations and Future Work 39

2 Methods and Pipeline Optimisation 44

2.1 Introduction **44**
2.1.1 Computational Analysis 45
2.1.1 Aims 46

2.2 Methods **48**
2.2.1 Experimental Protocol 48
2.2.2 Mapping 50
2.2.3 Data processing and quality control 53
2.2.4 Normalisation 57

2.2.5 Data Integration 57
2.2.6 Clustering 59
2.2.7 Dimensionality Reduction 60
2.2.8 Cell-type Annotation 62
2.2.9 Differential Expression Analysis 62
2.2.10 Functional annotation 64
2.2.11 High-performance computing (HPC) 65
2.3 Pipeline Optimisation 66
2.3.1 Ambient RNA correction 66
2.3.2 Inflation of Type 1 Error in differential Expression analysis 73
2.4 Conclusion 82
2.4.1 Limitations and Future Work 84
3 Single-nucleus RNA-sequencing of mouse models of Epilepsy 89
3.1 Introduction 89
3.1.1 Gene expression and signature reversion 90
3.1.2 Drug repurposing 91
3.1.3 Mouse models of epilepsy 91
3.1.4 Aims & Objectives 94
3.2 Methods 96
3.2.1 Experimental mouse models of Epilepsy 96
3.2.2 Tissue sectioning 97
3.2.3 Isolation of Nuclei 98
3.2.4 Data Generation 99
3.2.5 Mapping 100
3.2.6 Data Processing and Quality Control 100
3.2.7 Functional enrichment analysis 102

3.2.8 Enrichment of mendelian epilepsy genes 103
3.2.9 Association of heritability 103
3.2.10 CRISPR brain analysis 109
3.2.11 Enrichment of DEGs 109
3.2.12 Intersection with databases 111
3.2.13 Computational Tools 111
3.2.14 Code availability 112
3.3 Results 112
3.3.1 Study summary 112
3.3.2 Annotating cell-types 114
3.3.3 Identification of differentially expressed genes in epilepsy 116
3.3.4 Functional enrichment of cell-type specific DEGs 120
3.3.5 Gene list enrichment results 125
3.3.6 Integration of differentially expressed genes with genetic data 127
3.3.7 Perturbation analysis with CRISPRbrain 133
3.3.8 Intersections with interaction databases 138
3.4 Discussion 142
3.4.1 Limitations and Future Work 146
4 Single-nucleus RNA-sequencing of Human Epilepsy 149
4.1 Introduction 149
4.1.1 Epilepsy Surgery 150
4.1.2 Epigenetics 152
4.1.3 Aims & Objectives 154
4.2 Methods 156
4.2.1 Sample Selection 156
4.2.2 Tissue sectioning 156
4.2.3 Isolation of nuclei 157

4.2.4 Genotyping 157
4.2.5 Data Generation 158
4.2.6 Mapping 160
4.2.7 Data Processing and Quality Control 161
4.2.8 Weighted Nearest Neighbor (WNN) 163
4.2.9 Downstream Analysis 163
4.2.10 Genes4Epilepsy 164
4.2.11 Differential Accessibility 164
4.2.12 Computational Tools 165
4.2.13 Code availability 165
4.3 Results 166
4.3.1 Study summary 166
4.3.2 Annotating cell-types 167
4.3.3 Identification of differentially expressed genes in epilepsy 169
4.3.4 Functional enrichment of cell-type specific DEGs 174
4.3.4 Gene list enrichment results 178
4.3.5 Integration of differentially expressed genes with genetic data 180
4.3.7 Comparison to mouse data 185
4.3.8 Differential accessibility analysis 188
4.3.9 Intersection of DE and DA 191
4.3.10 CRISPR BRAIN results 192
4.3.11 Intersections with interaction databases 196
4.3.12 Comparison to mouse 200
4.4 Discussion 202
4.4.1 Limitations and Future Work 204

5 Identification of Dysplastic Neurons 208

5.1 Introduction **208**

5.1.1 FCD pathology 209

5.1.2 Dysplastic Neurones 210

5.1.3 mTOR pathway 211

5.1.4 Aims 213

5.2 Methods 214

5.2.1 Mouse model of FCD 214

5.2.2 BBmap 215

5.2.3 Human FCD samples 218

5.2.4 Downstream analysis 218

5.3 Results 219

5.3.1 Clustering of dysplastic neurons 219

5.3.2 Differential Expression of dysplastic neurons 227

5.4 Discussion 235

5.4.1 Limitations and Future Work 237

6 Discussion 238

6.1 Cell-types in Epilepsy **238**

6.2 Genetics of Epilepsy **241**

6.3 Alternative mechanisms **242**

6.4 Drug discovery **244**

6.5 Limitations **244**

6.6 Future Work **246**

References 249

Appendix 282

List of Figures

Chapter 1

1.1 ILAE Classification of epilepsy. 21

1.2 Cell types of the brain. 34

Chapter 2

2.1 Overview of the single-cell sequencing data processing pipeline. 43

2.2 Diagrams of the 10x Chromium Next GEM workflow. 45

2.3 Knee plot to show the ranked UMI per barcode. 48

2.4 Violin plots used for filtering. 52

2.5 UMAP to show the clustering of the samples using the specified parameters. 52

2.6 Functional annotation of Gene Ontology (GO) terms across DEGs in ambient RNA. . . .64

2.7 UMAPs showing the clustering of mouse samples before and after ambient RNA correction. 66

2.8 UMAPs showing the expression of the gene Syt1. 68

2.9 Measures of heterogeneity in single-cell data. 73

2.10 Scatter plots to show the correlation of Log2FC between methods of DE. 75

2.11 Sample size calculations. 77

Chapter 3

3.1 Brain regions used in mouse analysis. 89

3.2 Cell-type annotation of mouse data. 105

3.3 UMAP plots to visualisation cell variables. 106

3.4 PCA plots of mouse samples. 107

3.5 Numbers of DE genes in mouse data. 109

3.6 Functional enrichment of GO terms across DEGs in mouse study. 113

3.7 Enrichment of known epilepsy genes in mouse study. 117

3.8 Enrichment of epilepsy common SNP-heritability in mouse study. 120

3.9 Enrichment of epilepsy rare-variants in mouse study. 122

3.10 Significant results of CRISPRbrain analysis in mouse study. 128

Chapter 4

4.1 Diagram of histones. 141

4.2 Brain regions used in human study. 145

4.3 10x multiome experimental workflow. 147

4.4 Cell-type annotation of human analysis. 156

4.5 Numbers of DEGs in human analysis. 158

4.6 Overlapping genes between comparisons. 159

4.7 Functional enrichment of GO terms across DEGs in human study 160

4.8 Enrichment of known epilepsy genes in human study. 167

4.9 Enrichment of epilepsy common SNP-heritability in human study 170

4.10 Enrichment of epilepsy rare-variants in human study. 173

4.11 Overlap of DEGs between mouse and human comparisons. 175

4.12 Coverage plot of open chromatin of MTRNR2l1. 179

4.13 Coverage plot of open chromatin of MTRNR2l1. 180

4.14 Significant results of CRISPRbrain analysis in human data. 185

4.15 Human and human results of enrichments 192

Chapter 5

5.1 ILEA classification of FCD. 198

5.2 mTOR pathway. 201

5.3 Transfected Rheb sequences. 206

5.4 Clustering of all FCD samples. 209

5.5 Marker expression in mouse FCD samples. 211

5.6 Marker expression in human FCD samples. 212

5.7 Mutated Rheb expression in samples. 213

5.8 Distribution of transfected cells. 215

5.9 Functional enrichments of GO terms in marker genes in DNs. 217

5.10 Functional enrichments of GO terms in marker genes in DNs compared to all neurones. 220

5.11 Enrichment of epilepsy rare-variants in markers. 221

5.12 Overview of ML workflow. 223

List of Tables

Chapter 2

2.1 Numbers of DEGs from real cells vs empty droplets 63

2.2 Cell-type annotations before and after CellBender. 67

2.3 Numbers of DEGs before and after CellBender correction. 69

Chapter 3

3.1 Sample summary of mouse data. 105

3.2 Therapeutic strategy based on the outcome of CRISPR perturbation. 126

3.3 Intersections with IUPHARM database. 131

3.4 Intersections with PDE3 database. 132

Chapter 4

4.1 Sample summary of human data. 156

4.2 Numbers of intersecting DEGs across all comparisons. 162

4.3 Numbers of DA genes. 178

4.4 Therapeutic strategy based on the outcome of CRISPR perturbation. 182

4.4 Intersections with IUPHARM database. 188

4.6 Intersections with PDE3 database. 190

Chapter 5

5.1 Summary of mouse FCD samples. 204

5.2 Summary of human FCD samples. 207

5.3 Numbers of transfected neurons. 214

5.4 Top 15 genes DE in DNs. 216

5.5 Top 15 genes DE in DNs compared to all neurones. 219

Abbreviations

AED Anti-Epilepsy Drugs

ASD Autism Spectrum Disorder

ASM Anti-Seizure Medications

ATAC Assay for Transposase-Accessible Chromatin

BCL Base Call Files

BH Benjamini-Hochberg

CCA Canonical Correlation Analysis

DEE Developmental and Epileptic Encephalopathies

DEG Differentially Expressed Genes

EEG Electrocochleography

FCD Focal Cortical Dysplasia

FDR False Discovery rate

GEM Gel Bead-in-Emulsion

GGE Genetic Generalized Epilepsy

GO Gene Ontology

GSA Genome-wide SNP-based Pathway Analysis

GWAS Genome-Wide Association Studies

HS Hippocampal Sclerosis

ILAE The International League Against Epilepsy

iPSC induced Pluripotent Stem Cells

LD Linkage Disequilibrium

MAST Model-based Analysis of Single-cell Transcriptomics

MR Mendelian Randomization

MS Multiple Sclerosis

NAFE Non-acquired Focal Epilepsies

ORA Over Representation Analysis

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

PILO Pilocarpine induced post-status epileptic model

PTV Protein Truncating Variants

scRNA single-cell RNA-seq

SNP Single Nucleotide Polymorphisms

snRNA single-nucleus RNA-seq

SSSE Self Sustained Status Epilepticus

SUDEP Sudden Unexpected Death in Epilepsy

t-SNE t-distributed Stochastic Neighbor Embedding

TLE Temporal Lobe Epilepsy

UMAP Uniform Manifold Projection

UMI Unique Molecular Identifier

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

Chapter 1

Introduction

1.1 Epilepsy

Epilepsy is a common neurological disorder which is primarily characterized by recurrent seizures (1). Active epilepsy affects approximately 1 in 200 people, and more than 50 million people worldwide (2). It can occur at any age, with a higher incidence in younger (i.e. infants, children and young adults) and older (i.e. over 60) age groups, and it is also seen more frequently in low to middle income countries. A diagnosis of epilepsy can be life changing due to the disease's debilitating nature, and it is often accompanied by a range of comorbidities. Furthermore, patients with epilepsy are three times more likely to die prematurely than the general population. Beyond the physical and cognitive repercussions of seizures, patients also endure societal and psychological consequences, due to the prevailing stigma, as well as cognitive and behavioral complications. Despite the high disease burden, there is still a large unmet need for new epilepsy treatment, and 30% of patients with Epilepsy have seizures that cannot be controlled by medication (3–5).

1.1.1 Clinical symptoms and diagnosis

A person is defined as having epilepsy if they have two or more unprovoked seizures, which occur more than 24 hours apart, or a single seizure with a substantial risk of recurrence (1). The epileptic seizure is the primary symptom of epilepsy. Seizures can be highly disabling, and symptoms may include an abrupt abrupt, involuntary stiffening, followed by synchronous jerking and a loss of awareness as in the case of a generalized tonic clonic seizure. Focal

symptoms are usually observed as visual and auditory impairments (occipital and lateral temporal seizures respectively), accompanied by other symptoms such as cognitive disfunction and disruption in bladder and bowel control (2). The specific symptoms exhibited hinge on the origin of the abnormal neuronal activity within the brain. The International League Against Epilepsy (ILAE) classification framework is a tool that is frequently used in the diagnosis of individuals with seizures. According to this classification seizures can be categorized as being Focal, Generalised or Unknown. Focal seizures arise from a specific region or hemisphere of the brain, whereas Generalised seizures begin in a wide area over both hemispheres. In cases where the onset of a seizure cannot be determined, it falls under the "Unknown" classification (Figure 1.1) (6).

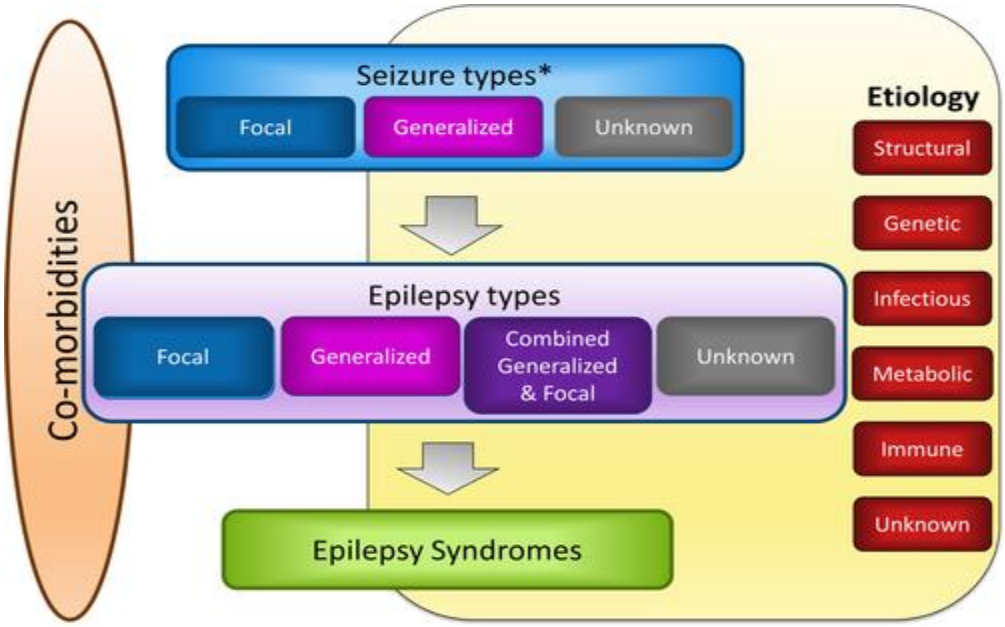


Figure 1.1 – Classification of Epileptic Seizures according to the International League Against Epilepsy (2017). This flowchart illustrates the categorization of seizures starting from their major types – Focal, Generalised and Unknown. These types can be further classified into more specific epilepsy types and epilepsy syndromes. Various aetiologies and comorbidities can be used to aid classification (Image Credit: Scheffer et al. 2017)(7).

As well as classifying the seizure type, the specific epilepsy syndrome can also be classified. Epilepsy can be more accurately described as a syndrome made up of many different epilepsy diseases due to the variation in the presentation of symptoms. The type of epilepsy can inform the treatment a patient receives. Diagnosis is based on several features such as seizure type, EEG, imaging, age of onset as well as age progression of the disease. Good classification of epilepsy syndromes can guide good treatment (6).

Co-morbidities of epilepsy include psychological problems such as anxiety, depression, learning difficulties and autism spectrum disorder (ASD), adding to the challenge of the disorder. Additionally, people with epilepsy have higher rates of injuries such as fractures and bruising from seizures, and epilepsy related deaths are often caused by indirect effects of seizures such as falls, drownings, adverse effects of anti-epilepsy drugs (AEDs), psychiatric drugs or aspiration pneumonia, all which could be prevented. Death in epilepsy can also be caused due to the direct effects of seizures such as in sudden unexpected death in epilepsy (SUDEP)(6,8,9).

Epilepsy can be genetic or acquired. Historically, epilepsies that are caused by largely genetic factors have been termed ‘idiopathic’, and epilepsies that are caused by acquired factors ‘symptomatic’. Symptomatic epilepsies can be acquired through infection or injury, such as traumatic brain injury, tumour, or stroke. This is likely why the rate of incidence is higher in low-income countries where the quality of underlying infrastructure may result in more infection, birth injury and traffic accidents, and where resources are scarce. Many acquired cases of epilepsy are preventable (25%). Other risk factors for epilepsy include cerebrovascular disease, tumour, alcohol abuse and developmental malformations (10,11).

1.1.2 Pathogenesis and aetiology

Seizures are caused by abnormal and uncontrollable electrical activity in the cortex which can be focal or generalized. Abnormal electrical activity involves the hypersynchronous firing of neurons and neuronal networks. An early view was that epilepsy is caused by an imbalance in the inhibitory and excitatory neurotransmission through the GABA and glutamate receptors, but this is now thought to be an over simplification (12,13). The aetiology of epilepsy is likely made up of a complex interaction between genetics and the environment, and the progression can be influenced by several factors including genetics, environment, brain structure, and the response to treatment. Some patients experience progressive deterioration of the condition while others experience steady seizures throughout their lifetimes (13,14). Understanding the pathogenesis and progression of epilepsy often involves studying the initiation of seizures and epileptogenesis, this is the process in which seizures develop from previously normal brain tissue. This involves studying the molecular, structural, cellular and circuit level alteration that lead to unprovoked seizures, usually initiated by a pathogenic event or a genetic alteration (6,15). Understanding the pathogenic alterations that occur during the process of epileptogenesis, and its progression is important in helping to find therapeutic interventions (16). There has been considerable progress in understanding epilepsy, and it has been shown that many of the genetic mutations involved in epilepsy are involved in synaptic transmission or ion channels. However due to the heterogeneity of the disease, the mechanisms involved in causing seizures are still poorly understood. A better understanding of the early stages of disease may help to find better interventions. There is also emerging evidence that implicates immune system dysregulation and inflammation in epilepsy pathogenesis (17). This can include autoimmune epilepsy associated with antibodies directed against the neuronal cell membrane (e.g., LGI1, CASPR2 etc) or cytoplasm (e.g.,

GAD65), as well as a role for immune dysregulation in the process of epileptogenesis following brain injury (18,19).

1.1.3 Genetics of Epilepsy

Epilepsy is a highly heritable condition with up to 70% of cases arising through no cause other than an inherited predisposition. This has been established both through traditional twin-based studies and also more recent genome-wide heritability analyses (20). Twin and family studies have shown that the heritability of epilepsy ranges between 35-70% depending on the type of epilepsy. Even epilepsies that have been acquired are thought to have some genetic factors, as shown by family history data (11,21). The heterogeneity of epilepsies can be attributed to the underlying genetic architecture. These can be Mendelian (or monogenic) in which a single gene (with a large effect size) is responsible for the inheritance of the gene, or complex, where several loci interact together and with environmental factors in order to cause the disease (22). Mendelian (or monogenic) epilepsies are rare are often begin in childhood, and these are predominantly channelopathies. CHRNA4, the first gene to be linked to epilepsy, was discovered in 1995 through linkage studies in affected families. It encodes a subunit of the acetylcholine receptor (nAChR) which is a ligand-gated ion channel (23,24). Since then, mutations have also been identified in several voltage-gated ion channels, in particular K+ and Na+ channels. One of the most common epilepsy genes, SCN1A which encodes for a neuronal sodium channel subunit is responsible for several epilepsies. The majority of SCN1A gene mutations cause Dravet syndrome, a severe epileptic encephalopathy which is resistant to medication (25). While monogenic epilepsies are rare individually, they account for a high proportion of epilepsy cases. Studying these forms of epilepsy has improved our understanding of the mechanisms involved in the disease by demonstrating how neuronal activity, ion channel function, and synaptic transmission are

disrupted to contribute to seizure generation. Technological advancements have also meant that genetic testing has become a part of epilepsy diagnosis and clinical practice in early onset forms of epilepsy or specific epilepsy syndromes such as Dravet, where causal mutations are known and well characterized (26).

As technology has advanced, our knowledge of the genes involved in epilepsy has widened. Targeted approaches such as whole exome sequencing (WES) and whole genome sequencing (WGS) of large cohorts have been used to detect high susceptibility variants in the general population. These have helped to identify de novo epilepsy genes particularly, meaning they occurred in the gametes or developing embryo. These account for some of the most severe forms of epilepsy (27). However, these de novo genes are rarely inherited, and do not explain the majority of epilepsy heritability (28). While rare variants have been shown to contribute to both generalized and focal epilepsy, large proportion of epilepsy heritability remains unexplained and cannot be accounted for by known genetic variants.

The majority of epilepsies are likely driven by more complex inheritance and remain poorly understood. Complex traits are believed to have multigenic susceptibility, mediated by common genetic variation such as in single nucleotide polymorphisms (SNPs). It is through that the additive effect of multiple small effect variants is what contributes to most family risk for epilepsy (28). This may also involve interactions between multiple genes as well as interactions between genetic predisposing and environmental factors. The genetic architecture of epilepsy is complicated with different types of epilepsy syndromes likely having distinct genetic underpinnings. This means the identification of complex genetic factors has been much less successful (22). The advent of genome-wide association studies (GWAS) have allowed for the discovery of common risk factors for epilepsy, as these allow for an unbiased

approach without the need to select candidate genes. GWAS are a type of study designed to identify common genetic variants (defined as variants with a population frequency > 1% (29) that are associated with complex diseases and phenotypes. GWAS typically involves testing the frequency of a variant across the entire genome between two populations of matched cases and controls where the cases have the trait of interest and the controls do not. Millions of SNPs are compared between groups to find variants that are more significantly present in cases. As disease effect sizes for common variants (SNPs) are usually small, GWAS usually require very large sample numbers and careful phenotyping. Due to the large number of statistical tests (one per SNP), the genome wide significance threshold must account for multiple comparisons to minimize type 1 error (false positive results). The larger the sample number, the better the ability to detect significant associations (30). A critical issue in the GWAS statistical toolbox is the avoidance of population stratification which can also introduce false positive disease associations due to genetic mismatches arising from systematic population or ethnicity differences between disease cases and controls.

The international against epilepsy (ILAE) consortium on complex epilepsies have led several large GWAS studies (31,32), the largest and most recent of which was published in 2022 identified 26 common epilepsy risk loci, 16 of which had not been previously identified. The GWAS consists of 29,944 cases and 52,538 controls, which is a great increase in sample size since the first ILAE epilepsy GWAS in 2014 with 8,696 cases and 26,157 controls. This GWAS compared Genetic Generalized Epilepsy (GGE), and Focal epilepsy independently as well as all epilepsies. The results of this GWAS showed different genetic architecture between focal and generalized forms of epilepsy. Some genes such SCN1A, have been implicated in through GWAS to be contributing to genetic susceptibility of common epilepsy as well as monogenic (33). GWAS analysis can be summarized into GWAS summary

statistics which can more readily used in downstream analysis. As well as identifying risk loci, GWAS summary statistics are also a powerful tool which allow for secondary downstream analysis methods such as LD (Linkage Disequilibrium) score regression, mendelian randomization (MR) , and colocalization (34).

Substantial advancement in technology and methodology has allowed for the identification of >500 genes associated with epilepsy and seizures, and ongoing research efforts continue to uncover more genes (35). Despite the discovery of a large number of epilepsy genes, our current understanding of the genetic architecture of epilepsy is still poor, with the more common forms thought to be polygenic with contributions from both common and rare variants. It is important to consider the contribution of both rare and common epilepsy variants to find convergent mechanism which contribute to the underlying biology (36).

A better understanding of the genetic architecture of epilepsy including a more complete picture of all the genetic risk factors would likely lead to better treatment choices through more personalized approaches, and targeted therapies as well as improving diagnosis and prevention strategies.

1.1.4 Treatment and Drugs

Many patients with epilepsy are able to live seizure free with the appropriate anti-seizure medication. Anti-Seizure Medications (ASMs) are usually the first line of treatment for patients with epilepsy, but as well as drugs epilepsy treatment may also include surgery, dietary control, or neurostimulation devices. The type of treatment prescribed is dependent on the type of epilepsy, the severity and frequency of seizures and the individuals overall health. ASMs work by controlling the symptoms of epilepsy and suppressing seizures, generally by stabilizing excitability. These aim to improve the quality of life of patients and minimise adverse side effects of the disease, however they are unable to prevent or cure the disease.

Traditional ASM discovery focuses on a limited number of targets that do not encapsulate the biological complexity of the disease (37–39). These usually work by stabilizing the abnormal electrical impulses in the brain, by targeting ion channels and modulating the flow of ions or synaptic neurotransmitter release, and do not target the underlying pathogenic mechanisms. Additionally, there are many ASMs where the mechanism of action is still not known, even after decades of clinical use. This means important aspects of the disease might not be fully considered. Often a combination of ASMs is needed to adequately control seizures (40,41). Despite the number of drugs available to surpass seizures in epilepsy, 1 in 3 patients have seizures that are resistant to treatment. This means that medications are unable to keep seizures under control. Drug-resistant epilepsy, also called refractory epilepsy imposes a substantial disease burden and highlights an urgent need for the development of novel epilepsy drugs. In order to advance the development of effective treatment, a broader range of therapeutic targets is needed that better encapsulates more aspects of the disease. The considerable genetic and phenotypic heterogeneity associated with epilepsy, means the pathophysiology of epilepsy is still insufficiently understood. This makes discovering new drug targets that are contributing to the disease, a highly challenging process. This gap in the knowledge that makes it difficult to develop suitable treatments that can effectively modify the disease without disrupting normal brain function. By expanding our understanding of the molecular pathways and interactions that cause epilepsy, we can find more targeted and transformative therapeutic approaches (41,42).

There have been considerable efforts to explain the drug resistant phenotype in epilepsy, however these insights have generally not been translatable to the clinic. Translating biological insights into clinical practice requires a good understanding of disease-causing

genes, variants and the cell-types through which these are acting, as well as a better understanding of mechanisms and pathways (43).

The usual process of drug discovery involves many stages. The first step is to identify a target – this is a molecule that plays an important role in reducing seizures or modulating disease. Examples of common drug targets are ion channels, receptors, and enzymes. After a target has been identified, a chemical compound must be identified that can modify the activity of the target. This is often done thought high throughput screening (HTS), using in-vitro assays, animal models or virtual computer-based methods (44). The chosen compounds are then optimized for their pharmokinetic and safety profiles, to look out for any potential side effects and improve potency. These compounds then undergo preclinical and clinical testing to assess safety, and efficacy. This can involve testing the candidate drugs in animal models to evaluate their ability to prevent seizures, to assess their dosage requirements, and uncover potential adverse effects. The clinical trials involve testing the compound in humans. They consist of Phase 1, Phase 2 and Phase 3 trials. These are assessed, in a staged manner, for safety, dosage, pharmacokinetics, and potential side effects. If the compound is successful in Phase 3, it can then gain regulatory approval. This process is extremely long and expensive and it can take many years for a drug to be approved. Very few drugs make it all the way from preclinical proof-of-concept to regulatory approval making individual drugs generally including ASMs very expensive to develop (45).

1.1.5 Subtypes of Epilepsy

This thesis will focus on two types of epilepsy: Temporal Lobe Epilepsy (TLE) and Focal Cortical Dysplasia (FCD).

1.1.5.1 Temporal Lobe Epilepsy (TLE)

TLE is the most common form of focal epilepsy and affects 1/5th patients with epilepsy. TLE arising from the amygdala and/or hippocampus is known as mesial temporal lobe epilepsy (mTLE) and when it is associated with atrophy and scarring of the mesial temporal structures it is referred to as mTLE with hippocampus sclerosis (MTLE-HS). MTLE-HS is also the most common form of drug refractory epilepsy and the most frequent pathological abnormality seen in focal drug refractory epilepsy. The structural and neuronal loss of neurons in the hippocampus seen in MTLE-HS, is the result of a shrunken hippocampus and gliosis. The exact cause of MTLE-HS is poorly understood, and is likely heterogeneous. Approximately one half of cases will have a prior history of complex febrile seizures in childhood, suggesting that for some there is an initiating insult leading to long-term changes that eventually manifest as adult TLE (46). While TLE is considered a focal epilepsy as it originates in a specific area within the brain, some individuals experience secondary generalisation of seizures which means they can spread to involve the whole brain (so called focal to generalized tonic clonic attacks) (47). TLE can occur as a result of many factors including genetic predisposition, brain injury, infection and neurological conditions.

1.1.5.2 Focal Cortical Dysplasia (FCD)

Focal cortical dysplasia is a developmental disorder characterized by structural malformations in the cortex. These malformations disrupt normal brain function and cause seizures. It is one of the most common causes of refractory epilepsy, particularly in children. The underlying molecular mechanism of FCD were poorly understood until relatively recently, where it has been linked to dysregulation in the mTOR pathway (48). A key pathological feature of FCD is dysplastic neurons. These are neurons which undergo abnormal development and organization in a specific region of the brain, typically the cortex. These neurones do not migrate properly during development and are found in disorganized

locations. They also show abnormal morphology and functional abnormalities which can contribute to seizure development. They may also show altered gene expression signatures however there is ongoing research in the field to uncover the mechanisms in which they develop and how they contribute to progression of FCD (49).

1.1.6 Modelling Epilepsy

Mouse models have been a key part of epilepsy research by advancing our understanding of the disease and by facilitating the development of new treatment, through providing insight into the mechanisms that underpin epilepsy. While they may not fully replicate the complexity of the disease, the seizures that mice exhibit are very similar to human seizures, and the pathways and genes involved are highly conserved (50). Mice are particularly advantageous to genetic studies as they are easy to manipulate genetically and allow for standardised experiments. They provide a valuable platform for assessing the effects of interventions on seizure frequency, duration, and severity. Additionally, they provide a tool to study molecular-level mechanisms, neuronal excitability, and abnormal network activity to shed light on epilepsy pathophysiology. Several methods can be employed to induce seizures in mice, including chemical convulsant, electrical stimulation, or genetic manipulation. They provide opportunities to study changes in neuronal excitability, aberrant network activity due to similar seizure types and duration. Similarly, to humans, they can experience generalised and focal seizures which induce similar symptoms and behaviour such as loss of consciousness and muscle contractions. The seizure presentation can also be manipulated by researchers allowing for study of a variety of different types of epilepsy (51–53). They also allow for easy monitoring through electrocochleography (EEG) and brain imaging to observe how seizures affect electrical activity and brain structure. Finally, mice respond to many anti-seizure drugs in a similar way to humans which allows for testing of the efficacy of drug

targets. Not all aspects of human epilepsy can be recapitulated in mice due to differences in brain structure and complexity. Mice models do not account for the complex genetic backgrounds of humans, and the complexity of seizures seen in humans is far greater. It is also challenging to observe the range of symptoms seen in humans, particularly those related to cognition and memory. Mouse models are also unlikely to account for the environmental factors that may play a role in human epilepsy. Overall, they have been instrumental tools in preclinical experiments (52,54).

1.2 Transcriptomics in epilepsy

The human brain is very complex and is made up of over 100 billion cells that are very intricately connected. These cells are highly specialized, diverse, and have distinct morphological and functional characteristics. Variation and diversity in the brain arises from distinct mechanisms that operate within each cell through gene expression, regulation and epigenetic modification of genes (55). The transcriptome encompasses the full complement of the transcripts present in a cell and offers a valuable insight to the molecular environment of a cell. Diseases arise when the complex interplay between genetics and the environment is disrupted. In disease, transcriptomics can provide a snapshot of the molecular landscape by showing the actively expressed genes and their abundance and gene expression is a good proxy for disease as it can be informative of the underlying processes leading to disease pathogenesis. This can facilitate the investigation of the underlying processes which can guide the identification targets in drug discovery (56). Transcriptomics has played a key role in epilepsy research by helping to identify candidate genes that are associated with epilepsy. It has also advanced our understanding of the molecular mechanisms that underlie epilepsy, by identifying dysregulated pathways and processes involved in epilepsy and epileptogenesis

(36,57–59). It has also been a useful tool in identifying biomarkers for diagnosis that correlate with disease severity.

1.2.1 Single-cell RNA-sequencing

Measuring gene expression has advanced rapidly over time. From using northern blots to assess the expression of a single gene, to more high throughput methods like DNA microarray and bulk RNA sequencing, where thousands of genes or even the whole genome can be simultaneously investigated. The more recent development of single-cell sequencing (scRNA-sequencing) technology has revolutionized our understanding of gene expression by providing the opportunity to look at the transcriptomic profile of each individual cell, highlighting differences in cellular populations that were previously undetected by bulk RNA-sequencing. Unlike bulk sequencing where gene expression is measured across multiple cells, scRNA-sequencing allows analysis of expression levels for each gene across a population of individual cells. This allows us to look at heterogonous cell populations as well as small and rare cell populations which could go undetected in analysis of pooled cells. It also means subtle but meaningful changes in the same cell-type can be detected (60,61). Dissecting the intricate cellular landscape of the brain at a single-cell resolution, allows for better characterize the identification of the specific cell-types of the brain and their transcriptional profiles, as well as possible disease associated sub populations (Figure 1.2) (61–63).

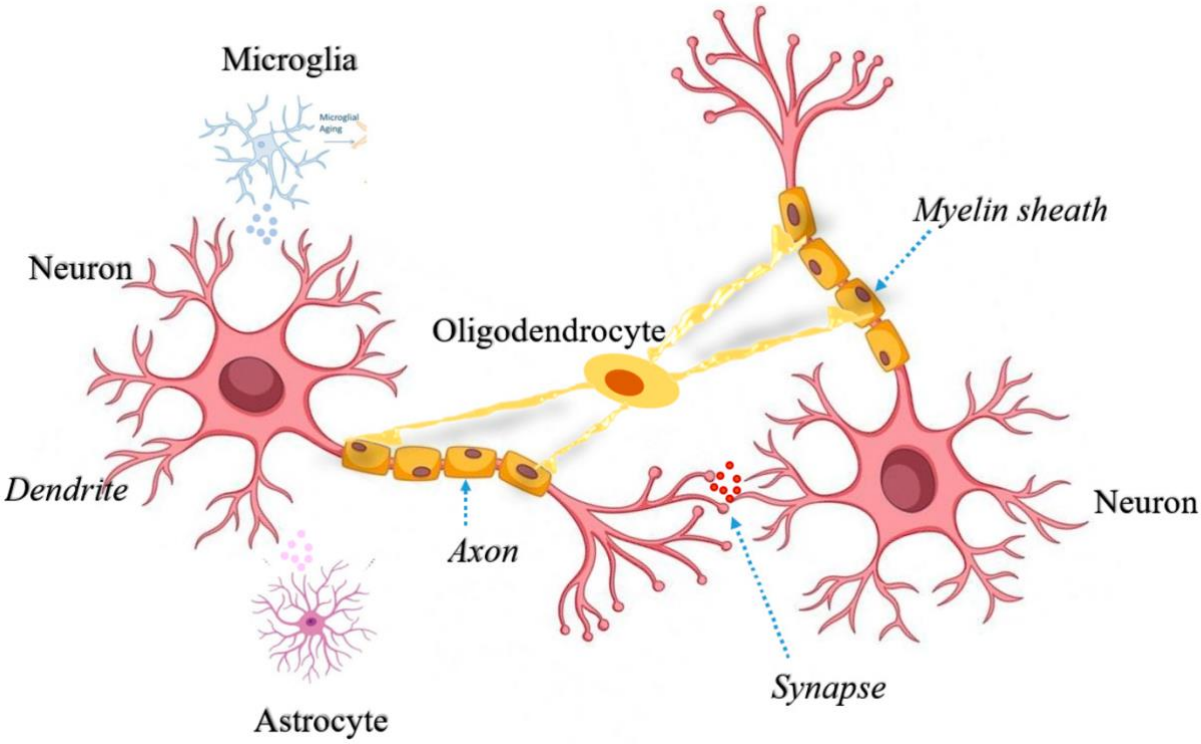


Figure 1.2 – Cellular composition of the brain. This figure depicts the key cell-types present in the brain including Neurons, Microglia, Oligodendrocytes and Astrocytes. This highlights the intricate connections between neurons and glial cell-types of the brain and the complex interactions between them. It also emphasises the diversity of cell-types in the brain and the need to study transcriptomics at a single-cell level (Image Credit: Ragupathy et al.,2023)(64).

A typical scRNA-sequencing workflow consists of single-cell dissociation, single-cell isolation, library construction and sequencing. In the first step, tissue is digested creating a single-cell suspension. The cells are then isolated and separately encapsulated within a microfluid droplet, along with the reagents required to construct a library. This is the process in which mRNA is captured, reverse transcribed and amplified. The droplet includes a unique barcode that will label the transcripts belonging to that cell. Reads can be pooled from multiple samples and traced back to the samples they came from, while also assigning each

mRNA transcript Unique Molecular Identifier (UMI) which allows us to distinguish between amplified reads and the absolute transcript count. Brain tissue can present challenges in single-cell sequencing as it is often supplied frozen and is not compatible with the methods, but neurons are also sensitive cells with long axons that can be easily damaged during dissociation, generating an artificial stress response. Some of these issues can be minimized by using single-nucleus RNA sequencing, as single nuclei are much easier to isolate and are more resistant to the mechanical stress. snRNA-sequencing data detects comparable numbers of transcripts as to scRNA-sequencing (98%) and gene expression values are highly correlated. The average variability between datasets is also similar. Therefore, this approach is often used instead of scRNA-sequencing. The terms single-cell and single-nucleus sequencing are often used interchangeably (63,65–68).

There are multiple platforms that are used for snRNA-sequencing, this project uses 10x 3' RNA-sequencing. This platform is a widely used technology for profiling gene expression in individual cells. The platform utilizes a unique approach called "GEMs" (Gel Bead-in-Emulsion) technology to capture and barcode individual cells. The 10x single-cell RNA-sequencing platform utilizes a 3' library preparation method, which refers to the capture and sequencing of the 3' end of each RNA molecule within the individual cells. A unique molecular barcode is added to the 3' end of each cDNA molecule during the reverse transcription step. This barcode serves as an identifier for each original RNA molecule and allows for the quantification of gene expression levels. Additionally, a unique molecular identifier (UMI) is also incorporated into each cDNA molecule to enable the removal of PCR duplicates and accurate counting of individual RNA molecules (68).

Single-cell analysis has undergone significant advancements, transitioning from a time-consuming and labour-intensive process to a faster and more efficient method, which allows

for higher throughput, enabling the profiling of a large number of individual cells in a cost-effective manner. One of the benefits of scRNA-sequencing is that it is a scalable approach meaning thousands of gene expression values in hundreds of thousands of cells can be measured, from as many individuals. Significant advancements have been made in our understanding of brain disease and function due to the advancements in scRNA-sequencing. Some notable examples of key discoveries that can be attributed to single-cell sequencing include the discover of disease associated microglia in Alzheimer's disease (69) and the characterisation of cell-types and sub types in the brain (70). As well as other discoveries in cancer and immune response such as the role of T-cells in cancer (71). In the brain it has advanced our knowledge of molecular phenotypes of neurons, as well as glial cells such as microglia, astrocytes, oligodendrocytes, and how they contribute to disease.

1.2.2 Cell-types in epilepsy

Seizures, characterized by abnormal and excessive neuronal activity, are believed to arise from hyperexcitability in the brain. This hyperexcitability can stem from various factors, including dysfunctions in ion channels that regulate neuronal firing, imbalances in neurotransmitter levels, synaptic abnormalities, and disrupted signalling pathways. While research on epilepsy has predominantly focused on neurons, the involvement of glial cell-types remains poorly understood, particularly at the transcriptional level. Given the intricate network of interconnected cell-types within the brain, it is highly likely that the development and maintenance of epilepsy (epileptogenesis) and the generation of recurrent seizures (ictogenesis) involves the contribution of multiple cell-types (72). Understanding the heterogeneity and functional roles of different cell-types is crucial for unravelling the underlying mechanisms of epilepsy and identifying potential therapeutic targets.

The role of glial cells, such as astrocytes, oligodendrocytes, and microglia, is increasingly recognized in modulating neuronal function and maintaining brain homeostasis. These glial cells are likely to contribute to epileptogenesis through their involvement in processes such as inflammatory responses, synaptic transmission, and regulation of ion concentrations, all which could be influencing epileptic activity. Astrocytes, for example, exhibit increased expression of glial fibrillary acidic protein (GFAP) and adapt to handle extracellular potassium overload. They can sense and respond to alterations in the extracellular microenvironment through signalling via neurotransmitter receptors. This ability allows them to modulate the immune response under pathological conditions associated with astrocytic gliosis (73). Microglia, on the other hand, have phagocytic functions and play a role in the regulation of inflammatory and immune responses. They survey the microenvironment for changes in signalling, pathogens, and injury, and can modulate synapse formation, elimination, and function. In the context of various neurological diseases, including epilepsy, microglia can be targeted by inflammatory molecules and cytokines (74).

The complex interactions between neurons, glial cells, and other cell-types with their distinct gene expression profiles and functions need to be comprehensively understood to decipher the mechanisms underlying epilepsy and identify potential therapeutic interventions (75,76).

1.3 Aims

Despite recent advances in our understanding of the genetic contributions to epilepsy, and past work on bulk brain tissue gene expression analyses, there are still significant gaps in our knowledge of the molecular and cellular processes underpinning chronic focal epilepsy. Yet, understanding the disease is crucial for the development of novel, mechanistically focussed disease-modifying therapies. The overarching aim of this thesis is therefore to increase the understanding of cell-types and pathways involved in epilepsy as well as highlight potential

gene candidates for drug targeting. By using single-cell sequencing of a variety of brain tissue samples from both human epilepsy and mouse models of epilepsy from different brain regions and across different types of epilepsy, this study aims to find commonalities and differences between species and epilepsy types that might point to convergent mechanisms of epilepsy that could be targeted for new therapy development.

The underlying hypothesis of this work therefore is that different focal epilepsies share convergent genes, pathways and/or cell-types. Through the analysis of cellular transcriptomes from snRNA-sequencing of epileptic mice and human epilepsy patients, this thesis aims to identify the convergent mechanisms that are perturbed in epilepsy. By integrating genetic data, the aim is to better elucidate disease-causing cell genes and pathways, and the cell-types through which they are acting. This increased understanding can help to facilitate the development of new therapeutic strategies that can improve the lives of people with epilepsy.

This project builds upon EpiNET1 and EpiNET2 (18,36), previous collaborations with UCB, that have focused on identifying gene expression signatures from bulk brain tissue samples derived primarily from the pilocarpine mouse model of TLE.. In the previous work, gene regulatory networks were built based on known interactions between transcription factors and their target genes and using co-expression analyses. These analyses identified sets of co-expressed genes that were largely cell-type specific and were used to optimize drug target identification using a framework that searched for common points of genetic regulation for the co-expressed gene set. Here, this work acts as an extension to these analyses by directly identifying cell-type specific gene expression profiles using snRNA-sequencing data and integrate these transcriptional profiles with genetic risk data to identify disease-context specific molecular pathways and cell-types that might be targeted therapeutically. This study

will use a broad range of bioinformatic tools and data-driven methodology to drive the identification of cell-type specific gene expression signatures and pathways involved in epilepsy.

1.3.1 Thesis outline

The specific objectives of this thesis are:

1. Optimise a developed pipeline for the analysis of large-scale snRNA-sequencing experiments, using a robust framework to efficiently process and analyse large amounts of data. This pipeline will enable the identification of meaningful insights from the transcriptomic profiles of individual cells. **(Chapter 2)**
2. Using snRNA-sequencing data from 3 mouse models of epilepsy, identify cell-type and disease specific differential expression signatures and pathways in epilepsy. Integrate this with genetic data to attribute epilepsy risk to specific cell-types. Use this knowledge to predict cell-type specific and disease context specific drug targets for disease modification in epilepsy. **(Chapter 3)**
3. Integrate results from mouse models of epilepsy with data from human patients with TLE and FCD to identify shared molecular signatures and disease mechanisms that converge in mouse and human epilepsy. Combine epigenetic data to provide insight into the regulation of the gene expression signatures and pathways. **(Chapter 4)**
4. Characterize the transcriptomic signatures of dysplastic neurones in focal cortical dysplasia in mice in order to find markers that can be applied to human datasets. **(Chapter 5)**

1.3.2 Limitations and Future Work

The focus of this thesis has primarily been on identifying converging genes, pathways, and cell-types with the rationale of uncovering shared mechanisms to increase the understanding of epilepsy. Finding common pathways and processes across different epilepsies could be an important step in enhancing our understanding of the fundamental biology of seizure generation and propagation which could lead to the identification of molecular targets involved in multiple types of epilepsy.

While drug-resistant epilepsies are thought to be highly heterogeneous (77), they could also potentially share underlying downstream molecular mechanisms. Several studies have demonstrated the overlap between the genetics of multiple types of epilepsy, such as mutations in ion channel genes like SCN1A (24,25), therefore it is reasonable to assume the existence of common underlying mechanisms. All epilepsies involve abnormal neuronal excitability and network dysfunction, suggesting that common pathways affecting synaptic transmission, plasticity, and neuronal connectivity may underlie different types (78). Similar environmental triggers might also lead to the development of epilepsy in the same way. Processes such as neuroinflammation, oxidative stress, and blood-brain barrier disruption have been implicated in various forms of epilepsy, contributing to neuronal damage and hyperexcitability (79–81). The efficacy of many antiepileptic drugs across multiple epilepsies further indicates that they target common mechanisms in seizures (40). Understanding these shared mechanisms can aid in developing broad-spectrum treatments, improving diagnostic accuracy, and advancing our overall knowledge of epilepsy as a complex neurological disorder.

However, one of the main challenges associated with epilepsy is the heterogeneity of the disease, related to seizure types, causes, and neurophysiology (34). Epilepsy is often referred

to as a collection of diseases, and an alternate hypothesis could be that distinct epilepsies arise from divergent molecular pathways given the wide range of genes that are known to play a role in epilepsy. Due to variations in etiology, onset, and disease progression, it is also likely that different types of epilepsy have distinct mechanisms. The complex inheritance patterns, involving multiple genes and gene-environment interactions, also adds another layer of complexity. This means the brain may be affected in different ways, leading to varied mechanisms of seizure generation.

Additionally, the time of onset of the disease or the time point at which the disease is studied could be a significant factor in the underlying mechanisms involved. Epilepsies that begin in childhood, such as focal cortical dysplasia (FCD), often have mechanisms rooted in genetic or developmental factors (82), whereas those that start in adulthood might be more related to neurodegenerative processes (10). Early-stage disease in patients might have disruptions in different pathways compared to later stages and chronic epilepsy where the brain may have undergone significant changes due to prolonged seizure activity. Nevertheless, there could still be convergence in pathways that drive the progression and maintenance of the epileptic state. While the time points and models used in this study are limited by tissue availability, this consideration should be addressed in future work and interpretation.

scRNAseq is a powerful tool allowing us to explore cellular heterogeneity of disease at a high resolution by profiling individual cells and uncovering specific cell populations within tissues avoiding the drawbacks of traditional sequencing which detects the average transcript level of populations (68). It allows us to study genes and their functional alterations at the single cell level and has played a significant role in uncovering cell biology in various fields such as, cancer, immunology, and development (69,71,83). This granularity is crucial in a

heterogeneous and complex condition like epilepsy, where different neuronal and glial cell populations may contribute uniquely to the disease pathology. However, there are still some critical limitations that should be addressed when using scRNAseq.

scRNA-seq generates high-dimensional data that contains a significant amount of noise which can impact the reliability of the data. Due to the low amount of input material, there is a limited sensitivity to detect low abundance RNA and low efficiency means only 10-40% of the transcriptome is detected (84). Additionally, the RNA concentration also affects the results, with low RNA yield potentially leading to incomplete or biased gene expression profiles limiting ability to study the genes expressed at various levels (85). scRNAseq can only detect actively transcribed genes, which means that time points and the levels of neuronal activity can influence the results. This limitation is particularly relevant in the study of epilepsy, where capturing the dynamic nature of gene expression during different stages of disease progression or seizure generation would be limited. Neuronal excitability and networks changes are dynamic processes which would benefit from the ability to capture changes in gene expression over time. The single-cell capture process also often disrupts cells. Particularly in tissues like the brain which are difficult to dissociate, the preparation of single nuclei can disrupt cellular integrity and alter cell activity, potentially introducing artifacts that do not reflect the true in vivo state of the cells and adding to the noise (86).

Additionally, in this study we use 3’ based sequencing protocols which can lead to bias where only the 3’ end of the transcripts are sequenced so we may miss information about the full length of the molecule(87). Important regions such as the 5' end, coding sequence, and potential splice variants may not be fully captured and you can’t consider alternative splicing and isoform information which may be helpful in understanding gene function. As

sequencing relies on the presence of poly(A) tails, non-polyadenylated RNAs (such as some histone mRNAs and certain non-coding RNAs) may be underrepresented or missed. This is relevant to epilepsy as isoform alternative splicing and isoform diversity play critical roles. Furthermore, as single-nuclei RNA-seq is used in this study as an alternative to scRNAseq, we may miss cytoplasmic RNAs, which are crucial for understanding neuronal function and disease mechanisms. This can lead to an incomplete picture of gene expression, especially for genes with significant post-transcriptional regulation or those predominantly expressed in the cytoplasm. In the context of epilepsy, mitochondrial genes play many roles in various cellular processes, including oxidative stress regulation and calcium homeostasis. Disruptions in mitochondrial genes could significantly contribute to the pathophysiology of epilepsy. Therefor it would be useful to also consider the interplay between mitochondrial and nuclear genes in understanding and addressing epileptic disorders (88).

Finally, in scRNA-seq the spatial context of a single cell is lost, this can result in a significant loss of important information about the microenvironment and cell-cell interactions. This loss of spatial context can limit our understanding of how cells function within their native environment, particularly in epilepsy, where the organization and interaction of neurons and glial cells within specific brain regions are critical. Therefore, while scRNA-seq provides valuable insights into the molecular characteristics of individual cells, it is essential to consider these limitations and complement scRNA-seq studies with additional experimental approaches.

Epilepsy is a disease characterized by electrophysiological signatures and network disruption. These electrical impulses, and spikes that are characteristic of seizure activity and are usually detected by techniques such as EEGs. While this cannot be measured with transcriptomic

data, it may give insight into the processes which regulate it (89). Integrating scRNAseq data with electrophysiological recordings could also help correlate gene expression profiles with functional outcomes, offering an insight into the gene expression signatures associated with specific electrophysiological states, and a view of the disease mechanisms at both the cellular and network levels. Correlations could help look at specific gene expression patterns in relation to rates, synaptic strength and offer a better understanding of how genetic and molecular changes manifest. Seizure frequency data was not available for these mice but in future should be used. Despite these challenges, scRNAseq can still be useful in reveal cell-type-specific changes in gene expression that underlie the abnormal neuronal excitability and mapping these molecular changes provides insights into the cellular basis of electrophysiological abnormalities.

Chapter 2

Methods and Pipeline Optimisation

2.1 Introduction

Single-cell sequencing (scRNA-sequencing) is a rapidly evolving tool that has become an increasingly critical in the study of disease. The complexity of biological tissue arises due to the unique and diverse cellular landscape within it, and assaying gene expression at a single-cell level allows for the investigation of heterogeneity between cells in a wide range of tissues and diseases (68). The decreasing cost of single-cell sequencing since its inception, has allowed this technology to become more widely accessible and frequently used. To date there are over 1,000 single-cell transcriptomic datasets published (90). Despite the rapid progress in technology, there are elements of single-cell studies, particularly in the analysis of the data, that remain challenging (91).

Properly processing and interpreting the data from single-cell experiments is vital. Given the sheer volume and granularity of the data generated by scRNA-sequencing, well developed workflows are essential for managing it. Since the surge in popularity, numerous bioinformatics tools have emerged that are specialised for different parts of the workflow. For this reason, it can be challenging to decide on the most appropriate tool, and the choice often relies on the dataset at hand. These pipelines encompass various steps from quality control, alignment, and normalization to more advanced analyses such as clustering, differential expression analysis, and trajectory inference. There are many challenges associated with the analysis of single-cell data, particularly in the processing of the datasets and application of the downstream analysis. As the best methods to use have not been universally decided across the field, and the this can be dependent on the tissue biology. Some key challenges

include; technical variability, sequencing depth, batch effect, dataset size and complexity (65).

2.1.1 Computational Analysis

As single-cell assays become more readily available, the effectiveness of their use is highly dependent on the ability to interpret and extract information from the data generated. The quality of computational and statistical methodologies plays an important role in extracting meaning from the data. As the field grows, and the number of tools available to analyse single-cell data grows. There are currently over 1,400 tools available to analyse single-cell data, these are catalogued in the scRNA-tools database (92,93) . For this reason, it has become increasingly difficult to navigate the single-cell analysis landscape and find the most suitable tools for a given study. While there have been efforts to benchmark some analysis such as clustering and integration (94–96), these methods can often be dataset specific and not perform consistently. Given the expanding scale of transcriptomic dataset, it is important to use workflows that ensure efficiency, reliability, and reproducibility. These pipelines should be applicable to multiple datasets and replicable with ease.

While computational methods can differ in terms of the statistical models applied, and how the data is represented, they largely encompass the same steps. The main steps involved in single-cell data processing include; data correction, quality control, normalization, feature selection, dimensionality reduction, integration, clustering, and visualisation.
As well as optimising analysis methods based on specific datasets, it is also important to adapt pipelines as new learnings and methodologies emerge. As the field has developed, new insights into the technology have arisen which present more accurate and refined ways to deal

with the data. This involves a deeper understanding of single-cell datasets and how the experimental technology impacts the interpretation (67,97).

2.1.1 Aims

This chapter will describe the analytical pipeline used to generate and analyse the single-cell datasets used in this thesis (Figure 2.1), and describe the packages and parameters that were used to perform the bioinformatic analysis in the following chapters.

This chapter will also highlight some key challenges that were faced in the development and optimisation of this pipeline and discuss how they impact the results. This will focus on methods that are not standardised across the field and where there is no clear consensus on how the analysis should be performed. This chapter will discuss two key and persisting challenges of single-cell analysis by summarising the previous approaches and how they've changed through the course of this study.

These include;

2.1.1.1 Ambient RNA correction

Droplet based RNA-sequencing techniques have been shown to contain high levels of ambient RNA. This occurs when mRNA from cells which have been broken in the dissociation process, get released into the cell suspension causing contamination in other cells. It has been recently shown that this contamination can skew results and cause misinterpretation of analysis (98).

2.1.1.2 Differential Expression

In single-cell analysis, commonly used techniques for performing differential expression analysis have involved using individual cells replicates in the test,

which can introduce a high rate of false positive results. This is because cells from the same samples are correlated (99).

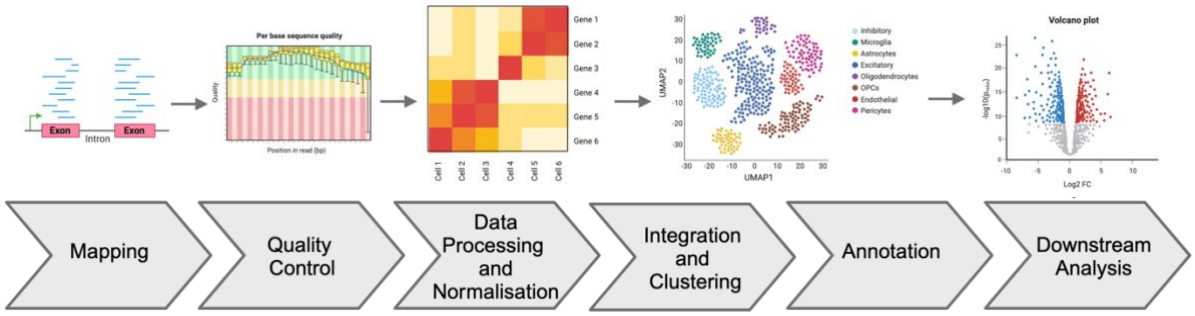


Figure 2.1 – Overview of the single-cell sequencing data processing pipeline. This illustrates the steps involved in processing scRNA-sequencing data from alignment to downstream analysis.

2.2 Methods

The snRNA-sequencing datasets used in the following chapters (Chapters 3&4) have been processed using this optimized pipeline. Following tissue sectioning (performed by UCB Pharma, the Imperial Brain Bank and the UCL Brain Bank), nuclei isolation was performed by Hun Jeong and Daniel Clode, technicians at Imperial College London. The following section describes the experimental protocol and the methodologies that were used in the analytical pipeline.

2.2.1 Experimental Protocol

Single-nucleus RNA-sequencing is designed to capture and analyse the gene expression of individual cells. The protocol begins with single-cell nuclei isolation and fixation which in this study was performed following the Krishnaswami et al protocol (100). The tissue is first homogenised, filtered and then centrifuged. A sucrose density gradient is prepared, with centrifugation to clean up the solution. This is followed by a series of washing steps and finally cells are counted before the 10x Genomics is used. This begins with the permeabilization of cells to allow enzymes and reagents to access the RNA transcripts. Next a single-cell suspension is prepared, cells are counted, and their viability is assessed using imaging. More details on the reagents used can be found in Chapter 3.

Following nuclei isolation and fixation, the 10x Genomics Chromium Next GEM Single-cell 3' Protocol (chemistry: Single-cell 3′ v2) is used. This involves 4 key steps; GEM Generation and Barcoding, Post GEM-RT Cleanup & cDNA Amplification, 3' Gene Expression Library Construction and finally sequencing. The chromium controller is first used to partition cells into Gel Bead-in-emulsion (GEMs) (Figure 2.2a). Each Gel Bead contains an Illumina TruSeq Read 1 primer, a 16 nt 10x Barcode to identify each cell, a 12 nt Unique Molecular Identifier (UMI) to quantify the number of mRNA molecules, and a 30 nt poly(dT) sequence

to capture poly-adenylated mRNA (Figure 2.2b). These GEMs also contain enzymes which facilitate the lysis of cells and release of mRNA. This is then reverse transcribed into cDNA, barcoded for cell identification, amplified, and prepared into a sequencing library.

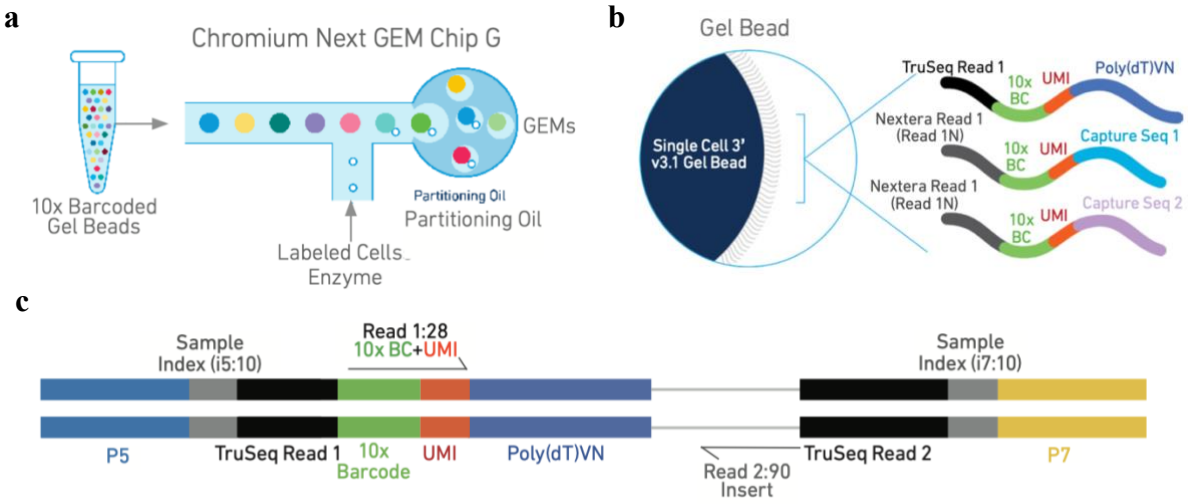


Figure 2.2 – Diagrams of the 10x Chromium Next GEM workflow. a) This illustrates the process of isolating single-cell within Gel Beads-in-emulsion (GEMs), which are formed by encapsulating barcoded gel beads and cells on the Next GEM chip. b) The barcodes are present within a gel bead which contains primers for amplification and unique barcodes for cell identification. c) Final library for single-cell RNA-sequencing; the UMI identifier + cell barcode are 28bp in total. This is followed by a 90bpcDNA insert.

Following reverse transcription, GEMs are broken and the first-strand cDNA is purified using Silane magnetic beads, followed by another round of PCR amplification to generate sufficient material for library construction. Next, cDNA us fragmented and selected by size (300-400bp) to optimise amplification capacity, Illumina-compatible sequencing adaptors are added through End Repair, A-tailing, and Adapter Ligation. P5, P7 and sample index adaptors and TruSeq Read 2 primers are also added. As the 3' assay is used the 10x libraries that are amplified in the final library represent the 3' end of the transcript which is where the

10x Barcode and Illumina Read 2 adaptor are ligated. Libraries are then assessed using Bioanalyzer and qPCR to guarantee the integrity of the samples for the sequencing. The library is then loaded onto an Illumina sequencer and sequencing is carried out using the Illumina NovaSeq 6000 system. The libraries consist of dual-indexed paired-end reads that start and end with P5 and P7 sequences. The 16bp 10x Barcode and 12bp UMI are located on Read 1 which allows for the identification and quantification of transcripts from individual cells. Read 2 is the cDNA fragment which the cell-type specific mRNA fragment. The i7 and i5 sequences are incorporated for sample identification during sequencing (Figure 1c).

2.2.2 Mapping

Once the prepared libraries have been sequenced, the outputted binary base call files (BCL) files are converted to FASTQ files using the Cell Ranger base calling software (SourceBioscience). The raw FASTQ files are then assessed for preliminary quality control metrics such as per-base sequence quality, per-sequence quality, per-base GC content, sequence length distribution and adapter content using FASTQC (101). This is because the accuracy of the downstream analysis is reliant on the quality of the input sequencing data. Following quality control checks, the reads are mapped to a reference genome or transcriptome using the 10x Software Cell Ranger (102). Cell Ranger is a tool designed by 10x Genomics to map single-cell data to the genome to produce gene expression matrices (http://software.10xgenomics.com/single-cell/overview/welcome.) It uses STAR, a splicing aware software that was developed for bulk data (103). Reads with the same cell barcode are grouped together and counted and reads with the same UMI are collapsed to create a gene by cell matrix that quantifies gene expression.

The reference used to map the samples in these analyses were pre-build and downloaded from the 10x website (https://support.10xgenomics.com/single-cell-gene-

expression/software/downloads/latest/). The latest versions of the references were used: GRCh38 for the human dataset and mm10 for the mouse. These use the updated Ensembl 98 gene annotations (104). The command cellranger count (v5.0.1) is used to quantify barcodes with the option the --include-introns which specifies that reads are mapped to intronic as well as exonic reads. For single-nucleus sequencing in particular, it is important to account for intronic reads as this primarily focuses on nuclear RNA rather than cytoplasmic RNA. Nuclear RNA is primarily unspliced pre-mRNA which includes introns. These reads account for 20-40% of mapped reads (105).

The output of mapping is an expression matrix in the sparse matrix format with each column representing the expression of a particular cell in a given gene (rows), quantified by UMIs. A sparse matrix is one in which most of the entries are zero. Storing data in sparse matrix format compresses the expression matrix as it avoids storing the zero values and processes only non-zero elements using significantly less memory and computational power. One of the major issues with snRNA-sequencing data is the presence of zeros in the expression matrix due to low amounts of starting material, despite genes being actively expressed. These zero counts are very dominant and account for over half of the entries in an expression matrix, which means that the data only captures a small percentage of the actual transcript. This is known as the "dropout" rate (106). Reasons for high dropouts rates could be biological, i.e. the gene was not expressed in the cell and therefore there are no transcripts, or technical, i.e. the gene was expressed but for some reason the transcript was lost prior to or during the sequencing. A challenge of this analysis is distinguishing between true biological zeros and technical ones.

One potential approach to address this is expression recovery, where statistical methods are employed to impute gene expression values for the data that were not directly observed but are estimated based on the rest of the data. This approach has also been found to have

limitations such introduction of bias and false signals which can mask real biology, and therefore it was not used in the subsequent analysis (107).

After transcripts have been mapped to barcodes, cells are called. This is usually done based on the number of unique UMIs or genes detected per barcode. This is to determine which barcodes represent real cells. A knee plot is a visualization tool used to help in the cell calling process. When the number of UMI per barcode is plotted in descending order, a sharp drop or "knee" is usually seen in the plot. This is usually used as the cut-off to distinguish between real cells (above the knee) and barcodes representing background noise (below the knee)(Figure 2.3). This is used to decide which cells are retained for further analysis (108).

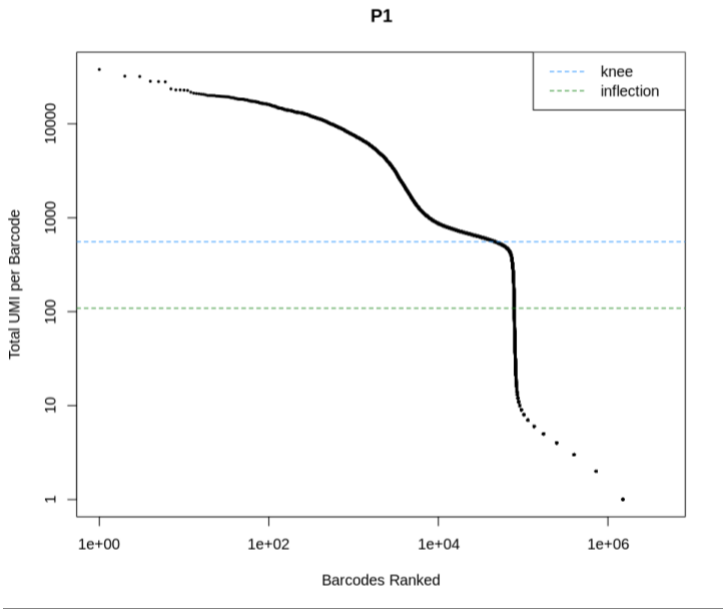


Figure 2.3 – Knee plot to show the ranked UMI per barcode in one scRNA-seq sample from mouse brain tissue. Cells above the "knee" line are retained for further analysis.

2.2.3 Data processing and quality control

Once an expression matrix is generated; this needs to be refined to remove low-quality or damaged cells that are not representative of true biology. Single-cell data can be very noisy and contain technical artifacts which mask genuine biological variability, and so it is crucial to remove and account for this technical variability while preserving biological variation. The first step in quality control is to remove counts from droplets that are either "empty", contaminated by ambient RNA or contain multiple cells (65).

2.2.3.1 Ambient RNA correction

SnRNA-sequencing involves the isolation of cells or nuclei, followed by encapsulation into microfluid droplets containing a unique barcode. When a cell or nuclei gets damaged in the isolation process, this causes the release of freely floating mRNA molecules termed "ambient RNA". These molecules can get captured alongside cells or nuclei, leading to those fragments being tagged with the same barcode as that cell, and contaminating expression profiles, which can cause inaccuracies in downstream analyses (65,98). Droplets which contain only ambient RNA can also be mistaken for cells. This can be particularly problematic when mRNA from different cell-types are caught in one droplet as it can make it difficult to distinguish between cell-types.

CellBender (v0.3.0) (109) is a computational tool that was developed to detect and remove contamination from ambient RNA, particularly from droplet based platforms. It uses an unsupervised deep generative model to remove systematic background contamination by learning the profile of noise counts from the data and predicting counts without the noise. This combines a probabilistic model with a deep neural network approach to learn the cell landscape. CellBender works by removing counts that are predicted to come from ambient

RNA molecules. The CellBender tool can be run on the command line using the command cellbender remove-background. It can also be optionally run with the GPU to speed up computational time. The input to CellBender is the unfiltered count matrix that is the output of mapping. In the analysis of the datasets in this thesis, it was run with the default pipeline, which is able to predict values for the parameters used based on the dataset. Quality control check were performed to endure the automatically determined priors matched the expected, as described in the documentation (https://cellbender.readthedocs.io/en/latest). The output of CellBender is a corrected count matrix which can then be used in downstream analysis.

2.2.3.2 Doublet Correction

Other technical artifacts that need to be corrected for are Doublets (or Multiplets). This refers to instances in the experimental protocol where two or more cells are encapsulated together in a single droplet. This is problematic as the counts will be sequenced as one cell and can confound the analysis, particularly when two heterogenous cell-types are captured together, this can lead to a mixed gene expression profile which may be mistaken for a new cell-type or state. One way to filter out these artifacts is by removing cells with an unexpectedly high UMI count, however there are several more sophisticated tools that have been developed to estimate and remove doublets based on their expression profiles. In these analyses, we use DoubletFinder (110), a computational tool that is designed to identify and remove doublets from snRNA-sequencing expression matrices. Following ambient RNA removal, the DoubletFinder package (v2.0.1) is used to remove doublets from the expression matrix. To do this, each sample must first be pre-processed through the standard Seurat pipeline which includes normalization, variable feature selection, scaling, and dimensionality reduction. DoubletFinder works by first generating simulated doublets and merging them with the snRNA-sequencing data that has been pre-processed in Seurat. It then estimated which of the

real cells have a high number of simulated cells as neighbors in the gene expression space. To run doublet finder, first the number of significant PCs must be calculated, and the optimal pK value needs to estimate based on the data. The pK value is the neighborhood size and requires the input of a pN value to be calculated. This is value defines the number of generated artificial doublets and should be based on an expected percentage based on observation. In snRNA-sequencing brain datasets, the previously observed Doublet proportion in the brain can be up to 7% (110), therefor this value was used to run DoubletFinder in the subsequent analyses.

2.2.3.3 Additional filtering

Further filtering steps are carried out to the gene x cell count matrix to remove low quality nuclei and uninformative genes that contribute to noise. Low quality cells are those which cannot be reliably included in the data and may not give meaningful insights. These may reflect technical artifacts rather than true biological results. There are a number of criteria used to determine these cells and the steps involve filtering of genes and filtering of cells based on several metrics such as number of UMI per cell, genes per cell and mitochondrial percentage.

First, the percentage of mitochondrial genes per cell is calculated and those with a high mitochondrial percentage are removed. This metric indicates the number of genes that are mapping to the mitochondrial genome for each nuclei. A high mitochondrial percentage is an indicator of a stressed or damaged cell that is in the process of dying and are therefore would provide low quality data. Secondly, as this is a single-nucleus protocol, we should not see any mitochondrial genes in the count matrices as these are cytoplasmic, however it is not uncommon to detect mitochondrial genes in snRNA-sequencing datasets as fragments of cytoplasm can be carries along in the experimental process and they can also be an indication

of contamination (111). The function PercentageFeatureSet() in Seurat is used to calculate the percentage of counts from a set of provided features. In these studies, a threshold of 5% mitochondrial content is used to filter cells and any cells above that are excluded. This was optimized based on visualization the data distribution as well as guidance from previous studies (Figure 2.4) (112).

Cells are also filtered based on their UMI counts, as cells with especially high or low counts are likely to be outliers which are attributed to empty drops or doublets or poor-quality cells. To determine the thresholds, violin plots are used to visualize the data distribution and select appropriate thresholds (Figure 2.4). We also apply gene filtering to remove those with sparse expression and are detected in only a handful of cells. These genes typically do not offer any additional information to the analysis and only contribute to noise in the dataset. In these studies, the thresholds used were nCount_RNA < 25000, nFeature_RNA < 7500 and mtPer <= 5.

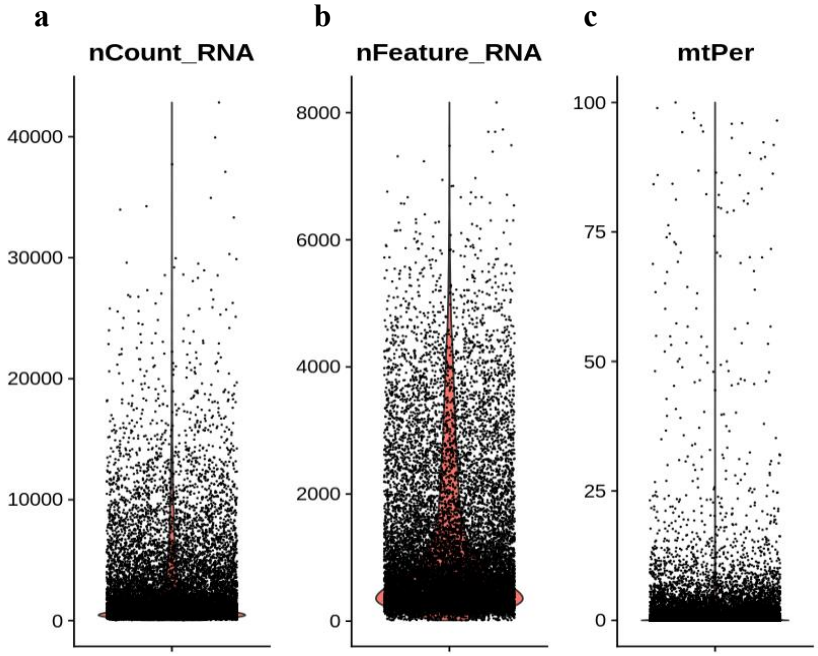


Figure 2.4 – Violin plots showing the distribution of a) Number of counts per cell b) Number of features per cell and c) Mitochondrial percentage per cell in one sample before filtering. These plots were used to determine the filtering thresholds for downstream analysis.

2.2.4 Normalisation

Following filtering, the count matrices of all samples are merged to form one unified count matrix. Due to the variability in the data, sequencing depths can differ, and data needs to be adjusted for differences that may have arisen solely because of sampling effect. This ensures that cells can be compared directly despite the differences in depth. Normalization helps to mitigate noise arising from technical variation and allows cells and samples to be more easily compared. Highly expressed genes often dominate the datasets, and normalization ensures genes can be compared by expression pattern rather than just absolute count (113,114) Data is normalized in LIGER (v1.0.0) (115) using the function normalize() to account for total gene expression across a cell. The LIGER function selectGenes() is then used to identify a subset of highly variable genes which are most representative of the biological variance in the dataset. Genes that are consistently expressed across cell-types are not as informative in the downstream analysis. The function identifies genes with a variance above a given variance threshold, based on the assumption of a Poisson distribution. Finally, the data is scaled using the function scaleNotCenter() to ensure that each gene has the same variance across samples.

2.2.5 Data Integration

Single-cell data that comes from multiple samples needs to be corrected for batch effect. This is because each sample may have technical variation that is unrelated to the underlying biological variation. Datasets produced under different sequencing conditions result in differences due to the experimental condition. This can cause spurious differences in the datasets that are not due to the biology, and make integrating samples difficult. The number of published datasets means that data integration problem is at the forefront of problems in single-cell analyses. Being able to use these pre-existing datasets can yield new information long after they have been published. There are many integration tools that have been

published to standardize datasets, and there are a number of benchmarking studies that have bene published showing that different datasets require different methods of integration (94). As well as data integration, we also need to account for batch effect. These are often attributed to differences in reagents, sample handling, processing times and equipment. Sequencing differences such as lanes, read depth, different flow cells, and read length can also cause discrepancies. These technical discrepancies can hide biological differences, leading to incorrect interpretations (114). To date there are >50 data integration methods available (93). It can be difficult to choose a method as these are often dataset specific. There are several benchmarking studies that have been done to facilitate and guide to a choice of method (65,93,94,116). These are helpful but the choice is often subjective.

In this study as well as multiple samples from the same condition being combined, samples from different conditions, epilepsy models and brain regions were all combined on a joint graph in order to find common cells. Combining datasets increases the power of the analyses to detect cell-types that are shared across samples. As these samples have been handled in distinct groups, integration of these datasets is essential to ensure they can be compared effectively (117). LIGER uses non-negative matrix factorization on the normalized and scaled datasets to adjust for unwanted variation. This identifies shared factors by summarizing the data as a set of metagenes (factors) and transforms the dataset in a low dimensional space to capture ethe primary sources of variation. These factors are identified which are unique to each dataset and shared across datasets. The shared factors are then aligned using canonical correlation analysis (CCA) to create a common factor space. This allows cells from different datasets to be compared directly. This is an iterative process to optimize the accuracy. This involves aligning the data across experiments, removing batch effects, and integrating the data into a single dataset. First k, which is the number of matrix factors is calculated. This is a kind of dimensionally reduction similar to PCs in a dataset.

This value can be optimized for each dataset. K was optimized separately for both mouse and human datasets but k=30 in both cases. Next there is a quantile normalization step using the quantile_norm() function which uses dataset, factor and cluster quantiles to fully integrate the data. This produces joint clustering and dimensionality reduction of the integrated datasets which can be used directly for downstream analyses (94,115)

2.2.6 Clustering

Clustering is a key step in the analysis of single-cell data, which groups cells with similar gene expression profiles into clusters. This grouping is essential for uncovering cellular heterogeneity and understanding the biological functions of the cells involved in the tissues being studies. To classify the cells into distinct groups based on their gene expression profiles, clustering algorithms can be used. There are a number of clustering algorithms available for single-cell data and the choice of algorithm and parameters can depend on dataset (65). This choice, particularly the resolution, can greatly impact the outcome of the clustering. Choosing a resolution that is too high may split one cell-type into multiple clusters, but a low resolution may not distinguish distinct cell-types and merge them into a single cluster. In this study, as the focus was on investigating major cell-type rather than subtypes, a lower resolution was chosen (res=0.3). This was also optimized by the ability to discriminate between cell-types and how clearly, they could be segregated. The Louvain clustering algorithm (118) was used due to its ability detect communities using a completely unsupervised approach with no prior assumptions needed, as well as its computational efficiency. This graph-based clustering method uses community detection within large networks., by representing cells as nodes where the edges between the nodes represent the similarity in gene expression profiles between cells. The algorithm seeks to quantify the density of edges within communities in comparison to the edges between communities. It

first looks for small communities by optimizing the density of the edges, and then aggregates nodes from the same community to build a new network, and this is repeated iteratively. Once the graph is constructed the algorithm identifies communities of cells based on similar expression profiles. Louvain clustering was implemented within LIGER using res=0.3 and otherwise default parameters.

2.2.7 Dimensionality Reduction

To effectively visualize single-cell data, dimensionality reduction techniques are first applied to facilitate the interpretation of the inherently high-dimensional datasets into a more comprehensible form. The high dimensionality caused by the large number of genes, can be difficult to interpret and obscure meaningful patterns so the data needs to first be simplified. When transferring high-dimensional data to a low dimensional space, it is essential to preserve as much of the relevant data structure as possible. Several methods can be used to reduce dimensionality for visualization including principal component analysis (PCA)(119) or t-distributed stochastic neighbor embedding (t-SNE)(120), and Uniform Manifold Projection (UMAP)(121). PCA (Principal Component Analysis) is usually the first step and condenses the data to a set of principle components that capture the primary variance in the dataset. UMAP and tSNE are then applied to reduce the data while preserving the intrinsic structure.

In this study, UMAP is used due to its robust ability to capture the underlying topological structure of the data. For each data point the algorithm identifies neighboring points similarly to nearest neighbors algorithms. These connections create a weighted graph, which emphasize the preservation of global structure. The graph is then constructed by representing the data in a low-dimensional space. Key parameters include; the number of neighbors each

point is connected to and minimum distance which controls how compacted the UMAP points are. These parameters were optimized according to the specific data used. UMAP was selected over t-SNE as it has the advantage of preserving and maintaining relative distances between clusters, which mean the embedding can be more interpretable. Visualizing the data on a graph is important as it allows us to explore the underlying biological data and reveal patterns and trends that are integral to understanding the complex structure of brain tissue (122). Figure 2.5 shows a UMAP with the adjusted parameters n_neighbors=20 and min_dist=0.3. n_neighbors specifies the number of points used in estimating the local manifold structure and min_dist controls how tightly points are compressed in the embeddings.

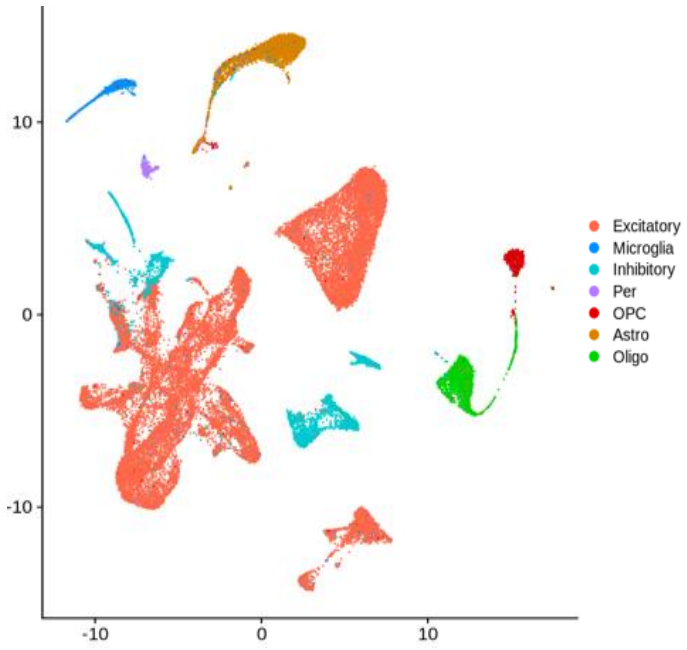


Figure 2.5 – UMAP to show the clustering of samples from one mouse model of epilepsy using the parameters n_neighbors=20, min_dist=0.3.

2.2.8 Cell-type Annotation

To assign specific cell identities to clusters, cluster specific markers were first identified using the Seurat function FindAllMarkers(). This applies a Wilcoxon rank sum test to identify uniquely differentially expressed genes (DEGs) between clusters. The parameters min.pct = 0.25 and logfc.threshold = 0.25 are used to restrict the analysis to only genes detected in 25% of cells, and only genes with a minimum log fold change of 0.25 log respectively. In addition, a false discovery rate (FDR) threshold of <1% was used to determine significance, and only positively DEGs are considered markers. Following the identification of DEGs for each cluster, a Fisher's exact test was conducted to assess the enrichment of known cell-type marker genes within the cluster specific markers (70,112). Cell-type annotations were assigned based the most significantly enriched set of marker genes. The set of marker genes used in this analysis were derived from two human single-cell datasets by Wang et al. (68,123,124). Verification of cell-type assignment was performed by visualising a heatmap of canonical cell-type marker genes, confirming their expression profiles within respective clusters. The canonical markers were as follows: Excitatory Neurons: SLC17A7, SATB2, VIP, LAMP5, Inhibitory Neurons: GAD1, GAD2, SOX6, PVALB; Astrocytes: AQP4, GJB6, FGFR3; Microglia: CTSS, C1QB, CSF1R; Oligodendrocyte Precursor Cells (OPC): CSPG4, PDGFRA, VCAN; Oligodendrocytes: MAG, MOG; Pericytes: PTGDS, ATP1A2, ITIH5, FLT1, DCN, PDGFRB; Endothelial Cells: ACTA2, KCNJ78, ZEB1.

2.2.9 Differential Expression Analysis

Differential expression analysis compares gene expression levels between different biological conditions or cell-types, aiming to identify genes whose expression levels vary significantly. This type of analysis can provide insight into the biological processes underlying cell differential and the molecular basis of disease.

To perform differential expression analysis, first genes that were expressed in ≤ 5 nuclei were first excluded from the analysis, to avoid anomalies due to low expression. Next, count matrices were extracted for each cell-type and aggregated per gene by summing values across samples to create a "pseudo-bulk" expression matrix. Following this, differential expression analysis was performed separately for each cell-type, across pairwise disease group comparison using the R-package DESeq2 (v1.4.0) (125). This tool, made for analysing bulk RNA-sequencing datasets, applies a negative binomial model by first estimating the count variance for each gene, accounting for both biological and technical variance across samples. This takes into account the over-dispersion of the sequencing data using shrinkage estimators for fold change. Covariates such as age, sex, and batch are also regressed out in the model. DESeq2 first models the raw counts using normalisation factors to account for library depth differences. It then uses the gene dispersion values and shrinks these estimates to generate more accurate estimates. After the binomial model is fit, The Wald test is used to perform the differential expression analysis.

The Log2 fold changes in gene expression are also calculated and associated p-value is assigned to each gene. DESeq2 also corrects for multiple testing using the Benjamini-Hochberg (BH) (126) procedure. This is an important step when analysing thousands of genes where the chances of false discovery are high. DESeq2 is a robust and reliable framework which is also computationally efficient in comparison to other methods of differential expression. The number of cells per sample was used as a covariate in the model to adjust for samples with larger cell number. The following design was used:

```
design = model.matrix(~condition + cellnumbers)
```

Sex and age of the samples were included as covariates where appropriate.

2.2.9.1 Multiple Testing

Multiple testing is the statistical challenge that arises when a large number of hypothesis tests are performed simultaneously. In the context of gene expression, this occurs when tens of thousands of genes are tested across conditions which can lead to an increased risk of false positives and genes appearing to be differentially expressed due to chance rather than due to true biological difference. Correction methods such as Bonferroni or Benjamini-Hochberg (126) are used to correct for false discovery rate. These adjust the p-values to account for the number of tests being performed, reducing the likelihood of false discoveries. The threshold used can help to control for type I error while also maintaining significant discovery. Handling of multiple testing is crucial in differential gene expressions studies due to the large number of tests being conducted.

2.2.10 Functional annotation

The R package WEB-based Gene SeT AnaLysis Toolkit *WebGestaltR* (0.4.6) (127) was used to perform over representation analysis for Gene Ontology (GO)(128) databases biological process, molecular function and cellular component. The DEGs from each pairwise comparison from each cell-type were tested for enrichment of the GO terms, against the background set of genes which was the number of cells expressed in each cell-type. The default WebGestalt parameters were used. These include GO terms were filtered by size to include terms with a minimum of 10 genes and a maximum of 1000 genes. Genes (see results). P-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) method (129). Terms were considered significantly enriched with an FDR value of >5%. The GOSemSim R (130) package (v 2.16.1) was used to measure of semantic similarity

(measure = “Wang”) and the rrvgo R package (v 1.1.4) was used to reduce terms as described in Feleke et al. (112).

2.2.11 High-performance computing (HPC)

The computational analysis in this project was performed on the Imperial High Performance Computer from the Imperial College Research Computing Service, DOI: 10.14469/hpc/2232. R (v4.1.0.0) was used in the majority of computational analysis this analysis unless otherwise stated. Conda was used to maintain all packages and versions within environments and to manage pipeline requirements and dependencies.

2.3 Pipeline Optimisation

This section explores two components of this pipeline that are critical in single-cell analysis. These were optimised specifically for the analysis of single-nucleus sequencing datasets from brain tissue. These address some of the most challenging aspects of single-cell analysis that remain unresolved, and these optimizations aim to enhance the accuracy and reliability of the results obtained from such complex biological data. The following optimisation analysis were performed on a subset of the datasets used in further chapters. These were samples from mice with Pilocarpine induced epilepsy (n=7) and matched sham mice (n=7). See Chapter 3 for more details on the generation of this mouse model.

2.3.1 Ambient RNA correction

Ambient RNA refers to the extracellular molecules of RNA that are not associated with a cell and can float freely to contaminate snRNA-sequencing droplets or adhere to the exterior of cells. This typically occurs during cell lysis and dissociation when cells break and RNA gets released into the surrounding environment. Despite the potential to significantly impact results, the issue of ambient RNA has received limited attention in the field. There are few approaches available to mitigate the effect of ambient RNA on snRNA-sequencing data, and these are rarely applied in standard single-cell studies, leading to a general lack of awareness of this problem. This can lead to unreliable findings and skewing of the gene expression profiles that lead to misattribution of RNA sequences to the wrong cells.

In a recent study by Caglayan et al. (98), it was shown that ambient RNA can significantly disrupt the reliability of single-nucleus RNA-sequencing studies that fail to account for it, leading to considerable misinterpretation of results. This contamination arises especially from more abundant cell-types, which can contaminate less abundant cells, and suggest that certain

genes are expressed across all cell-types. The authors reanalysed several previously published datasets and found them to be contaminated with ambient RNA, particularly in brain datasets. They showed that ambient RNA could contribute up to 50% of total RNA in some samples, primarily originating from neurons and affecting all glial cell-types. Neurons are also particularly susceptible to damage with creates even greater risk of ambient RNA release. This risks the misidentification of rarer glial cell-types. They also discovered previously unrecognized cell-types and subtypes, including rare populations which could actually be glial types contaminated with ambient RNA. Additionally, analytical tools which model RNA profiles might be inherently biased towards neuronal signatures, misinterpreting data from other cell-types. Despite the prevalence of this issue, and its potential to introduce false positive results and misinterpret results, it remains under-recognized. To address this challenge, several tools, including CellBender, SoupX, and DecontX, and have been developed to adjust expression matrices and mitigate the effects of ambient RNA contamination.

To explore the impact of ambient RNA in our generated datasets, first an approach was applied using the tool EmptyDrops (108). To test whether ambient RNA signals can drive results and cause false positives, EmptyDrops was used to first extract cells that can be considered “empty” based on their gene expression profiles. This means they may contain ambient RNA but not whole cells. The EmptyDrops algorithm finds and filters out “empty” droplets from scRNA-sequencing data by first modelling the distribution of the RNA content in the droplets, assuming that empty droplets follow a distinct distribution from real cells. A mixed model is fit to the data and a probability is given to estimate the likelihood of a droplet containing a cell.

First, the empty droplets and the real cells were separated out and aggregated by sample to give a “pseudo-bulk” expression matrix, one for empty drops and one for real cells.

Differential expression was then performed using DESeq2 at an FDR threshold of 5%. Table 2.1 shows the numbers of DE genes between Sham mice and mice with pilocarpine induced epilepsy using both real cells and empty drops. A much larger number of DE genes was seen in the real cells compared to empty drops replicating true biology. An intersect of the two sets of DE genes showed that there are 59 genes which are common between them.

	Number of DE genes
Real Cells	1823
Empty Drops	115
Intersect	59

Table 2.1 – Summary of differentially expressed genes from real cells contained in droplets vs empty droplets containing ambient RNA. A much larger number of DEGs can be seen in the analysis using real cells, with some overlapping results between both tests.

A Fisher's exact enrichment test also showed that the most highly variable genes within the empty drops samples are significantly enriched for neuronal markers (Wang et al.). Functional annotation (Figure 2.6) of these DEGs showed that terms related to both sets were mostly neuronal related, however the DEGs from the real cells showed epilepsy related terms such as glutamate receptor signaling, signal release and membrane receptor regulation, while the terms from the empty drops were more development related, and do not necessarily directly reflect epilepsy biology. This confirms the presence of mostly neuronal RNA molecules in the ambient RNA.

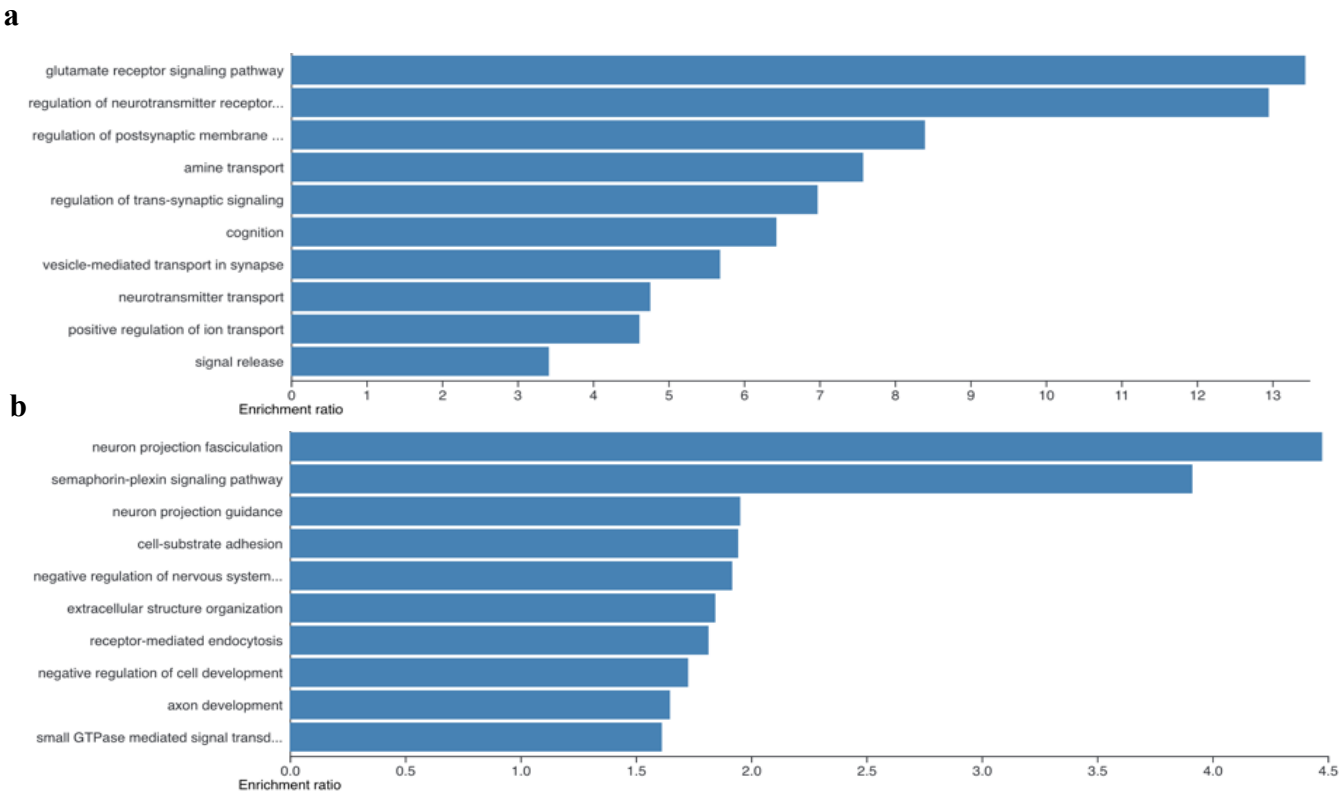


Figure 2.6 – Functional annotation of Gene Ontology (GO) terms across DEGs in a) real cells and b) "empty drops". The graphs show only terms that were enriched at FDR 5% and the y-axis represents the Ratio of the enrichment.

Next, we wanted to test the contamination of ambient RNA in the real cells. There are several available tools that correct the snRNA-sequencing count matrix for ambient RNA. Two of the most frequently used tools include SoupX(131), and CellBender. While SoupX is the most highly cited and used tool, it requires an input of the estimation of ambient RNA percentage and results are highly sensitive to this parameter selection. If this parameter is not adequately chosen, the data may be over or under corrected. Given the variability of brain tissue, utilizing software that demands exact parameter estimation presents challenges. As the outcome of the analysis is highly dependent on this value, it can be unreliable across multiple datasets.

CellBender is an alternative tool that aims to remove the effect of ambient RNA contamination by correcting the expression matrix using unsupervised learning. This algorithm is used to detect profiles of ambient RNA in the data and correcting for these profiles. The mouse samples were first filtered, clustered and annotated without ambient RNA correction (n=14), and then the final matrix of 34,257 cells was corrected for ambient RNA. Figure 2.7 shows the data before and after ambient RNA removal. These UMAPs show that the data after correction looks cleaner and more tightly clustered despite the same clustering parameters being used.

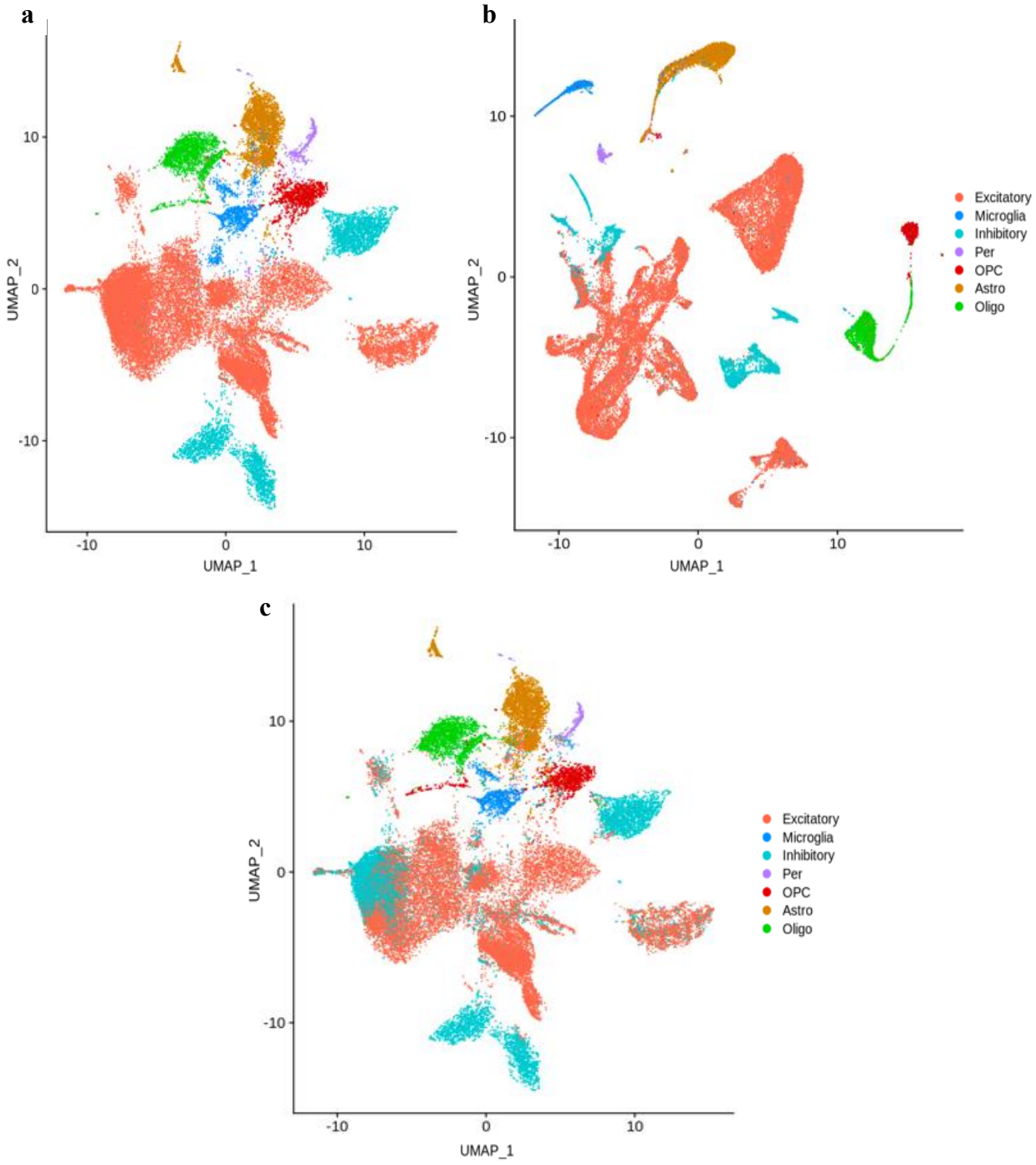


Figure 2.7 – UMAPs showing the clustering of 14 mouse samples before and after correction for ambient RNA. a) shows the clustering before ambient RNA correction b) shows the clustering after ambient RNA correction and c) shows the reannotation of the cells after ambient RNA correction.

The numbers of cells per cell-type before and after ambient RNA correction is shown in Table 2.2. This shows that removal of ambient RNA has caused many cells to be reannotated. Figure 2.7c shows the reannotation of the cells after ambient RNA correction. This shows that many of the reannotated cells are neuronal. Many inhibitory neurones had been labelled as excitatory prior to correction. This is likely because excitatory and inhibitory neurons share much gene expression and therefore are harder to distinguish when they have been contaminated. As the excitatory neurones are the most abundant cell-types, most of the ambient RNA is likely to be from these cell-types therefore masking the inhibitory neurons. The glial cell populations show less change after correction with the greatest change seen in Microglia and Pericytes.

	Excitatory	Inhibitory	Astro	Microglia	Oligo	OPC	Per
Pre-CellBender	22,734	3,452	3,272	1,145	2,018	1,110	526
Post-CellBender	16,939	9,713	3,373	842	1,939	1,088	363

Table 2.2 – Summary of differentially expressed genes from real cells contained in droplets vs empty droplets containing ambient RNA. A much larger number of DEGs can be seen in the analysis using real cells, with some overlapping results between both tests.

To asses whether the ambient RNA profiles were coming mostly from neuronal cells, we examined the expression of a neuronal marker gene *Syt1* which should only be expressed by neurones. Figure 2.8 shows the expression of *Syt1* in the cells before and after ambient RNA correction. Pre-ambient RNA correction, it was moderately expressed in the glial clusters, however post correction, *Syt1* expression is decreased in the glial cell clusters. This shows that neuronal ambient RNA was present and has been removed.

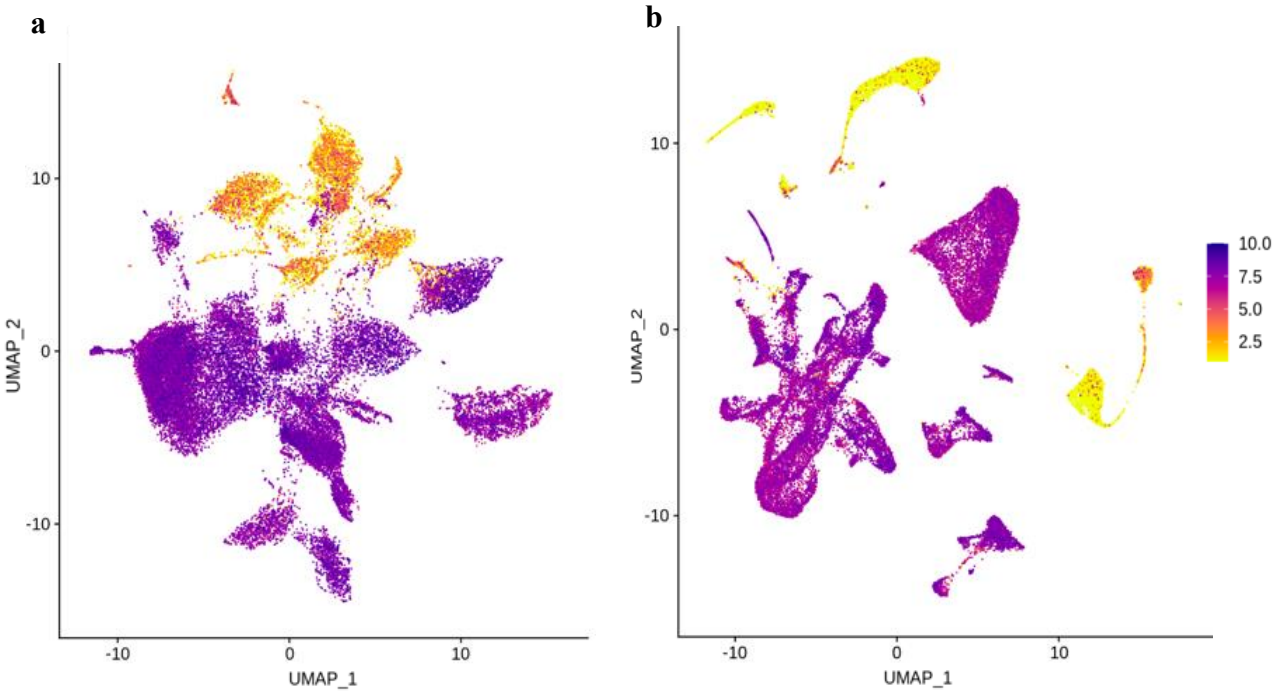


Figure 2.8 – UMAPs showing the expression of the gene Syt1 a) before and b) after ambient RNA correction. The expression of neuronal RNA in the glial cell clusters is decreased after CellBender.

Finally to tests the impact of ambient RNA on downstream analysis, differential expression analysis was performed on the cells before and after correction and reannotation. Table 2.3 shows the number of DEGs in the analysis before and after and the intersecting DEGs. The numbers of DEGs mostly decease after ambient RNA correction, suggesting that ambient RNA causes false positive results in the differential expression. There is a large overlap in

these DEGS (~50%), showing that there are many genes which are not detected prior to corrections.

	Excitatory	Inhibitory	Astro	Microglia	Oligo	OPC	Per
Pre-CellBender	1835	705	557	5	87	18	0
Post-CellBender	1234	247	421	1	50	13	0
Intersect	668	128	214	1	40	7	0

Table 2.3 – Table showing the number of differentially expressed genes before and after CellBender correction, and the intersection of these genes.

2.3.2 Inflation of Type 1 Error in differential Expression analysis

Differential expression analysis is a key analysis used in the study of transcriptomics that is used to help unravel the genetics of complex disease. Differential expression involves the comparison of levels of mRNA transcripts across different conditions to understand how genes are involved in health and disease. This can help provide insight into the molecular underpinnings of the disease and highlight which distinct genes are upregulated or downregulated in disease. This can enhance our understanding of the disease state and the processes which are affected by the disease. This type of analysis offers information about the genes that modulate disease and can offer important insights for drug discovery by targeting genes that are contributing to symptoms or perturbing normal function. Historically, differential expression analysis has relied on bulk RNA-sequencing, which averages the expression of signals across all cells in a tissue sample. This can potentially obscure the contributions of individual cell-types. With the advent of single-cell RNA-sequencing, we can now dissect the expression of multicellular tissue into single-cells, enabling a more precise identification of how different cell-types respond to disease. This

resolution allows for the characterisation of the complex, multidimensional response to biological perturbation, which in turn can provide a better understanding of disease mechanisms to facilitate he identification of therapeutic targets.

Reliable values in DE analysis are highly reliant on good quality control and filtering of the data so that there is no noise masking the true biological signal. As these data sets have different distributions and need to be modelled differently, the DE approach used must be developed accordingly. There are many bioinformatic tools that have been developed for both bulk RNA-sequencing DE analysis and single-cell DE analysis, and the choice of tool is often dependent on dataset. Single-cell data is sparse and has a high degree of technical noise and variability, and single-cell tools typically need to account for this sparsity. The data is skewed and a few cells express genes at high levels whereas most cells express the same genes at very low levels or not at all (132). Models such as zero-inflated and negative binomial which can handle the overdispersion in the data are typically used (133). One way of dealing with the sparsity of these datasets is by using dimensionality reduction methods and methods which take only the most highly variable genes; however, this can also be sensitive to normalisation effects which might affect the downstream analysis (114). This may also ignore smaller and rarer populations which are not defined by highly variable genes. Some people have suggested leveraging the sparsity of the data by using the dropout as binary values- this is suggested to be as useful in identifying cell-types as using the values themselves, but maybe not be useful in other downstream analysis. The same gene may be expressed in one cell-type and not in the other. Dropouts happen because we start with such sparse material as we are extracting it from only one cell (106). This means that the low amount of starting mRNA causes inefficient capture, which subsequently doesn't all get amplified in the PCR step. mRNA expression is also stochastic and dependent on different

processes and time points in the cell's life. The sequencing will only capture one time point in the cell which means mRNA levels will vary depending on the state of that cell.

Bulk data on the other hand is typically normally distributed after log transformation, as reads are aggregated across many cells which can stabilise the variance. Single-cell RNA-sequencing is also different in its structure and must be modelled and interpreted differently to bulk-data. While tools designed to perform differential expression on bulk data is well established, tools for single-cell data present a unique set of challenges. Typically tools designed for use on single-cell date treat single-cells as replicates, creating a situation where a group of cells taken from the same sample are treated as if they were independent samples. This can create a significant issue in the reliability of results as it violates the assumption of independence that is required for most statistical analyses. A paper by Zimmerman et al. (99) published in 2021 empirically documented the sued the correlation among measures from cells from individuals demonstrating how the current common practice leads to an overinflated type 1 error rate. A type 1 also known as a false positive, occurs in hypothesis testing when a statistical test falsely indicates that there is an association or difference, when there isn't one (134). Pseudo-replication, also refers to as subsampling, refers to observation that are nested within an experimental unit and are not statistically independent. While it is often possible to account for sample correlation within these models, this paper found that most single-cell studies do not, creating potentially unreliable results with high rates of false discovery.

MAST (Model-based Analysis of Single-cell Transcriptomics) (135) is a statistical framework designed for the differential expression analysis of single-cell data, and was designed to address the challenges of sparsity and noise presented by the data, particularly

dropouts. It uses a two-part hurdle model that models gene expression both continuously through the gene expression values and discretely through the cellular detection rate (CDR) which informs the fraction of genes that are detectable in each cell, acting as a proxy for technical factors such as dropouts.

This model typically does not include random effect for individual within the model as the most common approach in literature which leads of inaccurate results.

To assess the rate of false discovery generated when using MAST, differential expression was performed using mouse samples from the pilocarpine-induced model of epilepsy (n=7) and matched controls (n=7) for each cell-type. Additionally, this was repeated using the same samples but assigned to random groups as case and control and this was permuted 10 times. *Figure 2.9* shows the numbers of DEGs identified with the real groups (red) compared to the number of DEGs identified with the different sample permutations (grey). The values shown are significant at a threshold of FDR < 5% and log2FC of > 1.5. This showed that many genes are differentially expressed even with the random groups of samples, and at times the random groups are higher with random samples. This shows that DE is inflated when using MAST, as there should not be DE genes when samples are randomly assigned to case and control. The same study also highlighted that the MAST framework, when properly applied with samples as covariates, offers a more accurate differential expression analysis. However, many published studies do not apply this and often rely on default settings, which don't take sample covariates into account. Cells from the same samples are more likely to have correlated gene expression profiles which should be taken into account in the analysis. Correctly modelling the single-cell data is very important to distinguish true biological signal to produce reliable, reproducible, and biologically meaningful results. The high degree of heterogeneity in single-cell level data, makes detecting differential expression difficult at a

single-cell level as there is a high degree of technical and biological differences, making it harder to detect truly DE genes.

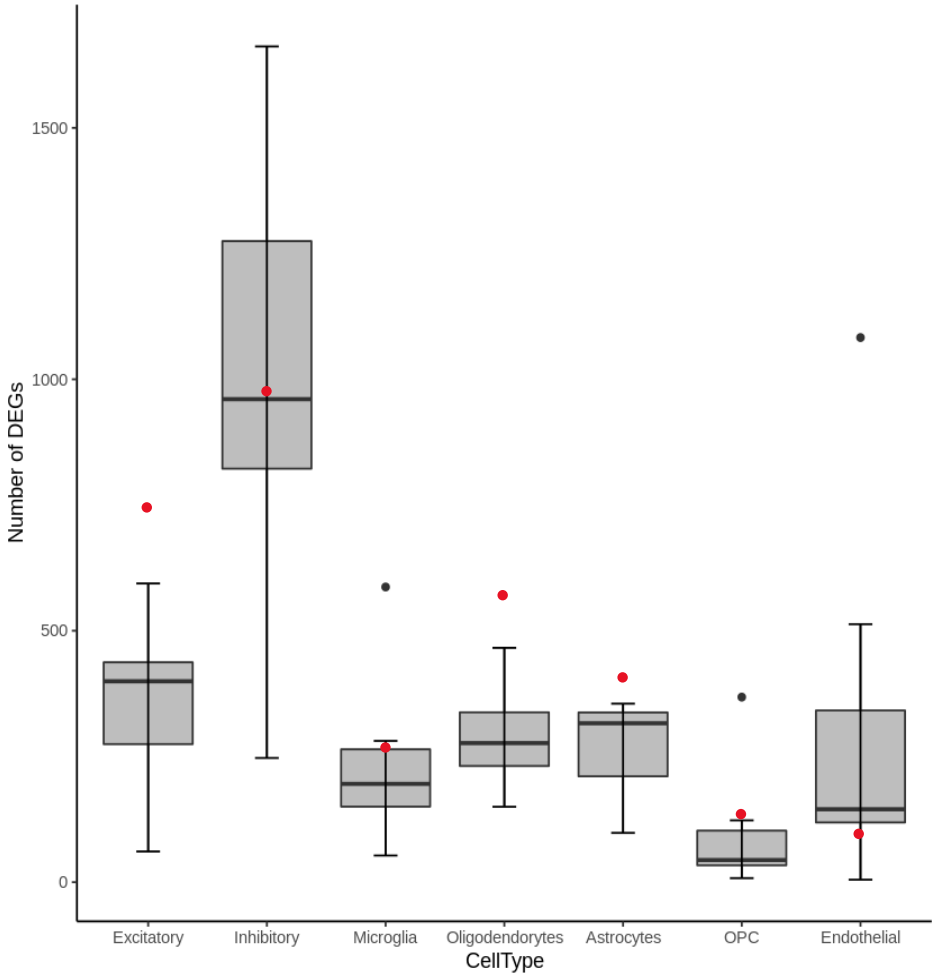


Figure 2.9 – Boxplot to show the number of DEGs using MAST when sample labels are randomly permuted per cell-type, compared to when correct sample labels are used (red).

To help explain these results, the coefficient of variance (CV) was used to asses the variability in the gene expression levels across the cells in these datasets. This is a standardised metric of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage and is defined as the ratio of the standard deviation σ to the mean μ. The formula for calculating the CV is below:

$$CV = \frac{\sigma}{\mu}$$

A higher CV indicates greater dispersion round the mean and more heterogeneity in the expression while a lower CV indicates more uniform expression. This was calculated both for all samples and for control and case samples separately (Figure 2.9). The CV of all samples (0.0642) was very close to that of cases (0.0602) and controls (0.0660) separately. This showed that there is just as much variation in means within case and control as across all subject, explaining why DEGs are seen even when sample labels are randomly shuffled.

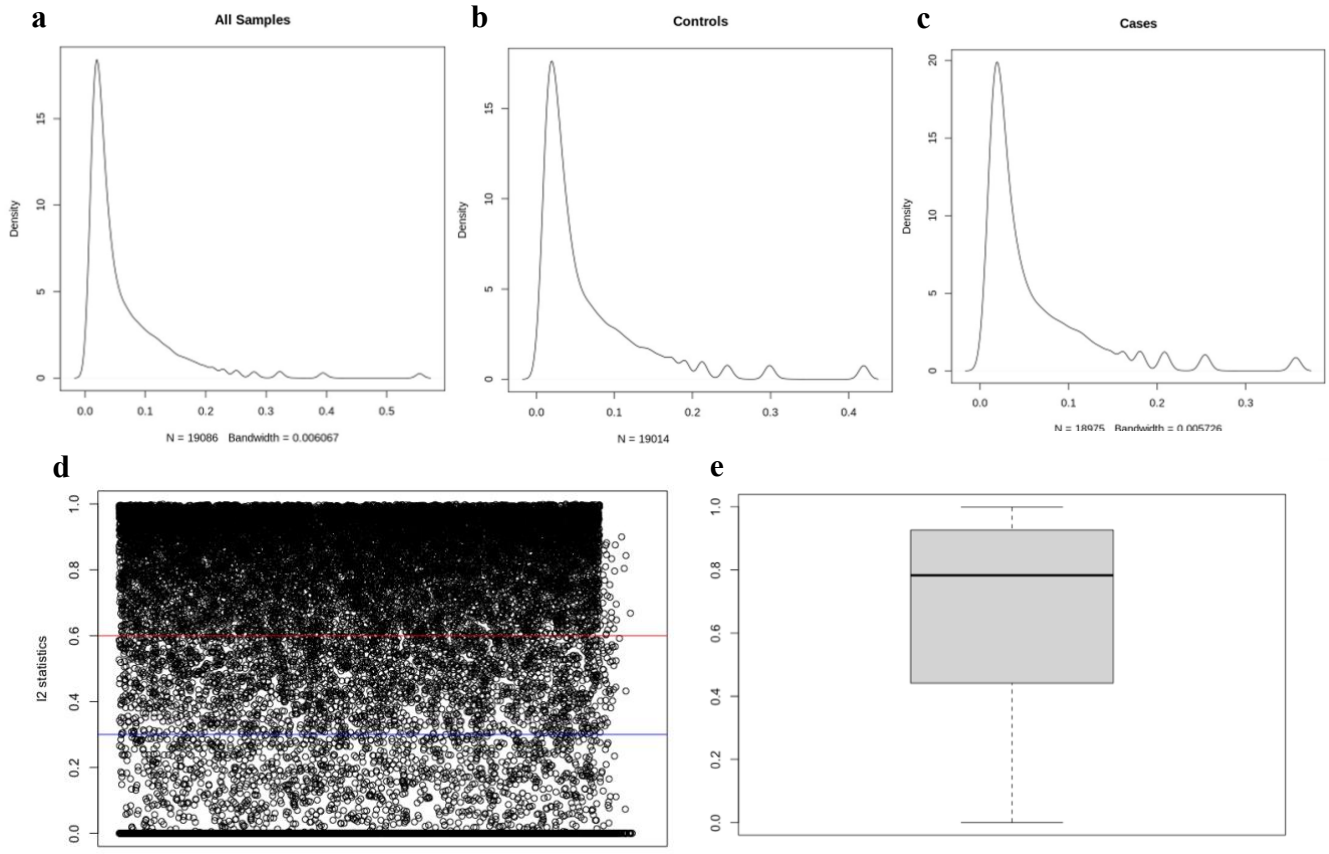


Figure 2.9 – Measures of heterogeneity in single-cell data. a) Coefficient of Variation of all samples b) all controls c) all cases d) Scatter plot showing heterogeneity using I^2 per gene across all samples e) Boxplot of I^2 per gene across all samples. All figures were generated using gene expression values from excitatory neurons.

The I^2 statistic is a measure used in meta-analyses to quantify the proportion of variation observed between studies that is due to heterogeneity rather than chance. I^2 describes the percentage of total variation across studies that is not attributable to sampling error. The range for I^2 is 0-100%, and a value of 0% indicates no observed heterogeneity, while higher values show increasing heterogeneity, a value >50% is considered to be heterogeneous. Figure 2.9d shows the distribution of I^2 across all cases and controls, and figure 6E shows these results as a boxplot. The mean value is 0.76 which shows very high variability of these genes.

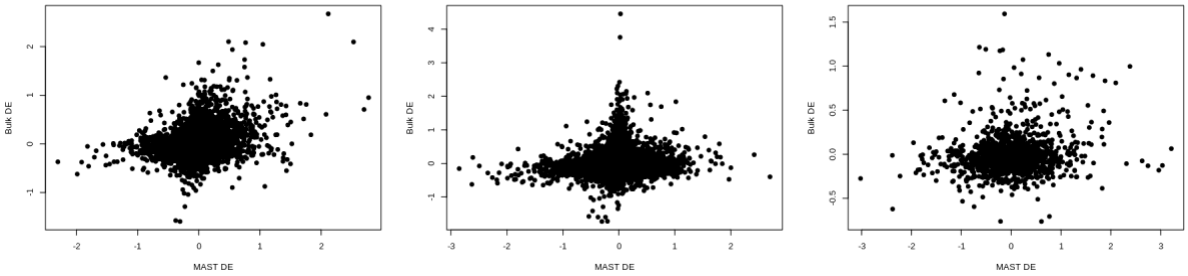


Figure 2.10 – Scatter plots to show the correlation of Log2FC when using MAST vs Bulk DE. a) Excitatory neurons b) Inhibitory neurons and c) Microglia.

To further asses the DE results produced by MAST, DEGs were compared to previously generated bulk RNA-sequencing from pilocarpine mice (n=200) (18) using the calculated Log2FC. While bulk data is an averaging of signal, the highest proportion of cells are neuronal and should correlate with the MAST single-cell results. Figure 5 shows the correlation of the bulk Log2FC vs the Log2FC from MAST for 3 cell-types. This shows that the corelation of Log2FC is low when using mast for differential expression when compared to bulk differential expression (Figure 2.10).

A suggested approach to this problem is to aggregate single-cell gene expression by summing values across all cells from one sample creating “pseudo-bulk” expression matrices. These

maintain the cell-type specific values but remove zero-inflation and within sample correlation. This can remove the noise that is present when looking at the data at a single-cell level, leading to clearer signals of true gene expression levels. It also removes much of the sparsity in the datasets which creates normally distributed datasets and allows for the usage of traditional DE analysis tools. This means the datasets are more robust and less heterogenous and which can confound results.

Pseudo-bulk is a more conservative approach and while they reduce false positives, they result in much fewer DE genes with underpowered studies that miss detecting genuinely differentially expressed genes. This underscores the necessity of nuanced approaches in single-cell differential expression analysis to balance the trade-off between detecting true positives and limiting false discoveries. As samples had already been generated, post-hoc power calculations were performed to see whether the DE is suitably powered. First 14 samples (7 cases and 7 control) were aggregated per cell-type by summing the values per gene across samples. This creates a "pseudo-bulk" matrix. The R package *ssizeRNA* (v1.3.2) (136) was used to perform sample size calculations. Figure 2.11 shows the result of the power calculations. To detect genes differentially expressed at Fold Change 1.5 or more, we need at least 26 samples (13 case, 13 control) and to detect genes differentially expressed at Fold Change 2 or more, we need at least 11 samples (6 case, 6 control).

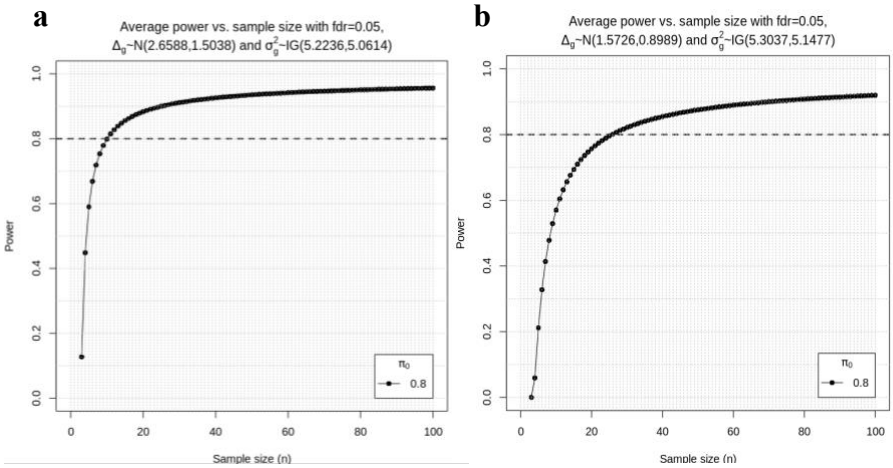


Figure 2.11 – Sample size calculations at a) Fold change 2 and b) Fold change 1.5 using ssizeRNA.

For the subsequent stages of analysis, the pseudo-bulk method was chosen as despite its conservative nature and the requirement for a larger sample size, the approach is deemed more precise and less prone to result inflation, thereby ensuring confidence in the validity of the findings.

2.4 Conclusion

In the rapidly advancing field of single-cell RNA sequencing (scRNA-sequencing), the use of analytical methods is critical to accurately interpret complex data. As technology evolves, there are several technical challenges still to address to ensure reliability and reproducibility. Datasets are increasing in size and the sparsity and variation in these datasets, as well as the complexity of biological systems needs to be well understood and accounted for (96). Challenges include: data size and sparsity, batch effect, dimensionality, noise, integration, annotating, computational resources. There is an increasing need for multidisciplinary approaches using computer science, statistics, machine learning and biology to refine data analysis workflows. The amount of information that is uncovered by single-cell data is unprecedented. With this advent of "big data", it is even more crucial to develop efficient methodology that can account for variation, complexity and volume of data.

This work emphasises the importance of considering extracellular ambient RNA when using droplet based snRNA-sequencing methods. This is particularly problematic in tissues like the brain, where the predominance of neuronal RNA can lead to misidentification of less abundant glial cell-types. Although tools such as SoupX, DecontX, and CellBender offer solutions to correct for ambient RNA, their application is not yet standard practice, leading to widespread under recognition of the issue (65,98). This gap in the methodology highlights the need for increased awareness and adoption of ambient RNA mitigation strategies to refine the accuracy of snRNA-sequencing data interpretation. It remains that the best way to minimise technical effects such as doublets and ambient RNA is by careful implementation and optimisation of the experimental pipeline, but these tools allow us to make full use of datasets even when the quality is not perfect.

In differential expression (DE) analysis within snRNA-sequencing, a major flaw arises from the neglect of intrinsic correlation within cells of the same individual. The current common practice of ignoring this correlation contradicts fundamental statistical assumptions, often leading to erroneous conclusions. This analysis complication is exacerbated by the fact that single-cell experiments commonly utilize cells from the same genetic and environmental background. Moreover, aggregation methods, while more conservative, provide an alternative means to control for zero-inflation and within-sample correlation, especially in unbalanced datasets which are frequent in snRNA-sequencing studies. Proper and careful statistical modelling are crucial to draw valid inferences in DE analysis, emphasizing the need for continued development and application of robust analytical tools in single-cell studies (65,137).

The integration of snRNA-sequencing data from different experiments is essential to correct for batch effect which can obscure biological signal. These are one of the biggest obstacles in snRNA-sequencing data. The method of integration we used indicated that the integrated dataset reflects actual biology rather than artificial clusters from technical discrepancies. Samples most often used in case control studies to discover quantitative changes between different groups or conditions plays an important role for understanding the molecular basis of phenotypic variation, and to unravel the genetics of complex diseases and traits causes. Evolution of tools originally developed for handful of samples, have had to expand to integrate 100s.

Furthermore, as single-cell technologies mature, the increasing complexity and volume of data require fundamental changes in data access, management, and infrastructure alongside specialized methods to facilitate scalable analyses. Around two thirds of tools are not

available from a standard centralized software repository (CRAN, Bioconductor or PyPI) and are mostly available only from GitHub(138). Future work includes building this analytical pipeline by using workflow management software such as NextFlow (139). These tools ensure reproducibility as workflows can be executed in the same way across different computing environments. These also allow scalability to allow the processing of large datasets. Parallelisation which means independent tasks can be run simultaneously reducing the time requires, and error handling which saves time and resources (140). Overall, it is important that the computational biology field.

By acknowledging and addressing these challenges, we can continue to leverage the full potential of snRNA-sequencing, enhancing our understanding of the molecular mechanisms underlying health and disease. Standardising approaches will also contribute to the robustness and reliability of results while also keeping in mind data specificity.

2.4.1 Limitations and Future Work

In this study, we focused on classifying broad cell-types rather than examining specific subtypes or regional differences. As we wanted to look at convergent signatures between heterogenous regions and epilepsies which share heterogenous subtypes, we focused on clustering at the major cell-type level, which often have overall physiology or pathology. Major cell-types provide a foundational understanding of the overall organization and function of tissues or organs, and they represent distinct groups with common characteristics that are easier to study collectively. Major cell-types are usually easier to classify and characterize compared to subtypes, which can have more nuanced differences (141). This simplicity can help in establishing basic principles before delving into finer details. Additionally smaller sample sizes in this dataset limited the ability to look at subtypes.

However, future work will focus on examining subtypes, as this could enhance our understanding of the disease as different neuronal subtypes likely contribute in distinct ways to the initiation and propagation of seizures. By studying these subtypes we can identify more specific pathways and mechanisms involved in epilepsy. Neuronal subtypes in particular have distinct gene expression profiles that correlate with their specific functions and roles in the brain (142), these subtypes affect pathology differently and may be involved in different circuits within the brain. In the context of the differential expression analysis, differences between control and epilepsy brains may be obscured when studying broader categories and subtypes may help reveal those more specific differences in neuronal clusters to enhance the resolution and specificity of findings. While collapsing sub-types increases the number of cells tested per cluster, this may not increase power for differential expression as it may mask regions and layer specific affects that may be more likely to be clear when looking at subtypes.

Similarly, collapsing analysis across brain regions and cortical layers also helped to reduce complexity in this analysis as we found that variation between cell-types was greater than variation between brain regions but this comes with similar limitations. This approach allows us to find patterns and fundamental principles that may be conserved across diverse brain regions, however, by combining data from distinct brain regions or cortical layers, we may overlook critical region-specific or layer-specific differences in gene expression.

Particularly in the contexts of epilepsy where specific neuronal subtypes and cortical layers play critical roles. For example, in temporal lobe epilepsy (TLE), CA1 and CA3 regions in the hippocampus are known to be susceptible to neuronal loss, gliosis and excitotoxicity and the dentate gyrus, contributes to regulating excitability and often showing mossy fibre sprouting which are linked to seizure propagation (143). Studying the hippocampus

separately may give more of an insight into initiation of seizure activity. The L2/L3 neurons in the cortex facilitate the propagation of seizure activity across cortical layers, whereas other layers have inhibitory properties crucial in suppressing epileptic discharges (78). Therefore, collapsing analysis across brain regions and cortical layers, risks masking region-specific and layer-specific effects which may be important for understanding the biological relevance of these differences in epilepsy. This is particularly important in neuronal cells compared to other cell-types such as astrocytes where diversity is not as high (142). Inhibitory neurones in particular are crucial drivers of epilepsy, which regulate neuronal excitability and maintaining the balance between excitation and inhibition within neural circuits. Understanding the specific roles and vulnerabilities of these inhibitory neurons would give a more granular understanding of neuronal function in epilepsy(144). Dysfunction or loss of specific subtypes of inhibitory neurons can contribute to the initiation, propagation, and severity of epileptic seizures. Various subtypes of inhibitory interneurons, such as parvalbumin-positive (PV+), somatostatin-positive (SST+), and vasoactive intestinal peptide-positive (VIP+) cells, exhibit distinct roles in modulating network activity. PV+ interneurons, known for their fast-spiking properties, are critical for synchronizing network oscillations and suppressing aberrant neuronal firing. SST+ interneurons regulate the timing and duration of inhibition and have been shown to contribute to epileptogenesis, while VIP+ interneurons modulate disinhibition and synaptic plasticity (145).

In this study, methods were used to correct for ambient RNA expression in the data. Ambient RNA correction methods enhance the accuracy and reliability of RNA sequencing data by reducing background noise and artifacts introduced by extracellular RNA and enables clearer interpretation of gene expression patterns by focusing on intracellular RNA originating from cells of interest rather than non-specific RNA contamination. While this is beneficial in

removing background noise, it could also eliminates genes that should not be present (146,147).

A limitation of these methods is that they may also introduce some bias and could lead to overcorrection, inadvertently removing genuine biological signals. For example, one way of testing the presence of ambient RNA is by looking for genes typically associated with specific cell-types that are not expected to be present in other cell-types e.g. the presence of neuronal markers in glia. However, this is likely an oversimplification as neuronal markers can exhibit dynamic expression patterns across diverse cell-types and non-neuronal cells may transiently or under certain conditions express neuronal genes. These may be remnants of developmental programs, responses to environmental cues, or through cellular reprogramming events (148).

These occurrences can contribute to functional diversity within cell populations and may play roles in adaptive responses or pathological conditions. Methods may not address this and overcorrect the data, so it is important to understand cellular diversity and function when applying these methods. Different ambient RNA correction algorithms or methods may introduce biases depending on their underlying assumptions and computational strategies. Some methods may disproportionately affect certain cell-types or clusters, skewing the relative proportions observed in the data.

In this data, we observe that the annotation of some cells changed after correction for ambient RNA, which subsequently affects the cell-type proportions before and after correction. This highlights the importance of determining which proportions more accurately reflect the true cellular composition and align with established tissue composition from literature and experiments. Future work will focus on investigating the accuracy of these corrected proportions to ensure that they are representative of the actual biological context. This may

involve comparing the corrected data to known cell-type distributions from other studies, or using alternative experimental validation techniques the proportion of specific cell-types. This may particularly be an issue when looking at subtypes which are more homogenous, and correcting for ambient RNA can reduce the observed transcriptional variability within cell populations.

This study uses pathway overrepresentation analysis identify enriched biological pathways or functions among differentially expressed genes to help add biological insight to the data. However these methods come with several limitations that are important to consider. Firstly, these methods rely on predefined gene sets and pathways, which may not capture novel or context-specific biological mechanisms that are unique to a particular experimental condition or disease state, such as epilepsy. Additionally, pathway databases can be biased towards well-studied pathways, potentially overlooking less characterized pathways that could be also relevant. Pathway analysis does not always consider the interactions and crosstalk between pathways, which are crucial for understanding complex biological diseases like epilepsy. Other methods which could be used to add functional annotation to genes include co-expression and network analysis. Network based methods like WCGNA (149) could provide more complementary insights into pathways underlying epilepsy and how genes are connected. Instead of focusing solely on predefined pathways, network-based approaches consider the interactions between genes within biological networks. This can reveal properties and functional modules that are not apparent from pathway enrichment alone. Additionally, methods such EWCE (Expression Weighted Cell-type Enrichment) could also be used to look at cell-type specificity. This is a statistical tool that uses gene lists such as disease associated gene lists from differential expression analysis to test for the enrichment of cell-types (150).

Chapter 3

Single-nucleus RNA-sequencing of mouse models of Epilepsy

3.1 Introduction

Understanding the intricate cellular and genetic landscape of epilepsy is key to the development of effective therapeutic strategies. The complex neuronal circuits and heterogeneity of the disease, makes epilepsy research challenging in comparison to other brain diseases. Advances in single-cell technologies have shed light on the distinct cell-types that constitute the brain and their unique roles in many diseases. This shift from studying brain tissue to dissecting cell populations is crucial in understanding this heterogeneity in epilepsy. There have been a limited number of studies to date that characterise epilepsy at a single-cell resolution (142,151,152), often focusing on specific subtypes of epilepsy and only including relatively modest sample numbers.

One third of patients with epilepsy have refractory epilepsy, which means they continue to have seizures even with medication (47). Developing drugs that better target the underlying mechanism of disease, and not just minimise seizures may enable the development of more effective treatment. This requires a more comprehensive understanding of epilepsy in a cellular context and a better understanding of the underlying pathways driving the disease.

3.1.1 Gene expression and signature reversion

Gene expression changes in epilepsy have been extensively studied by using bulk RNA-sequencing of brain tissue to assess the change across all cell-types; neuronal and non-neuronal. Several transcriptomic changes are known to take place in epilepsy including changes that lead to dysregulation of ion channel function, altered immune response, neurotransmitter imbalances, cell death and synaptic plasticity (153). However, the averaging of the signal in bulk sequencing can lead to an inadvertent masking and dilution of information. The granularity that can be gained by using single-cell transcriptomics can help to uncover differences in cellular function that haven't previously been identified.

Identifying single-cell gene expression signatures as drivers of epilepsy can facilitate drug discovery by targeting genes that have the potential to shift gene expression in a particular cell-type back towards a healthy state. This is based on the principle of "signature reversion" which aims to identify drugs that induce a transcriptional response that is counteracting that of the disease. This concept has been successfully used in other diseases such as multiple sclerosis (MS) and Alzheimer's Disease as well as epilepsy (154). Implementing this approach at a single-cell level could facilitate the discovery of novel strategies that modify the disease, by shifting gene expression profiles associated with the condition. In this study we look at three different mouse models of epilepsy and the transcriptomic changes that are induced. Using this comprehensive approach with multiple models, we aim to identify cell-types and pathways that are independent of the specific epilepsy type or underlying cause, and that converge across different types of epilepsy. This might increase the chances of prioritising drug targets that have more translatable potential for human disease.

This builds on previous work within the Johnson Lab, which used a gene regulatory predictive framework to identify a gene target Csf1R, a cell membrane receptor which when inhibited demonstrated the ability to restore a specific gene module's expression towards

health and which showed efficacy in animal models of epilepsy. Here, we aim to extend this approach to the use of single-cell gene expression data offering a unique approach to drug target identification for epilepsy and which has broad implications in many diseases (59).

3.1.2 Drug repurposing

A significant challenge in drug development for epilepsy is the failure of progressing targets into clinical trials. This is largely due to the lack of preclinical experimental models that adequately model the disease (41). Drug repurposing, also known as drug repositioning or reprofiling is a strategy that aims to use existing drugs for a purpose different than to the one they were originally developed for. This strategy can accelerate drug development as drugs do not have to be created from scratch which not only lowers costs but removes many long and tedious steps in the drug discovery pipeline. As safety profiles of the drugs are well established, this minimises safety testing and reduces risk which can speed up the approval process. This is a valuable approach in drug discovery that provides a risk- and cost-effective solution to the difficult drug discovery process.

3.1.3 Mouse models of epilepsy

There are a wide range of animal models that have been used for epilepsy, as there are many animals with a central nervous system and therefore can experience seizures. Mice have been particularly popular as they are very easy to access and have low costs. They also exhibit a significant degree of genetic similarity with humans and are easy to manipulate (155,156). Epilepsy can be induced in mice using chemo-convulsant such as pilocarpine or Kanic acid, genetic manipulation such as gene knockouts or activations, or by electrical stimulation. Each of the different models have their own advantages and limitations. Many studies have shown the validity of mice as pre-clinical models of human epilepsy (157). As well as

standardization, using mouse models in drug discovery is advantageous as it allows for higher throughput and lower costs which provides a tractable first step in the drug discovery process. Mouse models provide further advantages in drug development such as wider availably of tissue, smaller amounts needed in testing of compounds and shorter treatment duration. They also replicate neurobiological processes such as neuronal loss, neurogenesis, inflammation, gliosis, and molecular reorganisation often associated with human epilepsy.

In this study we use models of Temporal Lobe Epilepsy and Focal Cortical Dysplasia. Mouse models of temporal lobe epilepsy largely depend on the use of status epilepticus models (SE). The characteristics of TLE can be reproduced in the mouse model by inducing SE. These characteristics include; an epileptic foci, an "initial precipitating injury", a latent period, and the presence of hippocampal sclerosis (158).

The mouse models used in this study were Pilocarpine induced post-status epileptic model of TLE (PILO), a Sustained Status Epilepticus model (SSSE) temporal lobe epilepsy, and a genetic model of Focal Cortical Dysplasia (FCD).

3.1.3.1 Pilocarpine induced post-status epileptic model of TLE (PILO):

This is a chemo-convulsive model which involves a single injection of the drug Pilocarpine. This induces recurrent and spontaneous epileptic seizures in the mouse. Pilocarpine is a potent muscarinic agonist that promotes change that resemble TLE. Similarly to TLE, Pilocarpine will induce first a (a) an acute period where there is progressive accumulation of limbic seizures leading to status epilepticus (b) a silent period with a normalization of EEG and behavioural manifestations and (c) a chronic period characterized by spontaneous recurrent seizures.

3.1.3.2 Self Sustained Status Epilepticus (SSSE):

This is an electrical stimulation model which involves the implantation of an electrode into the hippocampus of the mice to induce a prolonged electrical current. These mice develop

seizures within the first month that are consistent with temporal lobe epilepsy. Brain tissue is then collected at 16 weeks post-SE (159). SSSE is a relatively newer model in comparison to chemo-convulsive models but can provide more opportunity and options. In this study two models of TLE were used to provide novel insights about the molecular mechanisms relevant to epilepsy.

3.1.3.3 Focal Cortical Dysplasia (FCD):

This is a genetically manipulated model of FCD, a developmental disorder which is associated with epilepsy. It is characterized by the presence of cortical malformations which are associated with increased seizure activity and intractable epilepsy. There are several types of FCD which are classified based on distinct histopathological features. Type IIa FCD is characterised by the presence of dysmorphic neurons. There is general agreement that FCD type IIa arises from the overexpression of the mTOR pathway resulting from somatic mutation in mTOR regulatory genes. In the mouse model of FCD used for this thesis, the mice were electroporated *in utero* with plasmids that contain a mutant *Rheb* gene, an activator of the mTOR pathway. The mTOR signalling pathway has a causal link with FCD. Clinically, the mice display strong, severe seizures, and pathologically, the mice manifest. This mouse model most closely resembles FCD type IIa (160). Characteristics of FCD type II include the generation of focal malformations, cortical dyslamination, and abnormal morphogenesis of neurones. FCD type IIa which displays an abnormal type of neuron called dysplastic neurones does not have other abnormal cell-types associated with FCD called balloon cells whereas FCD type IIb does have balloon cells (See Chapter 5). Unlike the models of TLE where the hippocampus is the most severely affected brain region, FCD affects the prefrontal cortex and malformation can be present over 50-90%- of the cortex (48).

3.1.4 Aims & Objectives

This chapter aims to identify and characterise the cell-types and pathways underlying the epilepsy disease state by leveraging snRNA-sequencing data from mouse models of epilepsy. Since gene expression may vary both as a consequence and a cause of a disease, we hypothesised that expression changes that have a causal (upstream) relationship to epilepsy may be enriched for genes which confer risk to epilepsy. To this end, gene expression changes in epilepsy will be integrated with summary data from epilepsy GWAS and rare variant analysis from Whole Exome Sequencing (WES) to find cell-types that are enriched for rare and common epilepsy heritability. Cell-type specific gene expression changes in epilepsy will then be integrated with gene perturbation data from CRISPRi/CRISPRa screens, and using a causal inference framework, used to predict potential therapeutic targets for epilepsy. These will then be intersected with databases of known drugs which have the potential to be repurposed for epilepsy. The specific objectives of this chapter are therefore:

1. Identify differentially expressed genes in the epileptic brain by comparing single-cell gene expression profiles between the disease and control samples across different animal models of epilepsy.
2. Uncover dysregulated pathways associated with epilepsy by testing for enrichment of known pathway genes within the differentially up- and down- regulated genes in epilepsy.
3. Identify cell-type/s that are enriched for common and rare epilepsy heritability by using epilepsy GWAS and rare variant analysis from WES to find cell-types that are enriched for rare and common epilepsy heritability.
4. Identify potential drug targets for epilepsy treatment by utilizing a publicly available perturbation dataset from CRISPRi/CRISPRa screens and a causal reasoning framework to predict potential causal gene targets.

5. Use drug databases to identify existing drugs that have potential to be repurposed for the treatment of epilepsy.

3.2 Methods

3.2.1 Experimental mouse models of Epilepsy

All mouse models were provided by UCB Pharma, and received in two batches. For each model, samples are provided from the epileptogenic zone from the epileptic mice as well as anatomically matched control brain tissue from littermate control mice.

All *in vivo* experiments were performed according to the National Rules on Animal Experiments in Belgium and to the guidelines of the European Community Council Directive. Analyses were conducted under Imperial College Research Ethics Committee approval.

3.2.1.1 Pilocarpine induced model of temporal lobe epilepsy

In both batches of pilocarpine mice, status epilepticus (SE) was induced by UCB pharma previously describe methods (161,162). Briefly, male mice were given a single injection of pilocarpine intraperitoneally (ip) (300 mg/kg), which induces generalized clonic–tonic seizures 10-45 minutes after the injection which leads to SE. SE is then interrupted after 3 hours by an injection of diazepam (10 mg/kg ip). Surviving mice show spontaneous recurring seizures within weeks and continue to have epilepsy for several weeks, before they are sacrificed.

3.2.1.2 Self-sustained status model of temporal lobe epilepsy

This model has been generated and described by UCB (159). Briefly, mice are implanted with a bipolar electrode into the right ventral hippocampus of the mice to induce a prolonged electrical stimulation. An inactive electrode was also implanted into the contralateral side. The surgery was conducted using anaesthesia and the electrodes held in place using assembly dental acrylic. The mice are then monitored using video electroencephalography

and the animals developed SSSE following a 90min stimulation. The animals developed SRS within the first month. To establish electrical current needed to induce a seizure exceeding ten seconds each mouse was given increasing electrical current (50 Hz, 1-s duration, 1-ms alternating current pulses). The development of SSSE was confirmed using an EEG trace. SSEE is then interrupted after 3 hours by an injection of diazepam (5 mg/kg ip).

3.2.1.3 Focal cortical dysplasia model

This FCD model was generated by UCB as previously described (160,163). In summary, mTOR activity is increased in the layer 2/3 neurons of the medial prefrontal cortex which causes tonic-clonic seizures. Using *in utero* electroporation, the activity of mTORC1 was increased by using *in utero* electroporation to express a plasmid which encoded a constitutively active form of the Rheb gene ($Rheb^{CA}$), which canonically activates the mTOR signalling pathway. The electroporation targeted a subset of neural progenitors at day E15 in the cortex which allows preferential targeting of neural progenitor cells that will generate layer 2/3 pyramidal neurones in the PFC (more details about this model can be found in chapter 5).

3.2.2 Tissue sectioning

Brain tissue samples were provided by UCB Pharma fresh frozen and processed at Imperial College London by Daniel Clode (Batch 1) and Jeong Hun Ko (Batch 2) (Table 1).

- For the PILO model – tissue was taken from the L&R hippocampus and combined. Tissue was also provided from anatomically matched control brain tissue from littermate control mice.
- For the SSSE model – tissue was taken from the hippocampus where the epilepsy was induced (right) and also the contralateral side (left). Matched tissue was also provided

from littermate control mice that were implanted with an inactive electrode, in the right side of the brain.

- For the FCD model – tissue was taken from the region FCD lesion (where the plasmids were transfected) as well as the contralateral side of the FCD brain. Sham mice were also provided that were transfected with control plasmids that did not contain *Rheb*.

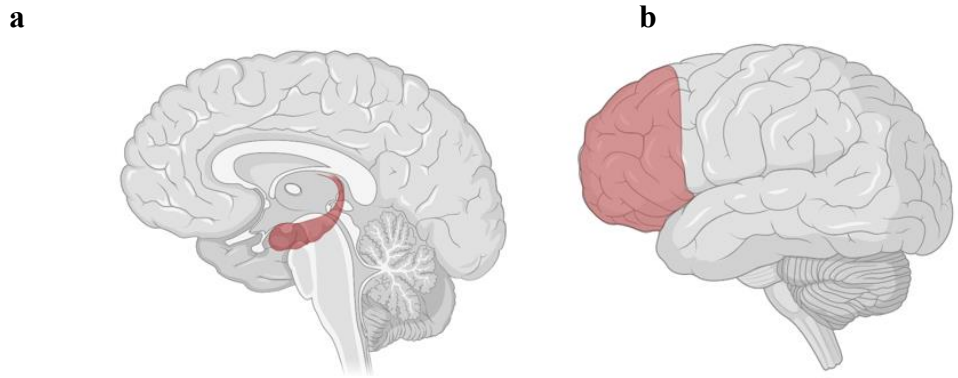


Figure 3.1 – These figures illustrates the brain regions used in this study. The regions highlighted in red represent the tissue sampled in this analysis. A) The structure shown in this figure is the Hippocampus which is located deep within the medial temporal lobe of the brain. B) The structure shown in this figure is the Pre-Frontal Cortex and is located in the frontal lobe of the brain directly behind the forehead.

3.2.3 Isolation of Nuclei

The nuclei isolation protocol that was used was adapted from the 10x genomics protocol (164), and the Krishnaswami et al protocol (100), except for a change to the douncing step which was increased to allow for better homogenization of the tissue.

Briefly, the tissue is homogenised in a pre-cooled 2 mL dounce homogeniser in 800 µL HB suspensions, using 5 strokes of the loose pestle and then 10–15 strokes with the tight pestle. The homogenate is then filtered with a 35 µm strainer cap and centrifuged for 8 mins at 100g. An additional clean-up step (density gradient centrifugation), was used with centrifugation of the layered nuclei/29% iodixanol solution at 13,000*g* for 40 min at 4 °C. The supernatant was then removed and washed with PBS buffer (PBS + 1% BSA + 0.2 U/ml RNAseIn), filtered through a BD Falcon tube with a cell strainer cap, centrifuged at 500*g* for 5 min at 4 °C and washed again. Nuclei were counted using an LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems, L20001) and using Acridine orange dye to stain nuclei. Nuclei encapsulation was performed using the 10x Genomics Chromium Single-cell Reagent Kits Protocol (chemistry: Single-cell 3′ v2).

3.2.4 Data Generation

snRNA-sequencing data was generated using the 10X Single-cell Next GEM Chip targeting a minimum of 6,000 nuclei per sample in batch 1 and 10,000 nuclei per sample in batch 2 and libraries were prepared using the Chromium Single-cell 3′ Library and Gel Bead v3 kit according to manufacturer's instructions. Briefly, this involves 8 cycles of cDNA amplification and 14 cycles of final indexing PCR. cDNA concentrations were measured using Qubit dsDNA HS Assay Kit and cDNA and library preparations were assessed using the Bioanalyzer High-Sensitivity DNA Kit. cDNA libraries were sequenced using the Illumina NovaSeq 6000 system at a minimum sequencing depth of 30,000 paired-end reads per nucleus on 4 flow cells. The sequenced reads were processed using Cell Ranger (v 3.0.2). Library preparation and sequencing was performed externally by Source Bioscience. The Illumina sequencers generate raw base call (BCL) files, which were then demultiplexed using *cellranger mkfast*q from 10x's Cell Ranger software which generates FASTQ files from the

BCL files. This function uses Illumina's bcl2fastq function with a few modifications to support 10x Genomics libraries.

3.2.5 Mapping

Sequenced reads (FASTQ) were aligned to the genome using Cell Ranger (v 5.0.1) and mapped to the mm10 (GENCODE vM23/Ensembl 98) reference genome using gene annotations from Ensembl v98.After the sequence quality of the raw reads is checked, as well as the GC content, adaptors, k-mers. Pre-built references were downloaded from the Cell Ranger website. The reference transcriptome used in the mapping downloaded from https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest/. Reads were mapped to primary transcripts including introns as well as exons using the --include-introns option in Cell Ranger, as this has previously been shown to detect a greater number of genes per nucleus. Counts were quantified as UMI counts, across each of the 85 sequenced samples.

3.2.6 Data Processing and Quality Control

A detailed description of the computational pipeline used in this thesis can be found in Chapter 2. Briefly, processing and quality control of single-cell data involve a series of steps to ensure the reliability and accuracy of the dataset:

1. Ambient RNA is first removed from the raw count matrix using the tool CellBender. This is free floating RNA which comes from damaged cells and get encapsulated along with cells in the microfluid droplets. This can cause noise in the data and skew the gene expression signatures and interfere with downstream analyses (147).

2. Doublets are then removed using DoubletFinder. Doublets or multiplets occur when two or more cells are encapsulated within a microfluid droplet. This also inflated gene expression profiled and leads to misleading data, particularly when the doublet compromises of heterogenous cells (110).
3. Further filtering is applied to the data using Seurat (v4) (117) to filter out low-quality cells which have very low count and uninformative genes which are not expressed in many cells. These are filtered out to improve data quality and remove noise.
4. Cells with high mitochondrial content (>5%) are removed as high mitochondrial content is a sign of poor quality or damaged cells (65).
5. Seurat is used to normalise the data by applying the default normalisation method which involves log normalisation and then scaled to factor of 10,000. Normalization methods account for differences in sequencing depth between cells to ensure comparability (117).
6. When analysing datasets that are generated across multiple experimental batches or conditions, batch effects can introduce unwanted variations. Batch correction and integration is applied using Harmony to correct across batches and LIGER to correct across condition (115,165).
7. Dimensionality reduction is applied, first by principal component analysis (PCA) and then Louvain clustering in Seurat to cluster the cells. UMAP are then used to visualize the data and map high-dimensional single-cell data into a two-dimensional space.
8. Clusters are then annotated to cell-types using known marker gene. First the unique markers per cluster are identified and a hypergeometric test is applied to test enrichment of these markers against list of know cell-type markers (Wang) (123). This process allows for the identification of specific cell-types within the dataset. The

annotations are then confirmed by plotting heatmaps of known marker genes to visualise the expression per cluster.

9. Differential expression analysis performed using DESeq2. First the expression values per cell per gene are summed across samples to create a pseudo-bulk expression matrix. Pairwise differential expression is applied in case and control in each model to identify genes that show significant expression differences in cell populations in different disease states. This approach has been shown to be more accurate in analysing single-cell data, however can produce more conservative results (125,166).

3.2.7 Functional enrichment analysis

The R package WEB-based Gene SeT AnaLysis Toolkit *WebGestaltR* (0.4.6) (127) was used to perform over representation analysis for Gene Ontology (GO)(128) databases biological process, molecular function and cellular component. The DEGs from each pairwise comparison from each cell-type were tested for enrichment of the GO terms, against the background set of genes which was the number of cells expressed in each cell-type. The default WebGestalt parameters were used. These include GO terms were filtered by size to include terms with a minimum of 10 genes and a maximum of 1000 genes. Genes (see results). P-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) method (129). Terms were considered significantly enriched with an FDR value of >5%. The GOSemSim R (130) package (v 2.16.1) was used to measure of semantic similarity (measure = “Wang”) and the rrvgo R package (v 1.1.4) was used to reduce terms as described in Feleke et al. (112).

3.2.8 Enrichment of mendelian epilepsy genes

The PanelApp (https://panelapp.genomicsengland.co.uk/panels/) is a database of 340 (as of 2023) gene panels covering a wide range of diseases and representing nearly 50% of Online Mendelian Inheritance in Man (OMIM) phenotypes. The gene list used in this analysis was from the panel "Early onset or syndromic epilepsy" which was merged from panels for Epileptic encephalopathy, Familial Focal Epilepsies, Familial Genetic Generalised Epilepsies and Genetic Epilepsies with Febrile Seizures. This panel consists of 792 genes which have evidence of association with epilepsy. These are split by a traffic light system into *Green*, *Amber* and *Red*, depending on the level of evidence. For this analysis we have only used the *Green* genes (420) which have a high level of evidence for disease-gene association and are authorised to be used for genome interpretation by Genomics England. These lists are regularly curated and signed off by NHS Genomic Medicine Service (167).

A Fishers's Exact test was used to test for statical enrichment of the overlaps between the differentially expressed genes in each cell-type in each pairwise comparison. This is a test which is used to determine the significance of association between two categorical variables. A contingency table is created which displays the frequencies or counts of observations for each of the two variables. This is a hypergeometric test and a p-value is produced which represents the probability that the gene lists are enriched. The significance level used is 5%.

3.2.9 Association of heritability

Heritability (h2) is the proportion of variance of a trait that can be attributed to genetic factors which is the additive effect of common genetic variants. We can use SNP-base heritability analysis to associate cell-types to heritability of epilepsy. Cellular heterogeneity may play a part in shaping how genetic variations manifest phenotypically.

3.2.9.1 GWAS dataset

The GWAS dataset used in this analysis is the most recent epilepsy GWAS published in 2022 by the ILAE (33) which comprises of a total 29,944 of deeply phenotype epilepsy cases and 52, 538 controls primarily of European descent (92%). 4.9 million SNPs were tested.
The cases were recruited from tertiary referral centres and of the cases 16,384 people had focal epilepsy (FE) and 7,407 people had GGE classified by neurologists. The meta-analysis was performed separately for a) 'all epilepsy' b) FE and c) GGE. There were 26 genome-wide significant loci in total 4 of which came from the 'all epilepsy' category, and 22 loci from the GGE category. There were no genome wide significant hits for the FE category. The summary statistics from this GWAS were used for the following LDSC analysis. The summary statistics include key results from the analysis including the SNP, and its chromosome and position, the association p-value, the effect size (usually as Beta coefficient or Odds Ratio), the standard error, z-score and the sample size.

3.2.9.2 s-LDSC

Stratified LDSC (Linkage Disequilibrium Score regression) (v 1.0.1) is a method designed for partitioning the heritability from complex traits using GWAS summary statistics. Genome-wide SNP heritability can be partitioned across categories of annotations in order to attribute disease to specific categories. It can be applied to cell-type specific sets of genes to identify disease relevant tissues and cell-types across diseases. In this analysis annotation categories of genes were defined as the genes differentially up- and down- regulated genes in each cell-type in each pairwise comparison in the different mouse models.
Statisfied LDSC is based on the principle that genetic variants are often in Linkage Disequilibrium (LD) meaning they are in close proximity to each other on the genome and will be inherited together. s-LDSC accounts for LD structures which means the association of

a SNP includes the effects of all other SNPs in LD as SNPs tend to be in LD blocks. It is important to account for this as SNPs with higher LD are likely to have higher association as they may tag multiple other SNPs. The window size around the gene bodies is defined as 100kb in this analysis.

s-LDSC controls for annotations in the baseline model which is comprised of 53 different types of genetic architecture such as functional regions and histone marks. The baseline model is constructed using a reference panel of genetic variants and captures the LD structure of the genome, HapMap Project Phase 3 (HapMap3) SNPs and 1000 Genomes Project Phase 3 European population SNPs were used for the reference panels and only SNPs with MAF >5% were partitioned. This analysis also excludes the MHC region (chr6: 25,000,000–34,000,000, GRCh38) which exhibits complex LD patterns which do not meet the assumptions of LDSC. The categories tested are constructed using binary annotations where SNPs overlapping a category are assigned a value of 1 and SNPs not in a category are assigned a value of 0. The same SNPs are used in constructing the annotation categories as were used for building the baseline model. The trait summary statistics are then regressed against the baseline model which is testing whether the observed associations are greater than what would be expected under the baseline model. In this analysis where the aim is to identify cell-types critical to disease heritability, we are testing whether an annotation category significantly contributing to trait heritability under the baseline model. In this analysis LDSC was applied to test whether the cell-type specific genes from the differential expression analysis showed enrichment for epilepsy heritability, which might imply a causal role in epilepsy. This is based on the presumption that for cell-types causally involved in a complex trait, have gene expression that predicts heritability; however the gene expression may be compensatory.

The summary statistics used in this analysis are the ILAE GWAS studies conducted in 2022, which included individuals of European ancestry as present in the LD reference panel. Significant association suggest that the gene set is enriched for genetic susceptibility to a disease or trait beyond the random expectation.

For all stratified LDSC analyses, coefficient p-values are reported. A regression coefficient is calculated and converted into a p-value using the R function pnorm() and multiple testing correction is performed using Benjamini-Hochburg. These p-values test whether the regression coefficient of an annotation is significantly positive, indicating that the annotation positively contributes to trait heritability, conditional upon the LDSC baseline model. It is noteworthy however, that compensatory, homeostatic gene expression changes that seek to mitigate the occurrence of seizures, may also be enriched for epilepsy heritability, but with different directionality, to changes that contribute increased risk for seizures, and that functionally, these pathways may also overlap with pathways perturbed in epilepsy (i.e., reduced and increased liability to seizures are equal).

3.2.9.3 EPI25K dataset

This dataset is a whole-exome sequencing (WES) case-control study which leveraged familial cases of epilepsy from the EPI25 Collaborative. The EPI25 Collaborative is an ongoing effort to collect data from patients from three major classes of epilepsy; Developmental and epileptic encephalopathies (DEE), generalized genetic epilepsy (GGE) and NAFE (non-acquired focal epilepsy). DEE is a severe form of epilepsy that begins in early life and causes intractable seizures and developmental disability. It is usually monogenic (autosomal dominant, recessive and X-linked) and many de novo variants have been identified for DEE. GGE and NAFE are more common types of epilepsy and are

generally less severe. GGE is characterized by seizures that involve both sides of the brain and NAFE affects a localised region. They account for 20-40% of epilepsies and are generally thought to be a result of polygenic inheritance, with some monogenic forms in a minority. This WES study aimed to find overlapping and distinct ultra-rare coding variants for these epilepsies by looking at burden at the gene-level and in candidate gene sets. This aims to look at the impact of rare genetic variants across epilepsy syndromes.

The DNA was collected from patients with epilepsy from 37 sites in Europe, North America, Australasia, and Asia with a total of 13,487 cases and matched 15,678 controls. The numbers of cases were GGE (n = 7,489), NAFE (n = 7,489), and DEEs (n = 2,007). Epilepsies used in this study were diagnosed based on the criteria consistent with the ILAE classification. 'Ultra-rare' variants are defined as those not seen in any non-neuro gnomAD population. Case-control comparisons were performed separately for each of the epilepsy disorders (DEEs, GGE, and NAFE) and for all the epilepsies combined. Collapsing analysis was performed to summarise the data into gene-level p-values using burden tests across the entire exome, to estimate the effect of rare, deleterious protein-coding variants in epilepsy. Briefly, a burden test is a statistical association analysis used to test the cumulative impact of multiple rare genetic variants within a gene or defined set of genes on a particular trait. The case-control status is regressed on the classes of ultra-rare variants aggregated across a target gene set.

Six key functional categories of coding variant were used in this analysis: Synonymous, Protein Truncating Variants (PTVs), missense, damaging missense, intolerant missense and damaging and intolerant missense. These are denoted as Synon, PTV, EMiss, DamMiss, EInt, and EIntDam in the analysis respectively. Protein-truncating variants (PTVs) were filtered with loss-of-function transcript effect estimator (LOFTEE), "damaging" missense variants

were filtered with REVEL (168) filter ≥ 0.5, and "intolerant" missense variants were filtered based on a missense tolerance ratio (MTR) filter of ≤ 0.78 which represents the most intolerant quartile of all regions in the exome. This allows for the assessment of multiple genetic variants which may have low allele frequency to be combined for greater significance, which is particularly useful when looking at rare variants.

3.2.9.4 GSA-SNP2

GSA-SNP2 (Genome-wide SNP-based pathway analysis) was used to perform gene-set enrichment analysis using the EP25k rare-burden analysis. This is a computational tool that uses summary level p-value data to test for the enrichment of pathways and gene sets. It does this by using a modified Z-statistic test and comparing the observed distribution to the expected distribution under the null hypothesis of no association. Each gene set is assumed to be a random collection of genes, therefore having a normal distribution. The modification of the Z-statistic is implemented to model the size of the gene set, and increase the power of the analysis. The tool also provides a type I error control by adjusting for the SNP counts in each gene. This method also adjusts for LD by removing adjacently located genes within the same gene-set that have a high positive correlation (>0.5) in the 1000 genomes dataset (on average this occurs in <1% of gene sets) (169). The following parameters were used:

--minset 5 *–maxset 2000* to filter out gene-sets <5 and >2000.

The inter-gene genotype correlations used were from European decent and the SNP-gene mapping data was from version GRCh37(hg19) to match that of the data.

3.2.10 CRISPR brain analysis

3.2.10.1 Dataset

The CRISPR brain dataset is a publicly available dataset made available by the Kafman lab and can be accessed via https://crisprbrain.org/. It contains data from functional genomic screens from human iPSC differentiated cell-types; Glutaminergic neurones and Microglia. The data used in this study is from the CROP-seq experiment. CROP-seq (CRISPR droplet sequencing) is a method that combined CRISPRi and CRISPRa with snRNA-sequencing to study gene function. They performed CRISPRi using 184 genes and CRISPRa using 100 genes that well curated from the DisGeneNet database and literature based on genes that were identified as hits in previous genome-wide screens for neurogenerative disease. CRISPRi is a technique used to silence the expression of a gene of interest using the CRISPR-cas9 to target the gene and inactivate it. CRISPRa uses CRISPR-cas9 to activate the expression of the gene. These methods were used to perturb the genes in glutamatergic neurones and microglia to provide information about gene function and cellular responses. For each gene perturbation, differentially expressed genes were calculated by comparing the perturbation cells to cells containing non-targeting control sgRNAs with t-tests using the diffxpy package in Python. Approximately 21,000 CRISPRi neurons and 21,000 CRISPRa neurons were used, with 48,000 mean reads per cell and 4,341 median genes per cell for CRISPRi and 36,000 mean reads and 3,100 median genes per cell for CRISPRa (170–172)**.** For each gene perturbation, we identified the genes differentially expressed for a at FDR value of < 0.05 and split the set of genes into those that are up regulated and down regulated by the perturbation.

3.2.11 Enrichment of DEGs

To predict drug targets that may modify epilepsy, a causal reasoning framework was applied similar to that was used in Srivastava et al. (59). The effect of the gene perturbation on

epilepsy was tested by performing a hypergeometric test to test the Enrichment of the up and down regulated DEGs in each case-control comparison against the up and down regulated genes from the perturbations, using only the genes expressed in the epilepsy snRNA-dataset as the background. The perturbations that were performed in the glutamatergic neurones were tested against the DEGs in the Excitatory and Inhibitory neurones, and the perturbations that were performed in the microglia tested against the Microglial DEGs. FDR was calculated using BH correction to account for the number of enrichment tests performed. Genes were considered to have a significant effect on epilepsy if they had an FDR > 0.05. Mouse genes were converted to human orthologues prior to the analysis using the biomarRt package (173). The results were also quantified according to two metrics; Activity and Weight. Activity was calculated as the proportion of genes perturbed by CRISPRi/a which are differentially expressed in epilepsy, and Weight was calculated as the proportion of genes differentially expressed in epilepsy which are perturbed.

1) **Activity**: the proportion of genes perturbed by CRISPRi/a which are differentially expressed in epilepsy
2) **Weight**: the proportion of genes differentially expressed in epilepsy which are perturbed by CRISPRi/a

The Impact was then calculated, this is the sum of the activity and weight. The regulators were sorted according to the sum of the ranks of their activity and weight.

3.2.12 Intersection with databases

3.2.12.1 The International Union of Basic and Clinical Pharmacology (IUPHAR)

The UIPHAR/BPS Guide to Pharmacology database (found at www.guidetopharmacology.org) contains expert curated information about ligands and the molecular interactions with their targets (174). It provides information on pharmacological targets including receptors, ion channels, enzymes, transporters and other protein targets, as well as a detailed description of their structure, function and signalling pathways and physiological roles. This database is regularly updated and curated and is a very valuable tool for researchers in drug discovery, pharmacology and therapeutic development.

3.2.12.2 Prescribable Drugs with Efficacy in Experimental Epilepsies (PDE3)

This database was curated as a resource for drug repurposing research in epilepsy. It was created using literature mining of Medline abstracts to identify existing FDA approved drugs that have the potential to be repurposed for the treatment of epilepsy in humans. The database contains data from published preclinical studies that have evaluated the effect of drugs in rodent models of epilepsy. It comprises of 173 drugs from 500 studies, and can be downloaded from www.liverpool.ac.uk/D3RE/PDE. The database also collects information on drugs, their dosages, the seizure types they were tested against, and the efficacy outcomes observed in the experimental models. This database can be useful for highlighting novel prescribed drug targets for epilepsy which have been tested and approved, and could potentially be repurposed for the treatment of epilepsy (175).

3.2.13 Computational Tools

The computational analysis in this project was performed on the Imperial High Performance Computer from the Imperial College Research Computing Service, DOI: 10.14469/hpc/2232

3.2.14 Code availability

The code used in this analysis can be found at https://github.com/Mybzd/RNA-sequencing_analysis/

3.3 Results

3.3.1 Study summary

Single-nucleus RNA sequencing was applied to 85 samples from three distinct mouse models of epilepsy and corresponding controls as detailed in Table 3.1. These models consist of PILO (n=28), SSSE (n=20) and FCD (n=44). Given that PILO and SSSE are representative models of temporal lobe epilepsy, which most severely affects the hippocampus, tissue samples from these models were specifically sampled from the hippocampus. The FCD samples were sourced from the Prefrontal Cortex, marking the location of the FCD lesion in the mice. Following processing and quality control filtering, a total of 541,051 single nuclei were identified across the 85 samples, which were taken from 71 individual mice. This resulted in an average of 6,365 cells per sample, with an average of 1,699 genes detected in each cell. The samples were prepared and sequenced in two batches, and subsequent analysis integrated these batches into a single graph. Samples were integrated using LIGER (115) to minimise variance arising from sample origin and brain region, and further using Harmony (165) to mitigate batch specific effects (see Chapter 2.2 for more detailed methods).

Individual cells were clustered using Louvain clustering, and cell-type assigned post-hoc to one of seven major cell-types (Figure 3.2a), using a hypergeometric test to test the enrichment of a cell-type specific marker gene set (176) relative to cluster specific markers (see Chapter 2.2). To validate the cluster cell-type annotations, canonical cell-type markers

(177) were used, and their expression visualised in a heatmap (Figure 3.2c). This demonstrated that the clusters expressed markers unique to their designated cell-types, and also demonstrates how cell-type is most accurately assigned by the use of multiple cell-type marker genes rather than reliance on a single canonical cell-type marker gene

BATCH 1

Model	**Pilocarpine (PILO)**		**Self Sustaining Status Epilepticus (SSSE)**			**Focal Cortical Dysplasia (FCD)**		
Sample Type	Pilo	Sham	Implanted SSSE	Unimplanted SSSE	Sham	Transfected FCD	Non-transfected FCD	Transfected Sham
Brain Region	Left + Right hippocampus	Left +Right hippocampus	Right hippocampus	Left hippocampus	Right hippocampus	Prefrontal Cortex	Prefrontal Cortex (contralateral)	Prefrontal Cortex
Sample Number	7	7	7	7	6	7	7	7

BATCH 2

Model	**Pilocarpine (PILO)**		**Focal Cortical Dysplasia (FCD)**	
Sample Type	Pilo	Control	Transfected FCD	Non-transfected Control
Brain Region	Left + Right Hippocampus	Left +Right Hippocampus	Prefrontal Cortex	Prefrontal Cortex
Sample Number	7	7	8	8

Table 3.1 – Overview of samples used in this analysis. These tables provide a detailed breakdown of the brain samples used in this analysis, categorized by two separate batches. Three distinct models were used in this analysis Pilocarpine (PILO), Self-Sustaining Status Epilepticus (SSSE), and Focal Cortical Dysplasia (FCD). This table provides information on the number of samples from each model and the brain region the samples were collected from.

3.3.2 Annotating cell-types

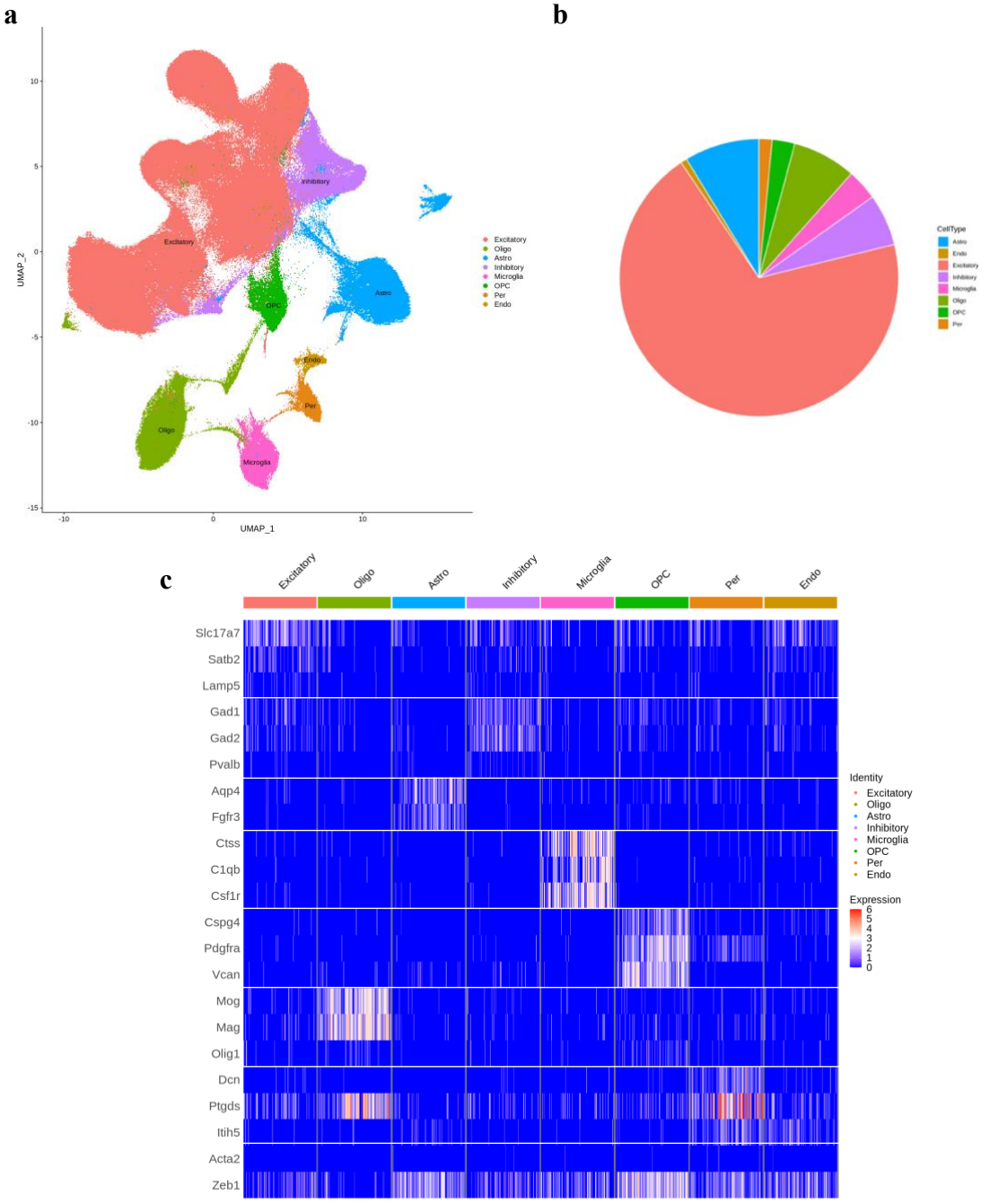


Figure 3.2 – Annotation of cell-types across brain regions. a) Visualization of all annotated cell-type clusters on a joint UMAP graph. Individual nuclei are colored by cell-type. b) Pie chart representing the proportions of each cell-type. The dominant cell-type is excitatory neurons followed by Astrocytes.c) A heatmap illustrating the gene expression levels (down sampled n=200 cells) of canonical cell-type markers across the cell-type clusters.

The cell-types are dominated by excitatory neurons and the proportions of cell-types align wit h the proportions seen in known literature (178). In total there were 375,599 excitatory neuro ns, 32,629 inhibitory neurons, 39,998 oligodendrocytes, 47,071 astrocytes, 19,425 microglia, 14,122 oligodendrocyte precursors (OPCs), and 3,958 Endothelial cells and 8,249 pericytes (Figure 2b). Visualisation of the cell-types by different variables shows that the primary sourc e of variations originated from the cell-types rather than the model (Figure 3.3a),brain regions (Figure 3.3b), or conditions (Figure 3.3c).

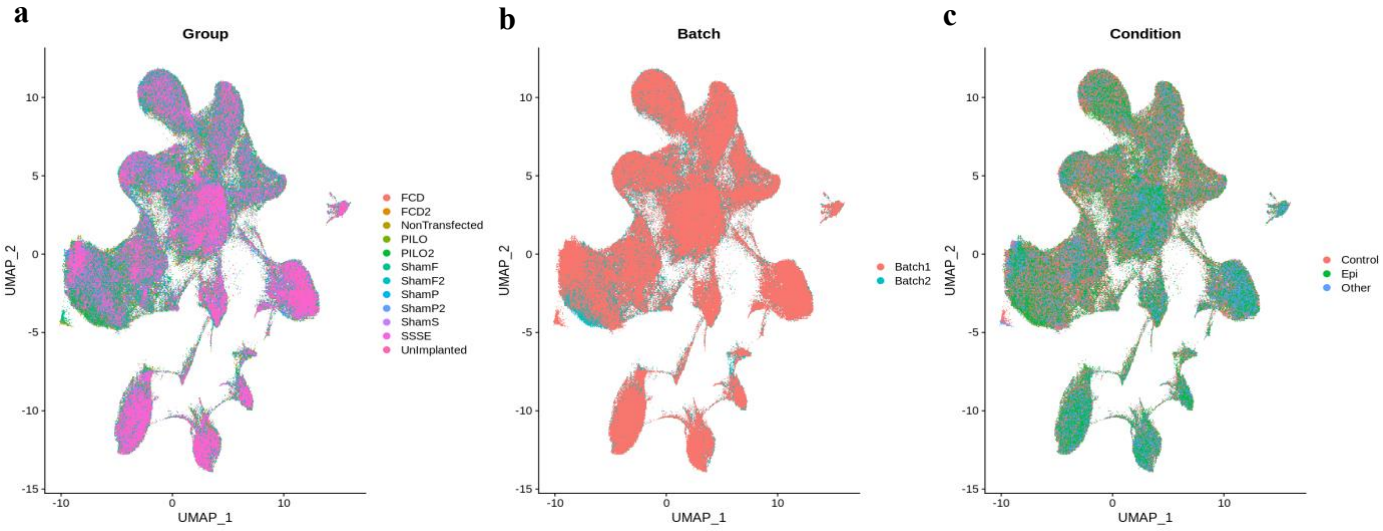


Figure 3.3 – These UMAP plots provide a visualisation of the dataset coloured by a) Model b) Batch c) Condition. The cells are well integrated by all 3 of these variables and show no distinct clustering pattern within group, batch and condition.

3.3.3 Identification of differentially expressed genes in epilepsy

To identify the cell-type specific transcriptional changes occurring in each of the epilepsy models, differential gene expression analysis was conducting using a pseudobulk approach and DESeq2 (125). Differential expression aims to characterize cell-type specific changes in gene expression in the disease. A "pseudo-bulk" approach is used, where gene counts are first aggregated per cell-type per subject (detailed in Chapter 2.2) to avoid the inflated type I error rate (i.e., where non-differentially expressed genes are falsely indicated as differentially expressed) arising from "pseudoreplication" (i.e., where dependent cells from the same individual are treated as independent observations)(179). The comparisons tested were PILO (n=14) vs Sham (n=14), SSSE (n=7) vs Sham (implanted) (n=6) and FCD (n=15) vs Control (n=15) (Table 1).

Principle Component Analysis (PCA) shows a predominant sample separation per cell-type based on brain region (PC1), followed by disease groups (PC2 and PC3) (Figure 3.4).

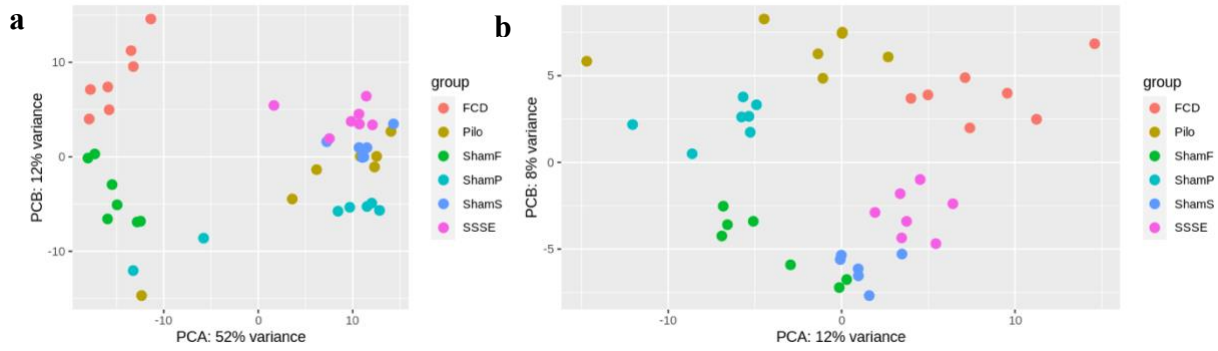


Figure 3.4 – PCA plots to show the sample separation by variance. a) PC1 vs PC2 – samples show separation across PC1 where the cortex samples separate from the hippocampus samples and accounts for 52% of the variation. b) PC2 vs PC3 – samples are split across both PC2 and PC3 where case/control samples separate mainly across PC2 which accounts for 12% of the variation.

For the differential expression analysis, only those genes expressed with a minimum of 3 counts per gene per cell-type per sample were included. Whilst single-cell differential expression analysis based on aggregated counts is a conservative method (minimising the Type 1 error), power to detect DEGs is very much dependent on sample size (as well as number of cells per cell-type – see Power Calculations in Methods). For this initial discovery phase of the analysis, we therefore considered genes as significantly differentially expressed (DE) at FDR of less than 0.1 (rather than the conventional FDR<0.05 threshold) in order to facilitate downstream functional and genetic enrichment analyses. The numbers of differentially expressed genes in each model in each cell-type are shown in Figure 3.5. The results are split into up- and down-regulated genes.

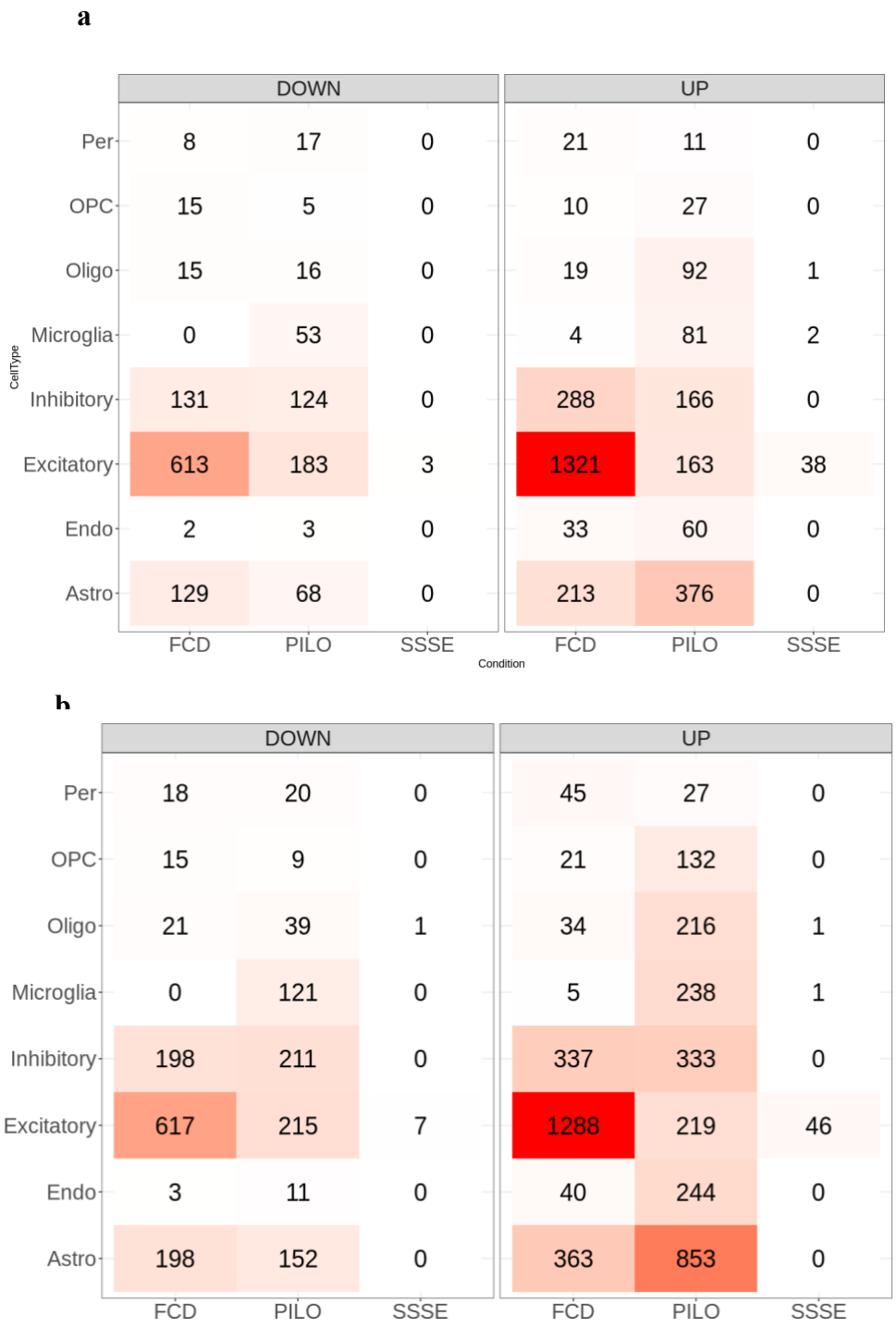


Figure 3.5 – Numbers of cell-type specific differentially expressed genes (DEGs). This heatmap represents the cell-type specific DEGs across the pairwise comparisons in the three mouse models. The right hand panel represents the up-regulated genes and the left hand panel represents the down-regulated genes. Genes are considered differentially expressed at a threshold of 10% FDR. The colour intensity represents the degree of differential expression. a) DEGs at FDR <5% b) DEGs at FDR <10%. All results from this analysis can be found in Appendix Table 1.

Differential expression is widespread across cell-types in both the PILO and FCD comparisons, with the highest number of DEGs seen in the FCD comparison. Notably there are very few DEGs seen in the SSSE comparison. Excitatory Neurons had the most prominent changes in gene expression in the FCD condition, 1,288 genes were upregulated, while 617 genes were downregulated, followed by Astrocytes with 563 DEGs in total. Conversely, in the PILO model Astrocytes had the highest number of DEGs with an upregulation of 853 genes and downregulation of 152. This is followed by Inhibitory neurones. Inhibitory Neurons displayed consistent differential expression across both FCD and PILO conditions, with 198 and 211 genes downregulated, and 337 and 333 upregulated respectively. In both models, there is more upregulation seen generally across cell-types.

Of the neuronal cell-types, Excitatory neurones are more highly differentially expressed in the FCD model where as Inhibitory neurones are more highly differentially expressed in the PILO model suggesting a disruption in different parts of neuron hyperexcitability. Possibly due to the limited power of the study, only one gene was seen to have the same expression pattern across all three models, *Palmd*, which was downregulated in excitatory neurones in all comparisons. This gene encodes a protein which is implicated in cytoskeleton regulation (180). Interestingly, Palmd has recently been shown to interact with Pdlim5 in excitatory neurons, with knockdown of either *PalmD* or *Pdlim5* resulting in reduced neuronal dendritic branching (181). To date, *Palmd* has not been implicated in epilepsy (either as cause or effect), whereas aberrant dendritic spine formations are a known cytoskeletal feature of the epileptic brain (182). It remains to be determined if aberrant dendritic morphology is a cause of neuronal hyperexcitability, or a homeostatic response (which could be compensatory or maladaptive to the epileptic state) but our finding that this is a universal feature of diverse epileptic aetiologies warrants further investigation.

When considering genes which are convergently DE across two of the three epilepsy models, a higher numbers of converging genes between the FCD and PILO comparisons is observed: 110 which are upregulated and 97 downregulated in excitatory neurones, 42 upregulated and 56 down regulated in inhibitory neurones and 41 upregulated and 22 downregulated in astrocytes. This overlap suggests converging molecular changes (which may be cause or effect) between these two types of epilepsy, despite originating in different brain regions.

3.3.4 Functional enrichment of cell-type specific DEGs

Functional enrichment of Gene Ontology (GO) terms provides valuable insights into the biological significance of the DEGs within each cell-type, and can shed light into the biological mechanisms underlying epilepsy (or occurring as a consequence of the disease). To annotate the DE gene sets, over-representation analysis (ORA) was used, to test the enrichment of GO terms related to biological processes, molecular functions, and cellular components compared to a chance expectation. Pathways are then grouped based on their overarching parental terms using sematic similarity analysis. Figure 3.6 shows the functional enrichment of genes that are differentially expressed (FDR<10%) across cell-types in models of epilepsy.

In the PILO comparison, significantly enriched pathways were identified in the genes DE in most cell-types, with the highest number of pathways in Astrocytes in the upregulated genes. These genes highlighted terms linked to the mitochondrial matrix, mRNA processing and the early endosome. These all suggest an organisational role in upregulated astrocytes. The glial cell-types (Astro, OPC and Microglia) showed a significant upregulation of the receptor complex, where as inhibitory neurones showed a downregulation.

Some terms such as "PDZ domain binding," "receptor complex," and "early endosome" appear frequently across different cell-types, suggesting their importance in the context of epilepsy. PZD domain binding which is downregulated in excitatory neurones and upregulated in inhibitory neurones, astrocytes and OPCs.

a

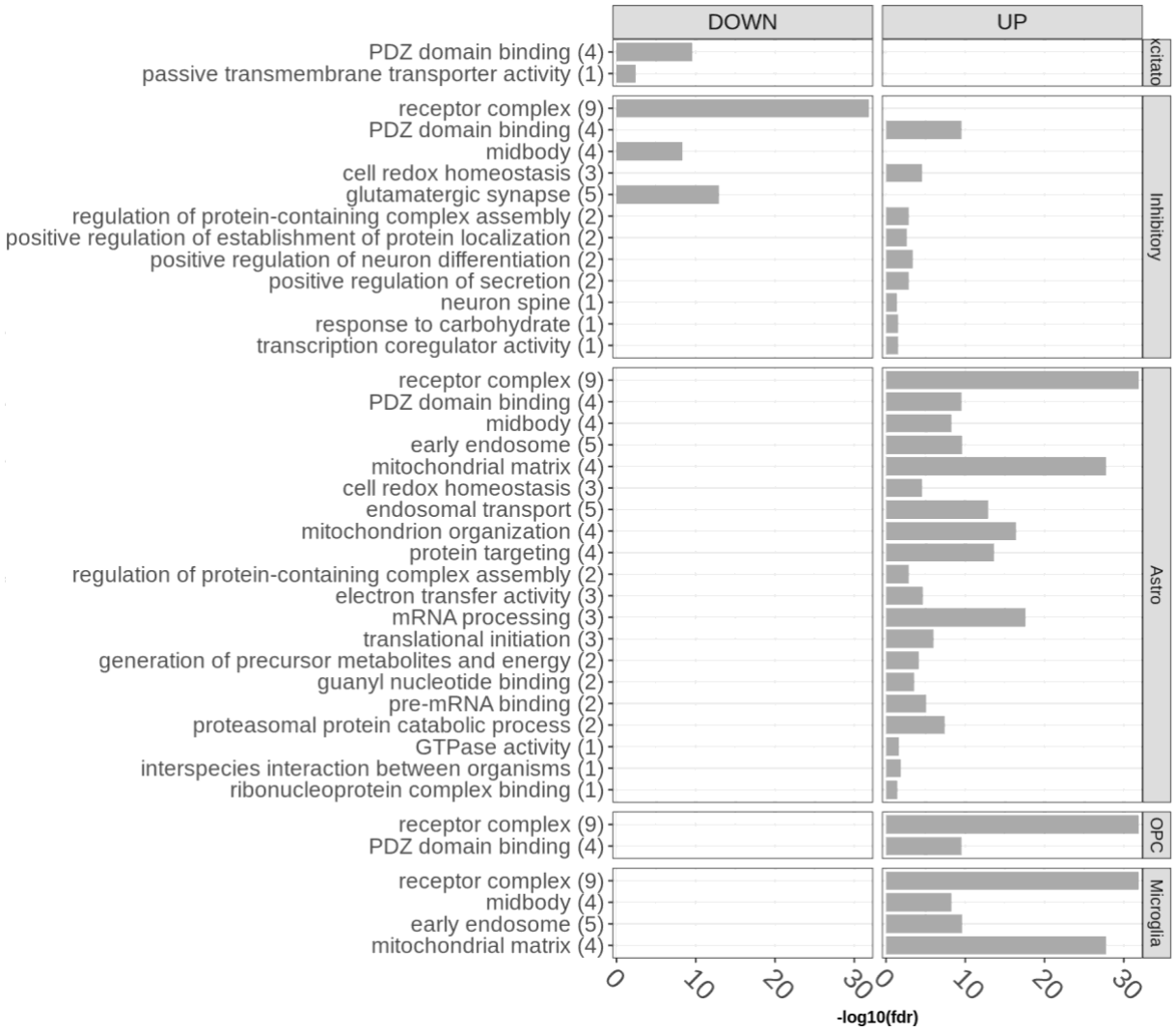
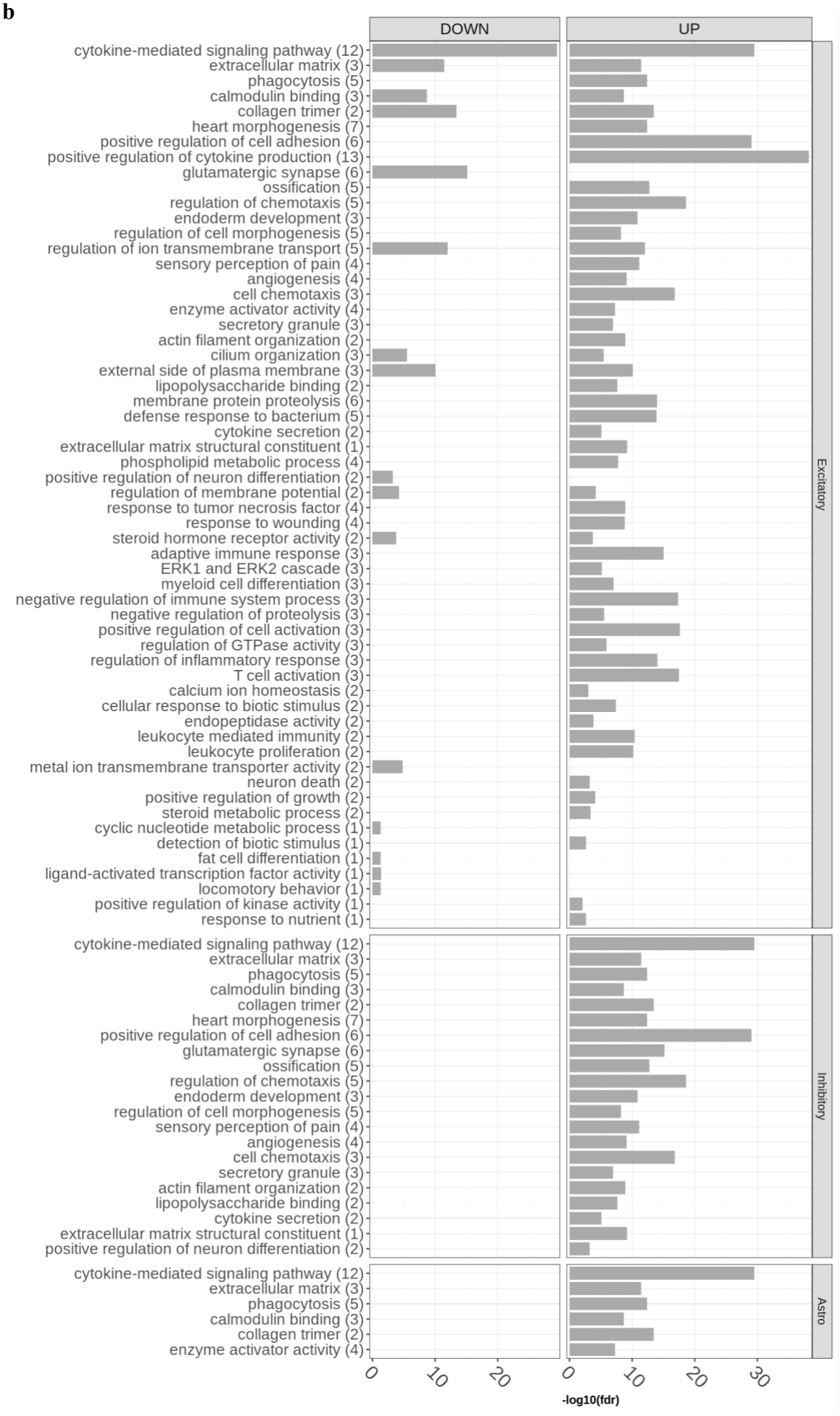


Figure 3.6 – Functional annotation of Gene Ontology (GO) terms across a) PILO DEGSs per cell-type, and b) FCD DEGs per cell-type. This graph shows the enrichment of reduced GO terms associated with cell-type specific up-(right) and down-(left) regulated DEGs in the pairwise comparisons in the different models. Sematic similarity was used to group original GO "child" terms, the numbers of which are shown in brackets on the y-axis. The x-axis represents the -log10(FDR) of the most significant "child" term associated with that parental term for each cell-type. All results from this analysis can be found in Appendix Table 2.

b


DOWN
UP
Excitatory
Inhibitory
Astro
cytokine-mediated signaling pathway (12)
extracellular matrix (3)
phagocytosis (5)
calmodulin binding (3)
collagen trimer (2)
heart morphogenesis (7)
positive regulation of cell adhesion (6)
positive regulation of cytokine production (13)
glutamatergic synapse (6)
ossification (5)
regulation of chemotaxis (5)
endoderm development (3)
regulation of cell morphogenesis (5)
regulation of ion transmembrane transport (5)
sensory perception of pain (4)
angiogenesis (4)
cell chemotaxis (3)
enzyme activator activity (4)
secretory granule (3)
actin filament organization (2)
cilium organization (3)
external side of plasma membrane (3)
lipopolysaccharide binding (2)
membrane protein proteolysis (6)
defense response to bacterium (5)
cytokine secretion (2)
extracellular matrix structural constituent (1)
phospholipid metabolic process (4)
positive regulation of neuron differentiation (2)
regulation of membrane potential (2)
response to tumor necrosis factor (4)
response to wounding (4)
steroid hormone receptor activity (2)
adaptive immune response (3)
ERK1 and ERK2 cascade (3)
myeloid cell differentiation (3)
negative regulation of immune system process (3)
negative regulation of proteolysis (3)
positive regulation of cell activation (3)
regulation of GTPase activity (3)
regulation of inflammatory response (3)
T cell activation (3)
calcium ion homeostasis (2)
cellular response to biotic stimulus (2)
endopeptidase activity (2)
leukocyte mediated immunity (2)
leukocyte proliferation (2)
metal ion transmembrane transporter activity (2)
neuron death (2)
positive regulation of growth (2)
steroid metabolic process (2)
cyclic nucleotide metabolic process (1)
detection of biotic stimulus (1)
fat cell differentiation (1)
ligand-activated transcription factor activity (1)
locomotory behavior (1)
positive regulation of kinase activity (1)
response to nutrient (1)
0
10
20
30
-log10(fdr)

In the FCD comparison, ORA demonstrated significant enrichment of terms in Excitatory, Inhibitory and Astrocytes. Particularly in excitatory neurones, there are a high number of terms enriched, consistent with the high number of DE genes. Notably, inflammatory terms are highly dysregulated in excitatory neurones with 11 upregulated terms related to the cytokine mediated signalling pathways, 13 terms related to positive regulation of cytokine production are upregulated, and cytokine secretion and phagocytosis are also upregulated. Activation of immune response and adaptive immune, immune response to antigenic stimuli, and leukocyte proliferations and leukocyte mediated immune response, t-cell activation response also upregulated. Cytokine-mediated signalling was not exclusive to excitatory neurons but was also observed in inhibitory neurons and astrocytes.
Neuronal terms are also enriched in the neurones. The glutaminergic synapse is upregulated in inhibitory neurones but downregulated in excitatory neurons in the FCD model. Positive regulation of neuron differentiation is also upregulated in inhibitory neurons and down regulated in excitatory neurons, and neuron death is upregulated in excitatory neurones. These results indicate an increased function of inhibitory neurons in response to epilepsy and a response from excitatory neurones to upregulate immune-related genes. The involvement of inhibitory neurones can be both anti-epileptic and pro-epileptic which makes physiological interpretation difficult (183).

Collectively, these results emphasise most crucially the involvement of excitatory, inhibitory, and astrocyte cells in both the PILO and FCD comparison. Astrocytes and Excitatory neurones do not share any overlapping terms in the two conditions, however inhibitory neurons in both conditions have upregulated positive regulation of neuron differentiation. Notably the glutaminergic synapse is downregulated in inhibitory neurones in PILO but upregulate in inhibitory neurones in FCD.

In contrast, the SSSE comparison lacked significant functional annotation terms. However, the upregulated genes in the excitatory neurones were found to be associated with immune response, similarly to the FCD excitatory neurons, and also cell-cell adhesion, featuring genes such as *Fndc3a, Igsf9, Itgb7*, and cytokine receptors like *Il15ra, Lepr*, and *Tgfbr2.*

3.3.5 Gene list enrichment results

To assess if the cell-type specific DE genes observed in epileptic mice are potentially disease related, we first tested for the enrichment of genes known to be associated with epilepsy. A hypergeometric test was applied to test enrichment against two distinct gene-sets; a list of "high-evidence" monogenic epilepsy genes from the Genomics England Epilepsy panel (167), and a list of genome-wide significant hits from epilepsy susceptibly GWAS (142). Corresponding FDR values are shown in Figure 3.7. The genes expressed in each cell-type were used as the background set for the enrichment.

In the PILO comparison, upregulated Astrocyte genes showed enrichment of monogenic epilepsy genes, however in the FCD comparison, it is the downregulated Astrocyte genes that are enriched, despite more DEGs being upregulated in total. Both PILO and FCD comparisons show enrichment of monogenic epilepsy genes in downregulated inhibitory neurons. Downregulated genes in excitatory neurones are also highly significant in the FCD comparison however not in the PILO comparison. There is also significant enrichment of monogenic epilepsy genes in the other glial cell-types (microglia and oligos) in PILO in the upregulated genes. For the complex epilepsy genes, found through GWAS studies, significant enrichment was only seen in downregulated genes in both the FCD and PILO comparisons. Common epilepsy genes are highly enriched in both neuronal cell-types in FCD and PILO. In the glial cell-types, enrichment is seen only in Astrocytes in PILO and OPCs and Pericytes in

FCD. There is no enrichment of common epilepsy genes in the upregulated DE suggesting common genes are mostly downregulated. This analysis confirms the presence of both common and monogenic epilepsy associated genes across multiple cell-types in the PILO and FCD comparisons. This analysis again highlights the involvement of Excitatory and Inhibitory neurones and Astrocytes, suggesting these are the driving epilepsy cell-types.

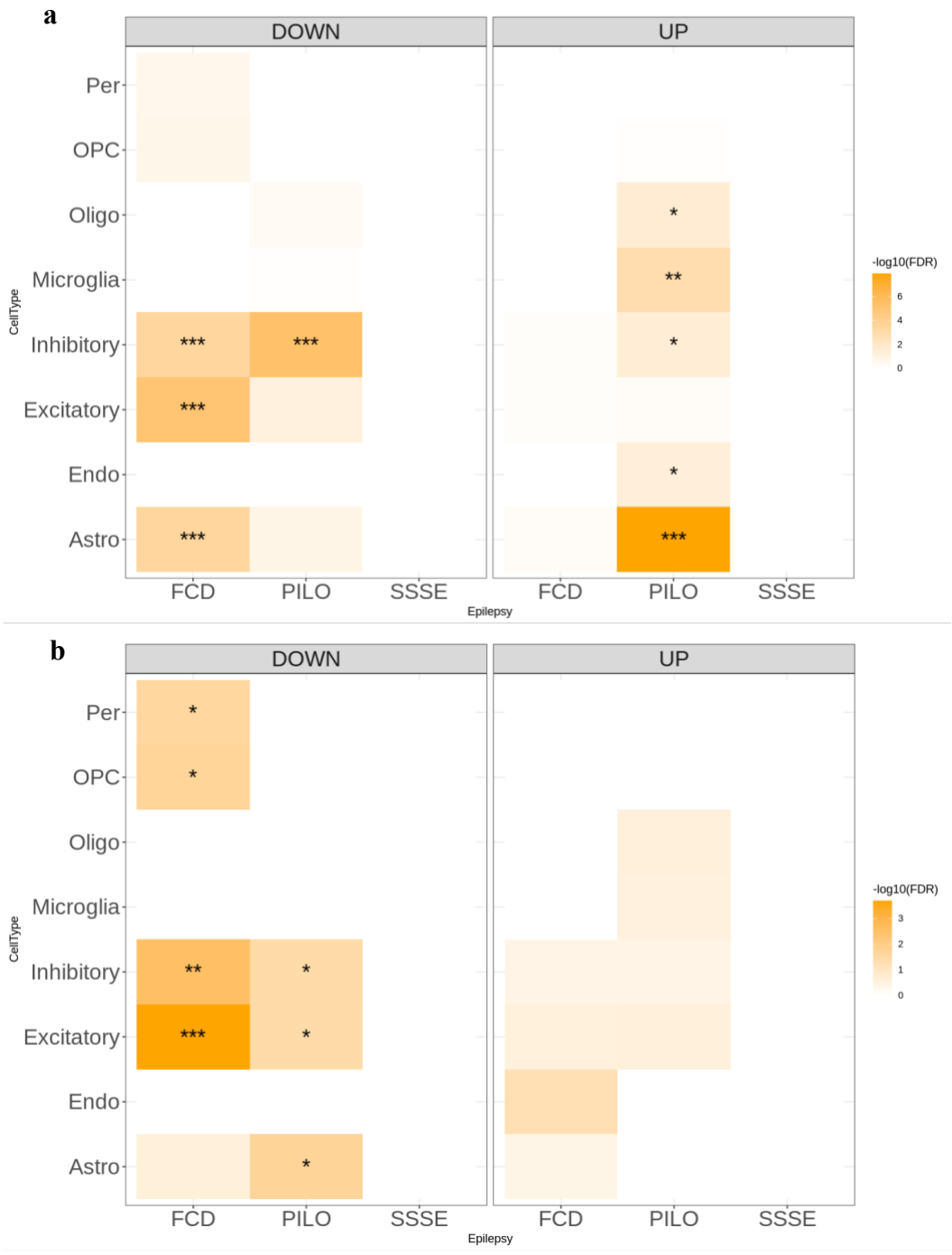


Figure 3.7 – Enrichment analysis epilepsy-related gene sets within the DEGs derived from the single-nucleus data. Heatmaps of enrichment FDR values from a Fishers exact test using A) a set of monogenic genes from Genomics England, which are categorised as "green" meaning they have a strong level of evidence linking them to epilepsy, and B) genes identified in previous epilepsy GWAS studies(184)*. These were tested for enrichment within the cell-type and model specific up-(right) and down-(left) regulated DEGs. Heatmaps are coloured by degree of significance and nominal significance is indicated using: ***FDR < 0.001; **FDR ≤ 0.01; *FDR ≤ 0.05.*

3.3.6 Integration of differentially expressed genes with genetic data

While cell-type specific DEGs can be related to molecular pathways and known epilepsy genes, it remains unclear whether the transcriptional changes seen are causing the disease or are just occurring as a consequence of the disease. Consequential gene expression can occur via a number of mechanisms, including inflammatory responses to recurrent seizures, or neuronal homeostatic processes potentially acting to mitigate the epileptic disease state (185). In this way, gene expression changes may serve to both mitigate or potentiate (the epileptic disease state, and these effects could be different in different cell-types. To try to get some understanding of the relationship of cell-type specific DEG to epilepsy, and ideally some insights into directionality, cell-type specific DEGs can be evaluated for underlying association with epilepsy heritability using stratified LD score regression (sLDSC) for GWAS data or enrichment tests for WES.

Using summary statistics from epilepsy GWAS, sLDSC can asses the enrichment of common genetic variants that are genetically associated with epilepsy. Provided there are no

unobserved confounding factors, then since genotype is independent of disease status, it is reasonable to assume that genetic variants associated with epilepsy are more likely to contribute riskto the disease.

LDSC was used to determine whether the proportion of SNP-heritability associated with the DEGs for each cell-type was greater than chance expectation, by partitioning heritability using the largest and most recent epilepsy susceptibility GWAS from the ILAE (33). This GWAS was analysed separately for both generalised and focal epilepsies as well as all forms of epilepsy. Enrichment of genetic association was tested for up- and down-regulated cell-type specific DEGs (FDR < 0.1) from each comparison (Figure 3.8).

Most significant enrichment is observed for generalised epilepsy susceptibility across the majority of cell-types. Despite PILO and FCD being models of focal epilepsies, enrichment of focal epilepsy heritability is only seen in one cell-type – upregulated excitatory neurones. Generalised epilepsy heritability is observed in both up- and down- regulated PILO and FCD neurones, as well as up- and down- regulated astrocytes in PILO and only down regulated genes in astrocytes in FCD. In PILO, all cell-types but Endothelial cells are enriched for generalised or all epilepsy in the upregulated genes. This analysis further indicated that excitatory, inhibitory neurones and astrocytes are heavily implicated, consistent with the cell-types previously identified in the gene list enrichment results.

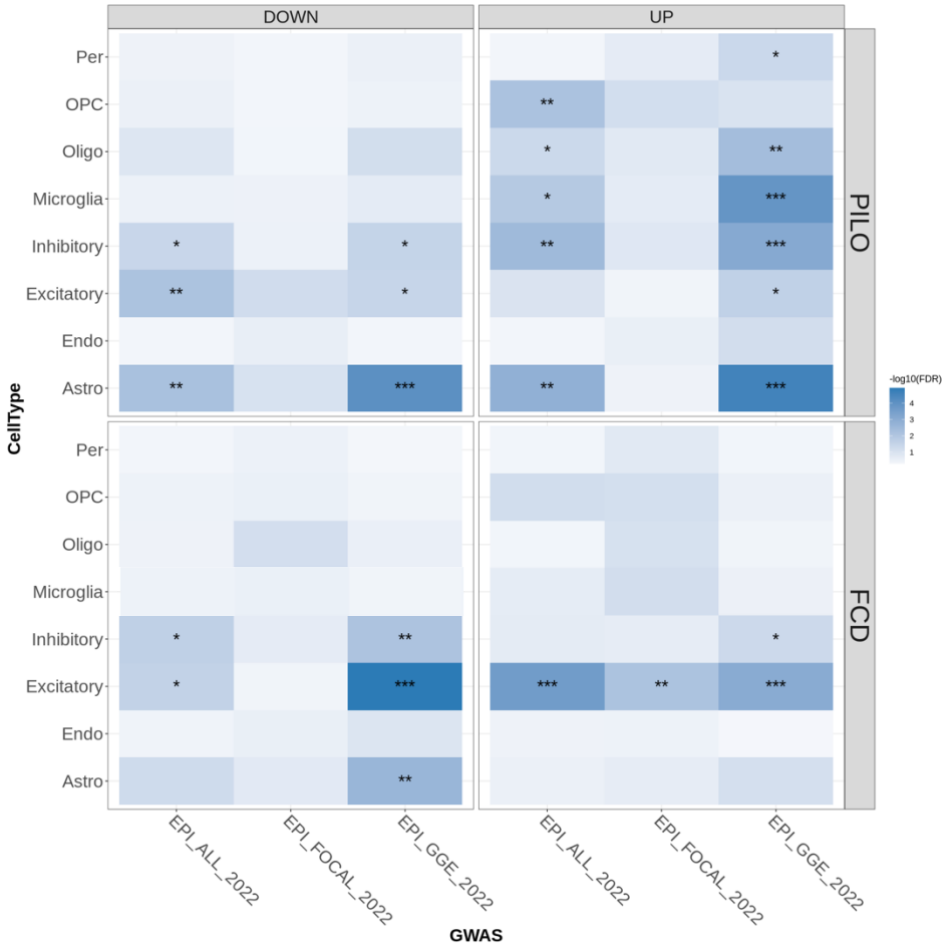


Figure 3.8 – Association of heritability analysis of cell-type specific DEGs using LD Score Regression. Cell-type and model specific DEGs were used to identify genetic association with epilepsy risk using ILAE 2022 GWAS studies. The analysis is run for the Focal Epilepsy, Generalised Epilepsy and All Epilepsies denoted EPI_FOCAL_2022, EPI_GGE_2022, and EPI_ALL_2022. Heatmaps are coloured by degree of significance and nominal significance is indicated using: ***FDR < 0.001; **FDR ≤ 0.01; *FDR ≤ 0.05.

To further explore the association of heritability within the differently expressed genes, we also looked at the enrichment of rare variants heritability as well as common variants.

The EPI25 dataset is a Whole Exome Sequencing (WES) dataset from 30,000 people with epilepsy coordinated by the ILAE Genomics consortium. The study considers epilepsy in 4

broad phenotypic categories severe developmental and epileptic encephalopathies (DEEs), genetic generalized epilepsy (GGE), non-acquired focal epilepsy (NAFE) and All epilepsy phenotypes. Summary level data was generated using collapsing analysis of deleterious rare variants (186). We considered gene-level results for the following categories of coding variant: Synonymous (Synon), Protein Truncating Variants (PTV), missense (EMiss), damaging missense (DamMiss), intolerant missense (EInt) and damaging and intolerant missense (EIntDam). This collapsing (or burden) analysis, produces a gene-level p-value of association to epilepsy or each gene in the genome.

Using these ranked gene level p-values we tested the enrichment of the DEGs for each cell-type in the PILO and FCD comparisons, using a tool called GSA-SNP2 (187) (Genome-wide SNP-based pathway analysis). GSA-SNP2 uses a modified Z-statistic test to compare the observed distribution of gene-level association p-values to the expected under the null hypothesis of no association. The null distribution is calculated by random set of genes of the same size. The Z-statistic is implemented to model the size of the gene set, avoiding false positive results arising from gene sets with a biased higher proportion of larger genes. The tool also provides a type I error control by adjusting for the SNP counts in each gene and adjusts for LD by removing adjacently located genes within the same gene-set that have a high positive correlation (>0.5) in the 1000 genomes dataset (on average this occurs in <1% of gene sets). We considered summary results for the clinical categories “All Epilepsy”, DEE, GGE and NAFE. The results are summarised in Figure 3.9.

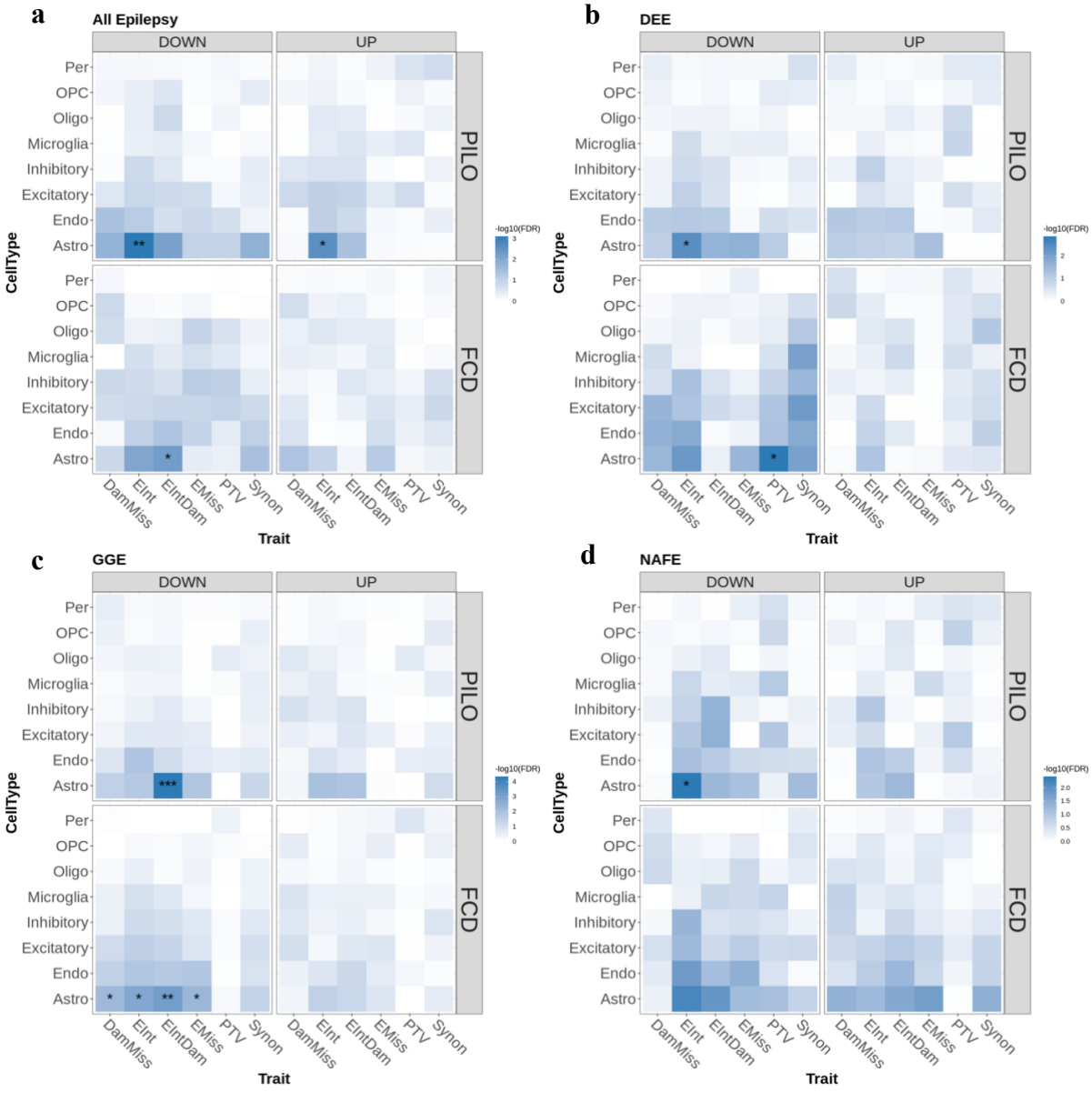


*Figure 3.9 – Analysis of enrichment of heritability of ultra-rare variants using GSA-SNP2. These heatmaps display results of heritability enrichment of ultra-rare variants associated with various epilepsy subtypes including (a) All Epilepsy, (b) developmental and epileptic encephalopathies (DEEs), (c) severe genetic generalized epilepsy (GGE), and (d) non-acquired focal epilepsy (NAFE). The following coding variants were included in the analysis: Synonymous (Synon), Protein Truncating Variants (PTV), missense (EMiss), damaging missense (DamMiss), intolerant missense (EInt) and damaging and intolerant missense (EIntDam).*Heatmaps are coloured by degree of significance and nominal significance is indicated using: ***FDR < 0.001; **FDR ≤ 0.01; *FDR ≤ 0.05.

The advantages of this analysis is that we can look at high penetrance variants which are deleterious, and are therefore direct impact on disease, allowing a tentative inference as to the causal relationship between the cell-type specific DEG and epilepsy. The results reveal a complex picture of cell-type specific differential expression in epilepsy. Of and by themselves, the results do not guide us on whether cell-type specific DEG are pro-epileptogenic or arise through compensatory homeostatic anti-epileptogenic mechanisms. For DEE mutations are intolerant in PILO and Protein Truncating in FCD. For GGE, mutations are primarily intolerant and damaging, whereas all missense mutations are implicated in FCD. However, at face value, the downregulation of genes in FCD and PILO which cause epilepsy when deleteriously mutated, suggests that their downregulation in astrocytes may be contributory to the hyperexcitable epileptic state.

3.3.7 Perturbation analysis with CRISPRbrain

The differential expression analysis revealed transcriptomic signatures that are cell-type specific and enriched in genes known to contribute risk to epilepsy. By examining the pathways associated with these genes, we gained insights into the mechanisms implicated in epilepsy. Instead of focusing solely on individual genes whose expression is altered in epilepsy, we can investigate these gene sets collectively to identify upstream regulators responsible for transcriptional shifts across multiple epilepsy-related genes. By targeting regulators of epilepsy-causing genes, we can potentially identify drug targets that address the broader pathways governing the disease rather than just single genes.

Building on these previous findings identifying key genes altered in epilepsy, we wanted to delve deeper into understanding the transcriptomic responses of these genes. CRISPRbrain is a dataset of functional genomics screens in differentiated human cell-types (172). Recent advances of CRISPR technology has allowed for large scale experiments, perturbing gene function using the CRISPR/cas9 system. This allows modification of the genome sequence to adjust the expression levels of specific genes to inhibit/activate specific genes. This can allow us to look at gene function and understand their roles in particular pathways and processes. This can be used to validate potential drug targets by understanding how their affects transcriptomic response.

In this analysis, the DEGS identified in the single-cell experiment were intersected with the CRISPRbrain database, a database of CRISPRi/a activation for 300 gene perturbations. Using a framework conceptualised by Srivatava et al.(59), the overlap of DEGs within the perturbation dataset was assessed to discern the shared functional impact. An enrichment test, using up- and down- regulated DEGs in epilepsy separately, was conducted on the genes

dysregulated through a gene perturbation. When the DEGs align with those affected by perturbation, it implies a potential regulatory function of the perturbed gene in the initiation of epilepsy onset. These findings can identify specific transcriptional changes that when disrupted, lead to transcriptional shifts mirroring those that occur in epilepsy. Analysing these changes at a cell-type specific level provides an important insight into how the cell functions in disease. The role of these genes, whether activating or inhibiting the disease depends on the directionality of the overlapping gene sets and the perturbation effect. Modulating the expression of these potentially disease-causing genes can help to reverse disease phenotype (or potentially augment if the directionality is reversed or misconstrued). The direction of the DE genes in conjunction with the CRISPR brain genes can then help inform the therapeutic strategy for the disease, determining whether a gene should be agonised or blocked in order to counteract the transcriptional effects of the disease. This framework is illustrated in Table 3.2 with directionality assumed on the basis of simple signature reversion.

DE direction	CRISPRbrain direction	Perturbation	Disease effect	Therapeutic strategy
↓	↓	CRISPRa	ACTIVATOR	Block / Inverse agonist
↑	↑	CRISPRa	ACTIVATOR	Block / Inverse agonist
↓	↑	CRISPRi	ACTIVATOR	Block / Inverse agonist
↑	↓	CRISPRi	ACTIVATOR	Block / Inverse agonist
↓	↑	CRISPRa	REPRESSOR	Agonist
↑	↓	CRISPRa	REPRESSOR	Agonist
↓	↓	CRISPRi	REPRESSOR	Agonist
↑	↑	CRISPRi	REPRESSOR	Agonist

Table 3.2 – This table shows how the overlap and directionality between differentially expressed genes in epilepsy, and gene perturbations suggests how the regulating gene is modulating the disease. This shows how the potential therapeutic strategy can be based on the outcome of the specific perturbation. Arrow pointing upwards (↑) represents upregulation, while arrow pointing downwards (↓) signifies downregulation.

The CRISPRbrain perturbation screens were performed in 2 cell-types; glutaminergic neurones and microglia as these are the only cell-types with transcriptional readouts following gene perturbation. In this analysis, the enrichment tests were performed on the DEGs from matching cell-types, with both inhibitory and excitatory DEGs being tested against perturbations in glutaminergic neurons.

In total 151 of the perturbed genes showed significant enrichment at 5% FDR in one or more model or cell-type. The most significant enrichment was observed in the CRISPRi perturbation of the gene *MAT2A* which demonstrated an overlap with the downregulated DEGs in the FCD excitatory neurones and as well as inhibitory neurones in both PILO and FCD (Figure 3.10). This suggests that an inactivation of *MAT2A* results in the downregulation of genes also down-regulated in epilepsy. Based on the reasoning of the previously described framework, activating *MAT2A* should potentially realign the downregulated genes in epilepsy closer to a healthy state, suggesting *MAT2A* as a disease repressor. A therapeutic strategy might then involve using an agonist. Methionine Adenosyltransferase 2A (*MAT2A*) is a protein coding gene that catalyses the production of S-adenosylmethionine from methionine and ATP, and lowered levels of this enzyme have been associated with gliomas (188). While this gene has not previously been directly linked to epilepsy, it has been shown that *MAT2A* is involved in the activation of the mTOR pathway through the regulation of DNA synthesis, which is known to cause FCD (163,189).

Functional annotation of the genes downregulated when *MAT2A* was knocked out showed 153 downregulated Biological Processes, including many related to synapse formation, neurotransmitter transport and signal release. Conversely, upregulated terms were related to mRNA processing, splicing and metabolic processes (Full FA in Appendix). S-adenosylmethionine has also been shown to have antiepileptic effects in epileptic rats therefore inhibition of this gene could be inducing epilepsy (190).

For each of the gene perturbations, two additional metrics were calculated. The 'Weight' of the gene regulator, which is determined by the proportion of genes under- or over- expressed in the disease, that are enriched in the set of genes dysregulated by the CRISPR perturbation. The 'Activity' of the gene was also calculated which is the proportion of the genes perturbed by the CRISPRi/a which are also differentially expressed in epilepsy. The Activity and the Weight are then combined and ranked. These were plotted against the log10 FDR of enrichment (Figure 3.10) which allows prioritisation of genes based on the strength of the predicted effect. *MATA2A* also has the highest ImpactRank showing that a large proportion of the significant genes are present in the DE and the perturbation.

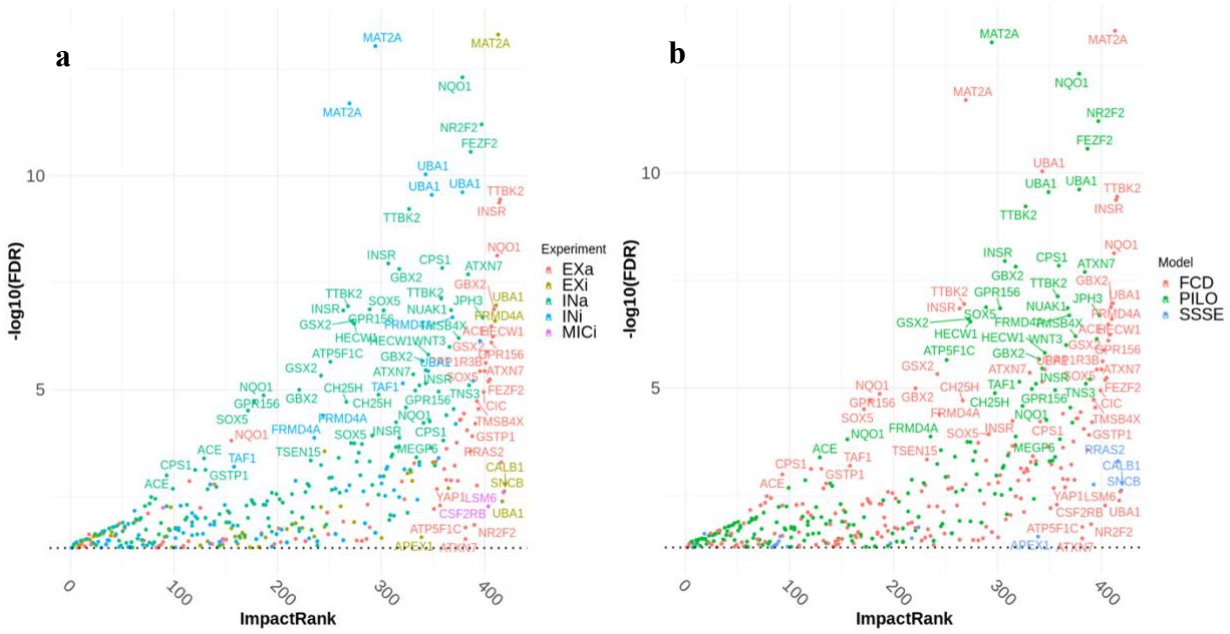


Figure 3.10 – Scatter plot to show the relationship between Impact Rank and -log10(FDR) for all significantly enriched gene perturbations, across different experimental models and conditions. a) The scatter plot is coloured by cell-type and CRISPR perturbation. Excitatory and Inhibitory neuron DEGs were tested against the glutaminergic neuron CRISPRi and CRISPRa. Microglia DEGs were tested against the microglia CRISPRi perturbations. b) The scatter plot is coloured by the mouse model the DEGs were identified in. Full results of this analysis can be found in Appendix Table 3.

Another highly significant perturbation which also had a high Impact Rank is *NQO1*. In this case, the gene was activated using CRISPRa, the downregulated genes upon *NQO1* activation were significantly enriched for those both up- and down- regulated in epilepsy. This suggests that *NQO1* could be acting as a disease activator or repressor. To gain more insight, a literature search was conducted. *NQO1* knockout mice have been shown to display spontaneous seizures due to alteration in intracellular redox status and altered glucose

metabolism (191). Functional annotation of the *NQO1* perturbation showed that cell redox homeostasis was upregulated biological process as well as other metabolic process. A large number of GO terms were downregulated, the majority of which are related to synaptic function. This suggest *NQO1* may be playing a role as an activator of epilepsy throughout metabolic abnormalities leading to seizure initiation.

While the majority of significant regulators are significant in excitatory or inhibitory neurones, there are five which were found to be significant in microglia; *CDK12, CSF2RB, EPOR, LSM6* and *MAPK14*.

This analysis has revealed potential regulators of epilepsy that might serve as potential therapeutic targets. Nonetheless, the challenges of prioritising and validating these targets remain. While these preliminary findings are promising, identifying and testing feasible drug candidates is a long and expensive process, highlighting the need for improved *in vitro* scalable functional readouts for novel candidate disease modifying therapies in epilepsy

3.3.8 Intersections with interaction databases

Repurposing existing drugs is a strategic approach to speed up the drug development process, given the long and costly process of identifying and testing new drugs. To identify pre-existing drug candidates, significant genes were cross-referenced with a database cataloguing known gene-ligand interactions, the IU/Pharm database. Using this database, 24 of the significant genes were found to have already established ligand interactions. With the aim of drug repurposing in mind, we focused on genes linked to already approved drugs. Out of the 24 identified genes, 9 genes had interactions with drugs that have already received approval for use in humans (Table 3.3).

Table 3.3 – This table shows the significant genes found to have pre-existing ligand interactions, and the number of the ligands from the IU/PHAR database. It further provides details on the specific ligands which are approved for use in humans. Additional information is given regarding whether these gene targets are known epilepsy genes and if they are listed in the PDE3 database.

Gene	Number of Known Ligands	Number of Approved Ligands	Approved Ligands	Known Epilepsy Gene	PDE3 Database
ACE	25	21	benazepril, benazeprilat, captopril, cilazapril, enalapril, enalaprilat, fosinopril, imidapril, imidaprilat, lisinopril, moexipril, perindopril, perindoprilat, quinaprilat, ramipril, ramiprilat, spirapril, spiraprilat, trandolapril, trandolaprilat, zofenopril	No	captopril, enalapril
ACHE	14	10	donepezil, echothiophate, edrophonium, galantamine, neostigmine, physostigmine, pralidoxime, pyridostigmine, rivastigmine, tacrine	No	No
ATM	9	0		No	No
BCL2	10	1	venetoclax	No	No
BCL2L1	6	1	venetoclax	No	No
CASP3	8	0		No	No
CCR5	37	1	maraviroc	No	No
CDK12	2	0		No	No
CREBBP	5	0		Yes	No
CTSD	5	0		Yes	No
EPOR	3	2	erythropoietin, peginesatide	No	No
GABRG3	2	0		No	No
HDAC9	13	4	belinostat, panobinostat, romidepsin, vorinostat	No	No
INSR	9	3	insulin, brigatinib, ceritinib		No
LSS	11	0		Yes	No
MAP3K12	6	0		No	No
MAPK14	31	0		No	No
NTRK1	30	2	gilteritinib, larotrectinib	No	No
NTRK2	14	0		Yes	No
NUAK1	7	0		No	No
SLC2A3	1	0		No	No
TAF1	1	0		No	No
TOP1	1	0		No	No
XIAP	9	0		No	No

To further validate these targets and their interactions, these gene targets were intersected with the PDE3 database. This is a curated database of all drugs that have been used and proven effective in the treatment of epilepsy in either humans or animal models of epilepsy(175). Data from the PDE3 database highlighted two drugs, Captopril and Enalapril, both of which target the ACE (Angiotensin-converting enzyme inhibitor) gene (Table 3.4). The framework identified ACE as significant in both Excitatory and Inhibitory neurons for the FCD and PILO comparisons. Given that the enrichment was observed in upregulated genes and the perturbation was an inhibition, this suggests that ACE plays an activating role in epilepsy and thus should be inhibited. These drugs have been prescribed for epilepsy management, demonstrating efficacy, especially when paired with another AED (192,193). Further literature research revealed that both Captopril and Enalapril, can supress epilepsy-induced activation in KA mice and counteract the abnormal interaction between astrocytes and microglia. This serves as a validation for this approach (194,195).

Name	Indications	First licensed (year)	Prescription items dispensed in England in 2016 (number)	Number of studies	Number of unique models	Models	AEDs enhanced
Captopril	Hypertension; chronic heart failure; diabetic nephropathy	1981	102,804	3	3	PTZ, AUD, Other	Carbamazepine, Lamotrigine, Topiramate, Valproate, Other
Enalapril	Hypertension; heart failure	1985	1,871,692	1	1	Other	Carbamazepine, Lamotrigine, Topiramate, Valproate, Other

Table 3.4 – This table shows two ACE inhibitors that were found to be effective in the treatment of epilepsy in the PDE3 database.

While the other approved drugs were not in the PDE3 database, and have not been previously been prescribed specially for the treatment of epilepsy a literature search showed that many of these gene candidates with approved ligand interactions, have prior associations with epilepsy. The *CCR5* (C-C chemokine receptor 5 antagonist) exhibits significant upregulation

in excitatory neurons in the PILO model, particularly in genes upregulated when the gene is knocked out. This implies that *CCR5* acts as an epilepsy activator, and its inhibition might alleviate symptoms. Maraviroc, an FDA-approved drug targeting *CCR5*, has demonstrated efficacy in preventing microglial activation and neuronal degeneration in mice with seizures. By inhibiting the *CCL5/CCR5* pathway with maraviroc, seizure symptoms were observed to decrease (196). Additionally, *HDAC9* (Inhibitor of histone deacetylases), a target of Vorinostat has also been shown to blocks seizures in zebrafish and rodents and could serve as promising candidates for drug development, potentially bypassing the extensive safety approval process as they are already approved (197).

3.4 Discussion

This work leverages single-nucleus sequencing data derived from three mouse models of epilepsy with the aim of identifying cell-types and underlying mechanisms associated with the disease. Furthermore, it seeks to utilize the cell-type specific differential expression information to identify drug targets which work by modifying the disease. By identifying specific genes and pathways that are dysregulated in disease, this information can be used as a potential framework for disease modifying drug target prioritisation. Given the intensive nature of the traditional drug approval process, this work makes use of publicly available databases to screen for existing drugs which can be reused for new therapeutic goals based on the framework outlined in this Chapter.

In summary, 85 samples from three models of epilepsy were combined on a joint UMAP and cell-type specific clusters were identified and annotated. Differential expression analysis was then performed for each pairwise comparison (PILO vs Sham, FCD vs Sham and SSSE vs Sham) to identify significantly dysregulated genes for each model and potential convergent DE across different epilepsy aetiologies. A conservative pseudobulk DE approach was used, as opposed to traditional single-cell DE methods which lead to a high rate of type 1 errors. The results of the differential expression analysis revealed significant differences in gene expression in all pairwise comparisons apart from SSSE which showed very few DE genes, even at a more liberal 10% FDR threshold. The lack of differential expression in SSSE raises the question of why this group did not exhibit significant changes in gene expression compared to the control group. One possible explanation is the highly variable gene expression within cells in each condition, which may obscure differences between conditions, especially considering the heterogeneity of cell-types. Technical noise could also contribute to the lack of differential expression observed. This is likely due to smaller sample numbers

for the SSSE model causing low statistical power to predict DE genes. Power calculations showed that a minimum of 23 samples are needed to provide 80% Power to detect fold-change >1.5 when using pseudo-bulk differential expression (see Chapter 2.3).

Functional annotation of the dysregulated genes using ORA analysis of GO terms highlighted the involvement of astrocytes in both models of epilepsy. Astrocytes have been shown to have many roles in the brain and in epilepsy including the homeostasis of potassium and glutamate as well as other synaptic functions. The upregulation of PDZ domain binding and receptor complex as well as other terms related to cell binding and communication, may highlight the close relationship between neurones and astrocytes and emphasise the upregulation of homeostasis at synapses likely as a response to seizures (73).

There are many immune GO terms enriched within neuronal terms, and there is significance of the term "cytokine-mediated signalling pathway" across all cell-types in FCD, highlighting the involvement of inflammation in these models. There is growing evidence into the role of inflammation in epilepsy and while this is usually associated with glial cells, there is a close interaction between neurons and the glial cells, as the brain has a complex cellular environment. Cytokines and chemokines are also involved in synaptic plasticity, neuronal differentiation, and migration (198). However, it can be unclear whether the inflammation seen is causing the seizures or is a result of the seizures. A key limitation of pathway analysis is to determine whether the observed associations are causal, consequential, or simply correlated with the phenotype, and enrichment analysis alone cannot distinguish the relationship between pathways. It is also worth noting that this type of analysis can be inflated by a few genes, and therefore any contamination of microglial genes within neuronal cell droplets could be driving some of these enrichments.

The question of whether changes seen in the differential expression analysis are indicative of the genes causing epilepsy or whether they are seen simply as a consequence of the disease due to compensatory mechanisms is key. Enrichment of known epilepsy causing genes within the DEGs showed that several cell-types were enriched for previously established disease-causing genes. This approach was used to account for genes with both rare and common variants. While this indicates association of epilepsy genes, these may be correlation and do not provide direct evidence for causation. The observed enrichment might be a result or a cause of a specific phenotype, or it might be coincidental.

To more directly implicate genes in epilepsy, association of heritability can provide evidence of a genetic basis to a trait. Significant association of heritability with a set of genes, implies those genes are genetically associated with the disease.

Firstly, LDSC was used in conjunction with the most recent epilepsy susceptibility GWAS to explore which cell-types were significantly associated with common epilepsy heritability. This showed significant association of several cell-types with common variants. To look at rare genetic variants, GSA-SNP2 was used alongside whole exome sequencing from EPI25K, to look at the association of rare variants within the cell-type specific genes. These integrative methodologies, which look at data from GWAS and WES, can offer more convincing inferences regarding the causative nature of the DE genes. In the context of the rare variant genetic analyses in particular, this usually leads to loss-of-function and deleterious mutations, which when enriched in downregulated genes, strongly indicates a causative function in the disease.

The vast phenotypic and genetic heterogeneity observed in epilepsy, particularly in subsets like GGE and NAFE, underscores the need to look at epilepsies as a whole, using integrative

methodologies to find convergent features. The primary convergent feature of the analyses in both models is cell-type specificity around astrocytes and neurones. There is evidence in both focal and generalised epilepsies, and in both rare and common types. Astrocytes and neurons (inhibitory in particular), have been empirically linked to both the genesis and progression of epilepsy (73,144)

Understanding the genetic basis of a disease creates opportunity for potential interventions through targeting a gene or a set of genes. To establish causation, it needs to be demonstrated that a change in the gene directly leads to a change in the trait, as association alone is not enough to prove this. Usually this can be done through experimental studies.

In this study, we achieved this through the use of CRISPRbrain, as a tool for identifying disease associated gene function. Perturbation data provides a provisional validation of these transcriptional shifts seen in the mouse models, and helps to elucidate whether a particular gene should be activated or inhibited to counteract disease-induced transcriptional shifts. When differentially expressed genes show enrichment within genes influenced by perturbation, it leads to the hypothesis that these DE genes play a causal role in epilepsy's progression.

This analysis identified several genes that could be targeted for therapeutic function as well a way to quantify the strength of these predictions. Several of these putative targets could be validated through intersection with databases of known drug targets and through previously conducted experimental studies. While this approach predicts numerous genes that can substantially influence the expression patterns in epilepsy, certain notable examples stand out. This provides a method to refine a selection of genes which can be taken forward for experimental validation. Finally by focusing on known drugs, this approach emphasizes the advantages of drug repurposing in the context of epilepsy research in speeding up discovery and reducing costs.

3.4.1 Limitations and Future Work

Incorporating preclinical mouse models into this analysis offers the ability to look at genes and pathways in a controlled experimental setting without the confounders of comorbidities and genetic variation, providing insights into molecular pathways and cellular interactions difficult to study directly in human samples. These models enable the manipulation of genetic, environmental, and pharmacological variables to elucidate relationships which can then be validated in human studies. Using mice also allows us to increase sample size and improve the power as mouse samples are more readily available and easier to obtain.

In this study, we saw very few differentially expressed genes in the SSSE model. The main comparison performed was between the SSSE mice which had been implanted with electrodes and treated with electrical stimulation versus the control mice which had been transplanted with electrodes but did not have electrical stimulation. This approach aimed to distinguish changes specifically induced by SSSE from those possibly influenced by surgical implantation. Additionally, samples from the contralateral side of the brain were included, highlighting the need for future studies to systematically compare gene expression changes between hemispheres. This comparative approach could elucidate whether observed effects are localized to one hemisphere or represent bilateral responses, thereby providing a more comprehensive understanding of the neurological impact of SSSE and its potential applications in clinical settings.

In the context of focal cortical dysplasia (FCD) and temporal lobe epilepsy (TLE) mouse models, a limitation arises when brain regions are analysed collectively, and neuronal subtypes are aggregated. FCD and TLE are characterized by heterogeneous pathological features that can vary spatially within the brain. Analysing brain regions together may obscure region-specific differences in gene expression patterns or cellular responses that are

critical for understanding disease mechanisms. Similarly, collapsing neuronal subtypes overlooks the diversity in cell populations that may have distinct roles in disease pathology or therapeutic responses. For instance, specific neuronal subtypes within affected regions like the hippocampus or neocortex may exhibit differential vulnerability to excitotoxicity or aberrant synaptic plasticity, which are central to epileptogenesis.

FCD which is characterized by abnormal cell proliferation and migration results in structural abnormalities in discrete brain regions (199). When analysing brain regions collectively, crucial variations between dysplastic and normal tissues may be obscured. Aggregating data from diverse brain regions could mask region-specific pathological changes and fail to capture the unique cellular and molecular characteristics associated with FCD. Similarly, temporal lobe epilepsy is characterized by seizures originating from the temporal lobe, with complex alterations in the hippocampus and surrounding structures. In TLE mouse models, neuronal subtypes within the temporal lobe can exhibit distinct pathophysiological changes such as aberrant excitability or altered neurotransmitter release(200). Aggregating data from different neuronal subtypes or entire brain regions might obscure these subtype-specific alterations and reduce the resolution of functional and structural insights

Future strategies could explore the association of immediate early genes associated with neuronal activity by exploring the enrichment of these genes within differentially expressed gene sets. Immediate early genes are genes which are rapidly and transiently expressed in response to neuronal activity and are associated with various synaptic processes. These genes have become widely used as molecular markers for neuronal populations which are active or undergoing plastic changes. These can relate which in response to neuronal stimuli and play pivotal roles in synaptic plasticity. Using an enrichment test with these markers, could help

better identify the sub-types of neurons that are active in the epileptic data and help associate a role to certain subclusters or indicate which neuronal populations could be targeted.

Chapter 4

Single-nucleus RNA-sequencing of Human Epilepsy

4.1 Introduction

Systems-level biology seeks to understand complex biological processes from a comprehensive data-driven perspective, focusing on how components of the system interact and converge (201). Within this context, a systems approach involves combining diverse types of biological data such genetic, transcriptomic, and epigenetic to gain a more comprehensive understanding of biological processes (202).

The approach taken in this thesis not only explores convergence across epilepsy models but also investigates the conserved mechanism underpinning epilepsy across different species. The use of mouse models in epilepsy has been instrumental in drug discovery but validating finding in humans is still essential for the successful translation of therapies. Human tissue offers several advantages over mouse as it allows for the study of complexity that is not found in mouse. It is also valuable in studying more disease and patient specific factors that are only relevant in certain forms of epilepsy and can facilitate personalised medicine approaches (42,203). Using data from both human and mouse studies not only increases the power of findings but can ensure more robust findings that are replicated across multiple systems. While animal models remain essential, when human tissue can be obtained, this complements these studies and plays a crucial role in advancing our understanding of disease and may help bridge the gap to clinical applications (202).

A convergent framework suggests that different types of epilepsy with different genetic mutations or epigenetic modifications, and in different species can identify fundamental pathways for disease that may be targeted for therapeutic development across multiple different types of epilepsy. By integrating different data types, this enhances the capability to identify core pathways implicated in disease, and provides a unified understanding of the underlying mechanisms of epilepsy and ictogenesis. Following the "signature reversion" principle, reversing these effects can be a potentially powerful therapeutic strategy, by counteracting the disease associated expression (154).

Previous work from the Johnson Lab (36) demonstrated the integration of diverse datasets including brain bulk RNA-sequencing data from human and mice, as well as GWAS and rare variant data, and showed that genetic susceptibility converged on a gene regulatory network 'M30'. This provided evidence for a connection between rare and common epilepsies which provided a potential strategy for the development of therapeutics.

Incorporating single-cell resolution and epigenetics into our analysis provides additional insights into mechanisms of gene regulation and these can influence disease susceptibility and response to therapy.

4.1.1 Epilepsy Surgery

Surgery is often performed in epilepsy patients with drug-resistant epilepsy, when AEDs are not effective in controlling seizures. In the case of focal seizures which originate from one part of the brain, surgery can offer a reduction in seizure frequency or a completely seizure free life, while minimising the reliance in AEDs. Certain types of epilepsy such as in hippocampal sclerosis and FCD have well-established surgical procedures that can be very successful, but these surgeries must be tailored based on the specific case with pre-surgery assessments (204).

4.1.1.1 Hippocampal Sclerosis (HS)

Hippocampal sclerosis (HS) is a common epilepsy pathology characterized by the selective loss and atrophy of neurones in the hippocampus particularly involving CA1 and CA4. It is also accompanied by gliosis, which is proliferation of glial cells, and other features include the dispersion of granule cell layer and mossy fibre sprouting. These changes disrupt the normal functioning of the hippocampus which is important in memory and learning, and they contribute seizure development (205). Hippocampal sclerosis is commonly present in patients with temporal lobe epilepsy (TLE), particularly mesial temporal lobe epilepsy, which is a subtype of epilepsy characterized by seizures that original from the mesial temporal lobe. Hippocampal sclerosis is often a marker for drug resistant epilepsy. For most patients surgical excision of the affected region is eventually required, and temporal lobectomy for hippocampal sclerosis is the most common epilepsy surgical procedure (206). The surgical procedure usually involves an anterior temporal lobectomy where a section of the anterior lateral temporal neocortex is removed, along with medial temporal lobe structures where the seizures are generated from. The prognosis after surgery is good, and studies have shown that between 50-70% of patients are seizure-free after several years (204). What lead to the development of sclerosis is unclear and understanding the mechanism associated with hippocampal sclerosis can be important in developing treatment for the management of intractable epilepsies.

4.1.1.2 Focal Cortical Dysplasia (FCD)

Focal cortical dysplasia is a developmental brain disorder characterised by the development of abnormal brain tissue in areas of the brain, and malformations of cortical development. These lead to delay in cognitive development and seizures. Focal cortical dysplasia is a heterogenous group of cortical lesions and several mutations within the mTOR pathway

genes have been associated with FCD, as well as insults such as infection, inflammation, or hypoxia. FCD is another leading cause of drug-resistant epilepsy and surgery is often required to live seizure-free (207). (More details on FCD can be found in Chapter 5.) The surgery may involves temporal or frontal lobectomy although in most cases usually a lesionectomy is performed to excise the malformed cortical region (208). After surgery, seizure free outcome ranges from 50-75% depending on a number of factors such as aetiologies, extent of lesion and types of surgery (205).

4.1.2 Epigenetics

In addition to transcriptomics, studying chromatic accessibility can provide a more comprehensive understanding of the complex mechanisms of gene regulation. Chromatin accessibility refers to the state of chromatin in the cells and its availability to transcription factors and other DNA-binding molecules. Chromatin is usually wrapped around histone proteins in structures called nucleosomes, which makes DNA inaccessible and this packing informs the accessibility. When chromatin is “open”, this means DNA is more accessible and can be accessible by machinery and actively transcribed. The “closed” state is compact and generally cannot be transcribed. Chromatin accessibility is very dynamic and changes in response to various things such as disease and development and potentially also seizures. It is crucial in gene regulation and proper cell function (86,209).

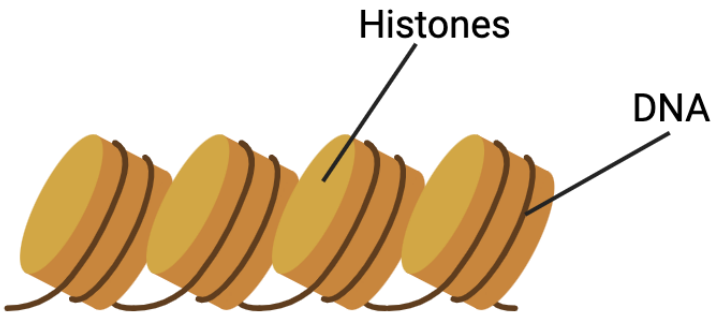


Figure 4.1 – Schematic representation of nucleosome structure. This figure illustrates how a segment of DNA is wound around histone proteins forming nucleosomes. Each nucleosome consists of eight histone proteins (octamer), which packages DNA in order to regulate gene accessibility and expression.

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a technique that can be used to map the open chromatin regions across the genome. ATAC-seq uses a transposable enzyme that inserts sequencing adaptors into the regions of the genome that are not wrapped around histones. The DNA is then fragmented and the regions with the adaptors can be amplified and sequenced (210). Using ATAC-seq at a single-cell level, in combination with RNA-sequencing, can help gain valuable insights into the relationship between chromatin accessibility and gene expression in different cell-types and tissues. The open regions of chromatin are usually regulatory regions such as enhancers and promotors and can be used to identify genes which are being actively transcribed. By integrating these two types, we can identify the regulatory regions that are associated with changes in gene expression and can provide insight into the underlying mechanism (211,212). Additionally, ATAC-seq can provide improved annotation of cell labels to provide a deeper understanding of cell-type heterogeneity, and can also help replicate differential expression analysis (213).

4.1.3 Aims & Objectives

Using samples from human epilepsy patients, this chapter aims to explore cellular and molecular mechanisms that contribute to the susceptibly of epilepsy. Building on work from mouse tissue, this chapter aims to compare results from the two different species across multiple types of epilepsy. By integrating snRNA-sequencing data with genomic insights from GWAS and WES, the aim is to prioritise cell-types that harbour genetic variants influencing epilepsy risk. Additionally, this data will be integrated with publicly available functional perturbation experiments using CRISPRi/a technologies, using a previously described framework of causal inference, we aim to highlight potential genes that could be targeted therapeutically for epilepsy. This information will then be intersected with information from drug databases to offer a pathway to identifying existing therapeutics that can be redirected for epilepsy treatment. This work will also be combined with chromatic accessibility data to gain insight into the epigenetic context and gain a convergent understanding of epilepsy.

Specific goals include:

1. Detecting genes that are differentially expressed in epileptic tissue, contrasting gene expression patterns between affected and control samples to detect cell-type-specific epilepsy signatures.
2. Identifying disrupted biological pathways in epilepsy by looking for significant representation of pathway-associated genes among those differentially expressed in the condition.
3. Pinpointing cell-types that are enriched for different types of epilepsy by correlating genetic data from epilepsy GWAS and WES with cell-type-specific gene expression profiles.

4. Determining potential therapeutic gene targets by analyzing data from CRISPRi/a screens within a causal analysis framework to predict genes with influential roles in epilepsy.
5. Using public databases to search for approved drugs that modulate the activity of identified target genes and that could be repurposed for epilepsy.
6. Combine this information with previously described findings from the mouse study to uncover broader convergent mechanisms that could offer better opportunities for therapeutic success.
7. Incorporate epigenetic information to potentially increase insights into the identified genes and cell-types and their regulation.

4.2 Methods

A detailed summary of the snRNA-sequencing analysis pipeline can be found in Chapter 2.2 and summarised again in Chapter 3.2.6.

4.2.1 Sample Selection

Control samples were received and snap-frozen in liquid nitrogen vapour for 20 minutes and then stored in -80C freezer for long term at the Imperial Brain Bank. Full ethical approvals were received and ethical regulations were agreed by the Imperial College research ethics committee regarding the use of human post-mortem tissue samples. Immunohistochemistry was undertaken on all samples and assessed for beta-amyloid, Tau, TDP43, alpha synuclein and p62. In the selection of controls, samples were excluded if patients had a pre-mortem history of neurological or psychiatric disease or if there was any evidence of neurodegenerative or other significant disease processes. Post-mortem samples were matched as closely as possible by age, sex and brain region to the epilepsy samples.
Samples from individuals with HS and FCD were obtained from the UCL Tissue Bank and were received fresh frozen. FCD samples were from patients with both Type IIa and Type IIb FCD.

4.2.2 Tissue sectioning

- For samples from patients with TLE – tissue was taken from both the hippocampus and the Middle Temporal Gyrus (MTG).
- For samples from patients with FCD – tissue was taken from the pre-frontal cortex (PFC).
- For control samples – tissue was taken from the hippocampus, MTG and PFC (from the same patients).

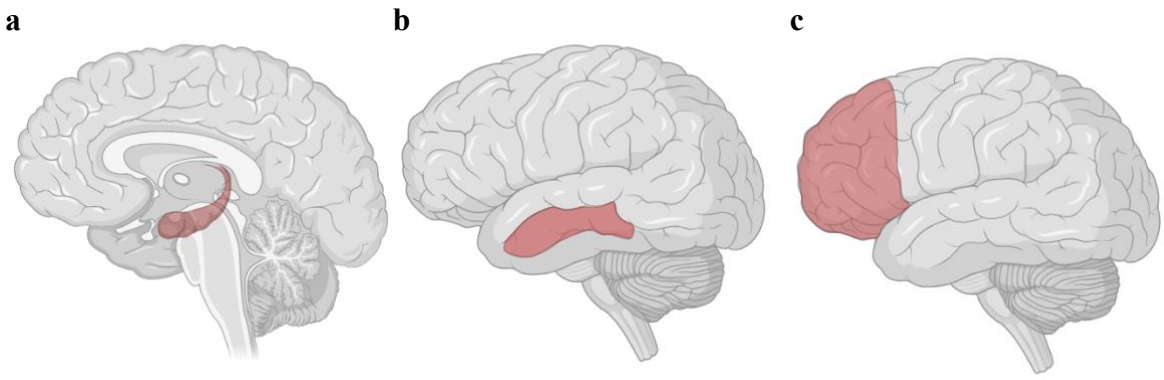


Figure 4.2 – Figures to illustrates the brain regions used in this study. The regions highlighted in red represent the tissue sampled in this analysis. a) The structure shown in this figure is the Hippocampus which is located deep within the medial temporal lobe of the brain. b) the Middle Temporal Gyrus which is located on the lateral side of the temporal lobe and c) the Pre-Frontal Cortex and is located in the frontal lobe of the brain directly behind the forehead.

4.2.3 Isolation of nuclei

The nuclei isolation protocol that was used was adapted from the 10x genomics protocol (164), and the Krishnaswami et al protocol (100). (Described in 3.2.3). Suspensions of single nuclei were prepared by Jeong Hun Ko (laboratory technician in the Johnson group).

4.2.4 Genotyping

DNA from all samples was genotyped using the Illumina 483 Global Screening Array v2.0 at Imperial College London. PLINK (version 1.90b6.18) (214) was used to call sample genotyped using the default settings by Yifei Yang at Imperial College. The genotyped data

was mapped to the hg38 reference genome using bcftools (version 1.9) with the fixref plugin (-m flip option) (215). Imputation was performed on the Michigan Imputation Server (version 1.6.3) using Haplotype Reference Consortium (version r1.1) (216)reference panel of European population with a pre-phasing using Eagle (version 2.4)(217). Only SNPs with imputation score r2 >0.8 were kept.

4.2.5 Data Generation

snRNA-sequencing and snATAC-seq data was generated using the 10x Chromium Next GEM Single-cell Multiome ATAC + Gene Expression kit, targeting 10,000 nuclei per sample. This kit profiles RNA-sequencing and ATAC-seq simultaneously from the same single nuclei. Nuclei are first transposed in a bulk solution with a Transposase which fragments accessible DNA and adds adaptor sequences. Nuclei are then portioned into Gel Beads-in-emulsions (GEMs) using a microfluidic chip. Each GEM contains unique 10x barcodes, which index both DNA and cDNA. Next lysis and reverse transcription of mRNA within GEMs produces barcoded cDNA while transposed DNA is also barcode.
Silane magnetic beads purify the mix and isolate barcode DNA and cDNA, which are then amplified using PCR with P7 sample index sequences. This results into types of library: ATAC libraries with double stranded DNA, with standard Illumina paired-end constructs with P5 and P7 adaptors, and gene expression libraries which contain a cDNA insert also with standard Illumina paired-end constructs with P5 and P7 adaptors. (Figure 4.3).
Libraries were sequenced using the Illumina NovaSeq 6000 system at a minimum sequencing depth of 25,000 paired-end reads per nucleus for GEX and ATAC. The data was sequenced across 31 lanes, with two samples pooled per lane (164).

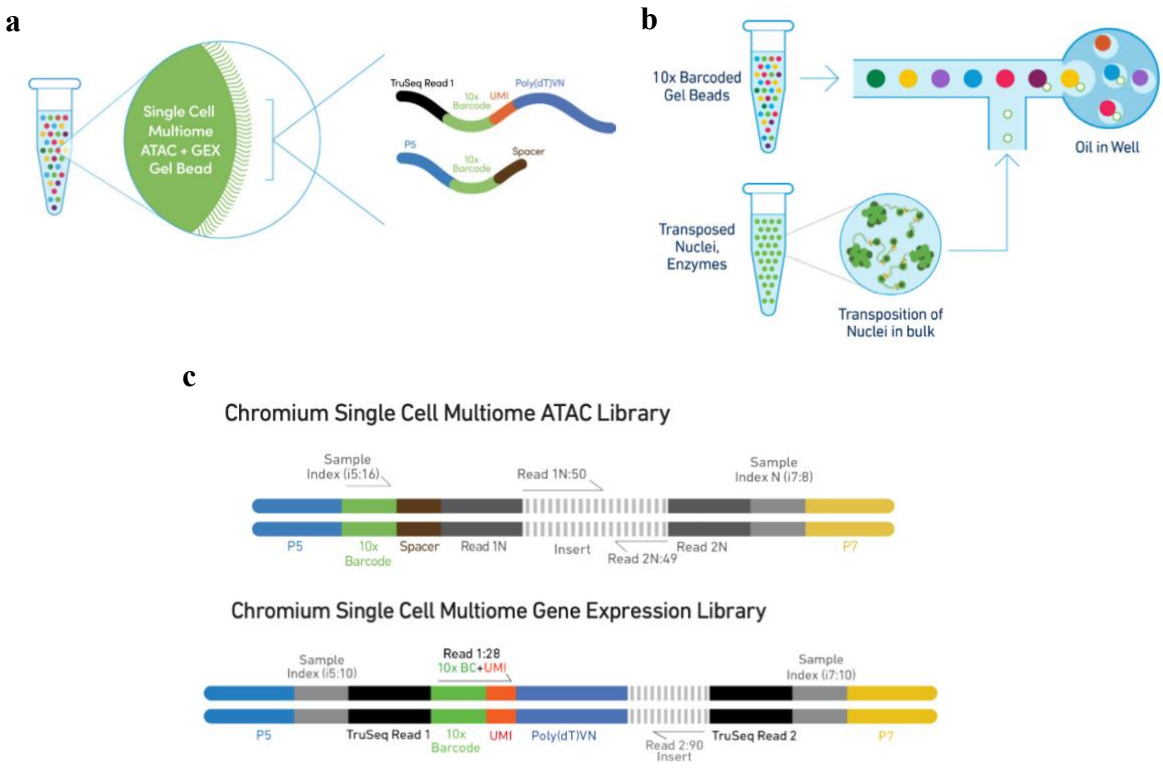


Figure 4.3 - Workflow for Chromium Single-cell Multiome ATAC + Gene Expression. (a) Depicts the 10x single-cell nuclei Barcoded Gel Beads. (b) Demonstrates the partitioning of transposed nuclei into Gel Beads-in-emulsion (GEMs) within a microfluidic chip. (c) Shows the library construction for ATAC and gene expression (GEX) sequencing. The top diagram represents the ATAC library and the bottom diagram illustrates the GEX library. This figure showcases the integrated approach of capturing both chromatin accessibility and gene expression profiles from the same single nuclei.

4.2.6 Mapping

Sequenced reads (FASTQ) are mapped to the genome using Cell Ranger ARC (v 2.0.0), a specialised analysis pipeline that processes Chromium Single-cell Multiome ATAC + Gene Expression sequencing data. The function *cellranger-arc count* is used to perform alignment, peak calling and barcode counting. Reads were mapped to to GRCh38 (GENCODE v32/Ensembl 98) genome using gene annotations from Ensembl v98 (104). The reference transcriptome used in the mapping downloaded from https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest/.

4.2.6.1 Demultiplexing

Sequenced reads were demultiplexed based on the genotypes of each sample using the Demuxlet algorithm with the default settings (218). Demuxlet implements a statistical model which uses the genetic variation in the samples to assign each cell in the pooled lanes to a sample. It does this by calculating the probability of observing reads overlapping a set of SNPs from each cell. It then uses maximum likelihood to determine the most likely source of each cell in a droplet. Demuxlet can also identify doublets which are later removed from the downstream analysis.

4.2.6.2 Peak Calling

Peak calling is the process in which regions of open chromatin are identified in snATAC-seq data. Once the reads have been sequenced and aligned to the genome, the Cell Ranger algorithm identifies regions with a higher density of fragments than expected and calls them as “open” regions. The background distribution of fragments is also modelled to identify a significant threshold for peaks and those below the threshold are removed. Once peaks have been identified, fragments are then quantified per peak.

4.2.7 Data Processing and Quality Control

snRNA-sequencing data and snATAC-seq data are initially processed individually and then combined after the data has been filtered and integrated separately as recommended by the Seurat pipeline.

4.2.7.1 RNA-sequencing

RNA-sequencing was processed according to the steps summarised in Chapter 3.2.6 and detailed in Chapter 2.2.

4.2.7.2 ATAC-seq

Processing and quality control of snRNA-sequencing data was performed using the following steps.

1. Filtering: Cells which have been filtered out during the snRNA-sequencing processing steps are first removed from the snATAC-seq data. As both data sets are coming from the same cells within the same droplet, if a droplet has been called as "empty", a doublet or low quality in one dataset, it should be removed from both.
2. Quality Control: Next, ATAC-seq QC metrics are calculated using the R package *Signac* (v1.5.0) (219) within the *Seurat* (v4.0.0) package. These include nucleosome binding and TSS enrichment scores which predict the quality of the ATAC-seq data. The nucleosome signal reflects the positioning if nucleosomes across the genome, nucleosomes should be less tightly packed in areas of the genome with more open chromatin. The approximate ratio of mono-nucleosomal to nucleosome-free fragments is calculated and cells with a score < 2 are filtered out. The Transcriptional Start Site (TSS) score is a metric that quantifies the degree of enrichment of sequencing reads around the transcriptional start sites of genes. This indicates the specificity of the chromatin accessibility assay and reflects how well the regulatory

regions are being captured. This was defined by the ENCODE project (220). Cells with TSS > 2 are filtered out. The blacklist fraction refers to the set of genomic regions that are typically excluded due to their potential to produce artefactual signals and decrease the quality of the data. Cells with a blacklist ration of >0.05 were also removed. These regions are defined by ENCODE for GRCh38 (220). Filtering is also applied to remove peaks with counts >30000 and <2500 as these contribute to noise and are unlikely informative about true signals of chromatin accessibility.

3. Normalisation: Signac is then used to perform term frequency-inverse document frequency (TF-IDF) normalization. This technique consists of two steps. The first is a normalisation of the binary accessibility matrix which is the "term frequency". This It then calculates how unique the accessibility of a region is across all cells in the data, this is the "inverse document frequency", cells with a higher "IDF" have are more unique and therefore have more impact in identifying particular cell-types. These scores are then used to transform the count accessibility matrix to normalise the profiles. This is then followed by a dimensionality reduction technique; singular value decomposition (SVD) which will reduce the data to the most important axes which is then be used downstream doe clustering (221).
4. Peak unification: Before merging multiple ATAC-seq samples, a unified set of peaks must be created. Peaks are unlikely to be exactly same from sample to sample if peak calling is performed individually and so a common set of peaks should be used across the whole dataset. To create a common peak set, the reduce() function from the *GenomicRanges* package is used to join peaks that overlap or are nearby (within 1000bp). The fragments are then re-quantified based on this new unified peak set.
5. Integration: Signac uses Latent Semantic Integration (LSI) to integrate snATAC-seq data across samples. LSI transforms the high-dimensional data into a lower

dimensional space by capturing the latent structure within the data. This is done by first calculating integration anchors, which are genomic regions with consistent accessibility patterns across samples. These are used as reference points to align the cells and samples. Accessibility profiles are first represented as vectors which quantify the accessibility of each region in each sample. Dimensionality reduction then retains the most informative features and shared patterns can be identified. The LSI embeddings capture the regulatory patterns which are shared across samples and these can be used to cluster cells.

4.2.8 Weighted Nearest Neighbor (WNN)

Once the snRNA-sequencing and snTAC-seq have been individually processed, Weighted Nearest Neighbor (WNN) integration is applied to integrate the two datasets using co-embedding. This is applied within *Seurat* package and can be used to combine many types of multimodal technology. WNN is an unsupervised approach that aims to harmonize multiple datasets by calculating "weights" that reflect the information of each modality and determine the relative importance. It starts by aligning the cells using a graph-based k-nearest neighbour approach for each dataset. It creates networks which are weighted using a combination of cell-cell similarity and dataset specific weights, and identifies a shared embedding space to align cells from different datasets based on their similarities to create a joint definition of cellular state (222). Once the data has been jointly clustered, the annotation steps described in 2.2.6 are applied to assign cell-types to clusters.

4.2.9 Downstream Analysis

The downstream analysis performed on the snRNA-sequencing data is performed as described in Chapter 3.2.8–3.2.9.

These include: Functional enrichment, enrichment of mendelian epilepsy genes, association of heritability, CRISPR brain analysis and intersection with public databases.

4.2.10 Genes4Epilepsy

Genes4Epilepsy is a resource that provides a comprehensive list of genes implicated in monogenic epilepsies compiled from various epilepsy gene panels (223). A master list was manually curated from four clinical diagnostic panels, and genes were annotated for inheritance patterns and associated broad epilepsy phenotypes. This identified over 900 genes with monogenic links to epilepsy, predominantly developmental and epileptic encephalopathies. There are also a small fraction of these genes which are associated with common epilepsy syndromes. This is list is publicly available and regularly updated and can be accessed through https://github.com/bahlolab/genes4epilepsy.

4.2.11 Differential Accessibility

Differential accessibility (DA) analysis is a computational method used to identify differences in chromatin accessibility between different conditions or experimental groups. This is a parallel analysis to differential expression analysis but using ATAC-seq data. The goal of differential accessibility analysis is to detect genomic regions that exhibit significant changes in chromatin accessibility, as this can indicate potential regulatory elements or functional differences associated with disease. This can provide valuable insights into the regulatory landscape of disease and identify potential candidate regions involved. The *Seurat* package was used to find differentially accessible peaks between the 3 pairwise comparisons in this study. This uses the Likelihood Ratio test to identify differentially accessible regions. After differentially accessible peaks are identified, the ClosestFeature() function in Seurat is used to assign peaks to their closest gene.

4.2.12 Computational Tools

The computational analysis in this project was performed on the Imperial High Performance Computer from the Imperial College Research Computing Service, DOI: 10.14469/hpc/2232.

4.2.13 Code availability

The code used in this analysis can be found at https://github.com/Mybzd/RNA-sequencing_analysis/.

4.3 Results

4.3.1 Study summary

In this study, combined snRNA-sequencing and snATAC-seq was performed on 53 patient samples representing two distinct forms of epilepsy, with tissue sourced from three different brain regions. Comparative control samples from corresponding brain regions were also included in the analysis as summarised in Table 1. For Hippocampal Sclerosis (HS), a focal epilepsy which originates in the hippocampus located in the medial temporal lobe of the brain, two brain regions were used; the hippocampus (n=9) and MTG (n=9). For Focal Cortical Dysplasia (FCD), which most commonly occurs in the frontal cortex, samples were sampled from the PFC (n=6) where FCD lesions usually occur. Control samples were taken from the same three brain regions of the patients; hippocampus (n=9), MTG (n=10), and PFC (n=10). Samples were pooled for sequencing and demultiplexed by their genotyped, before they were processed through a series of quality control and filtering steps. A total of 297,887 single nuclei were identified across the 53 samples, which were taken from 25 individuals. On average each sample yielded 5,048 cells, with an average of 1,289 genes detected per cell. The epilepsy samples were surgical fresh frozen and originated from the UCL Brain Bank and control samples were post-mortem and prepared by the Imperial Brain Bank. snRNA-sequencing and snATAC-seq samples were integrated on a joint graph using Weighted Nearest Neighbour (WNN) (222) to minimise variance arising from sample origin and brain region and clustered using the Louvain algorithm.

Group	Hippocampal Sclerosis (HS)		Hippocampal Sclerosis (MTG)		Focal Cortical Dysplasia (FCD)	
Sample Type	Hippocampal Sclerosis	Control	Hippocampal Sclerosis	Control	Focal Cortical Dysplasia	Control
Brain Region	Hippocampus	Hippocampus	Middle Temporal Gyrus	Middle Temporal Gyrus	Prefrontal Cortex	Prefrontal Cortex
Sample Number	9	9	9	10	6	10
Sex	5F 4M	5F 4M	5F 4M	4F 6M	3F 3M	4F 6M
Age	21-56	47-87	21-56	47-87	25-61	47-87

Table 4.1 – Overview of samples used in this study. This table provides a summary of the brain samples used in this analysis. Three comparisons were conspired in this study, split by epilepsy type and brain region; HS (from the hippocampus), HS (from the MTG), and FCD (from the PFC). This table provides information on the number of samples from each disease, the matched controls, brain region the samples were collected from and the age and sex of the patients.

4.3.2 Annotating cell-types

Cell-clusters were systematically assigned to eight major cell-types (Figure 4.4), using a hypergeometric test to test the enrichment of know cell-type markers (176) relative to cluster specific markers. To confirm the accuracy of the cluster assignments, canonical cell-type specific markers were visualised in a heatmap (Figure 4.4C) to provide confirmation that the clusters reflected the expression of unique markers from identified cell-types.

The cell-type composition within the analysed samples was primarily excitatory neurons, foll owed by oligodendrocytes. In total there were 121,263 excitatory neurons, 31,421 inhibitory neurons, 68,000 oligodendrocytes, 31,097 astrocytes, 11,591 microglia, 18,936 oligodendroc yte precursors (OPCs), and 3,958 Endothelial cells and 800 pericytes (Figure 4.4B).

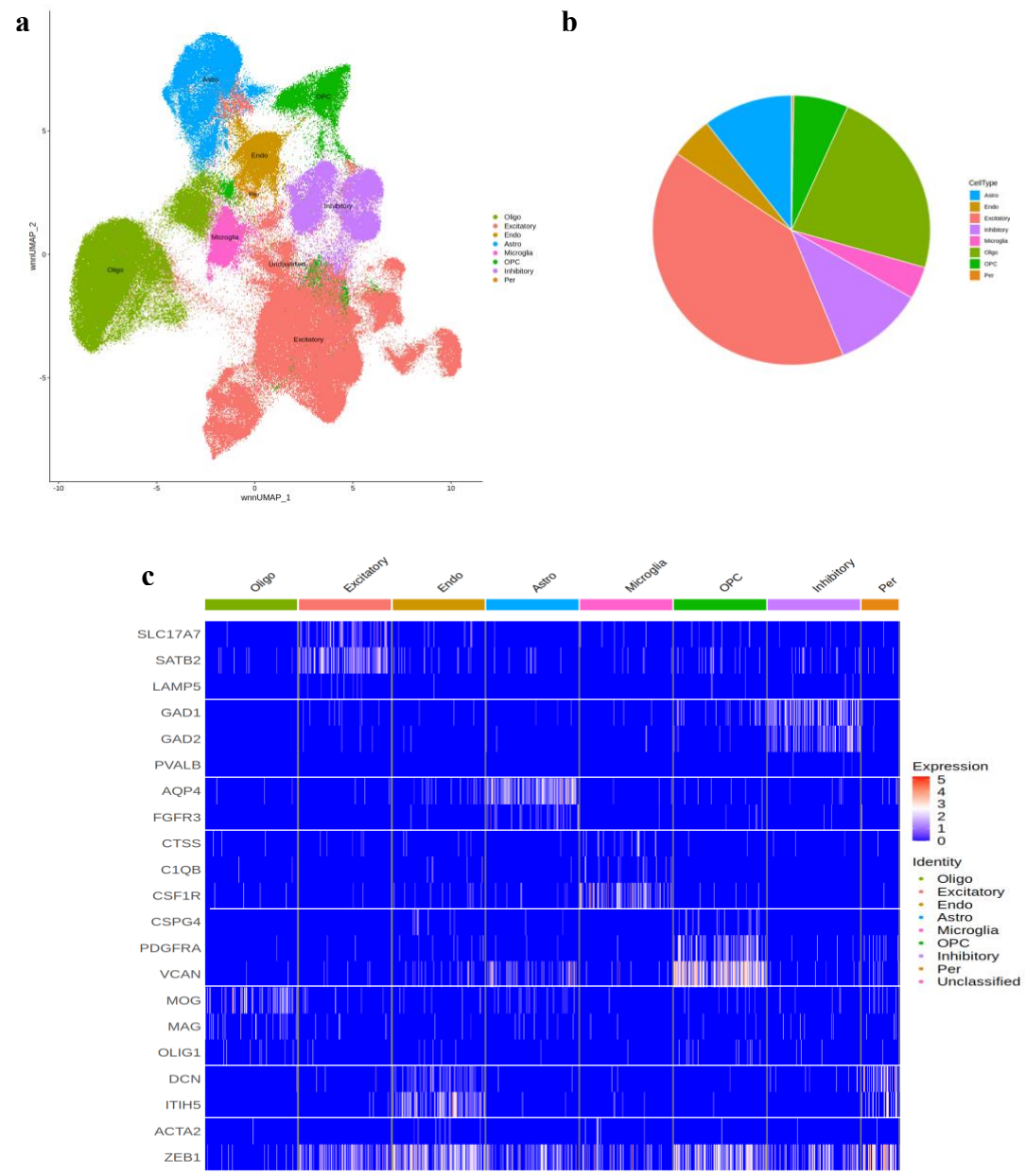


Figure 4.4 – Cell-type annotation of cell-types across brain regions. a) Joint UMAP to show the visualization of all annotated cell-type clusters after WNN embedding. Individual nuclei are colored by cell-type. b) Pie chart representing the proportions of each cell-type across the cells. The dominant cell-type is excitatory neurons followed by oligodendrocytes. c) A heatmap to show the gene expression levels of canonical cell-type markers across the cell-type clusters (down sampled n=200 cells).

4.3.3 Identification of differentially expressed genes in epilepsy

Following cell-type annotation differential expression (DE) analysis was performed in each cell-type separately across the three pairwise comparisons shown in Table 1. DE analysis aims to understand transcriptional changes associated with each cell-type in the context of disease. DESeq2 (125) was used to perform this analysis across the three pairwise disease-control comparisons. A "pseudo-bulk" approach is used to first sum gene counts across all cells per sample. The comparisons tested were HS (hippocampus) (n=9) vs Control (hippocampus) (n=9), HS (MTG) (n=9) vs Control (MTG) (n=10), and FCD (n=6) vs Control (PFC) (n=10). In this analysis, only genes with an expression of 3 counts per gene per sample were tested to minimise noise from lowly expressed, uninformative genes. Genes were considered significant if they were below the FDR threshold of 0.05. Figure 4.5 shows the number of DEGs in each comparison in each cell-type, by direction of change.

Significantly more DEGs are seen in FCD compared to the HS comparisons in most cell-types except Microglia and Pericytes, for both up- and down- regulated genes. Interestingly, endothelial cells exhibited the greatest changes in gene expression in FCD followed by excitatory neurones which had 1,280 DEGs in total. In HS, DE was less pronounced in both brain regions, however distinct patterns of gene expression could be seen in terms of the predominant cell-types. In the hippocampal comparison, the highest DEGs were seen in both neuronal cell-types as well as oligodendrocytes. In the MTG on the other hand, neuronal DE is very low and the highest number of DEGs are in the glial cell-types Astrocytes and Microglia, which are much lower in the hippocampus. This pattern suggest that the different brain regions could be contributing distinctly in the formation of hippocampal sclerosis, with neuronal abnormalities originating in the hippocampus and glial in the MTG. Generally, there was a trend toward more gene upregulation across all comparisons.

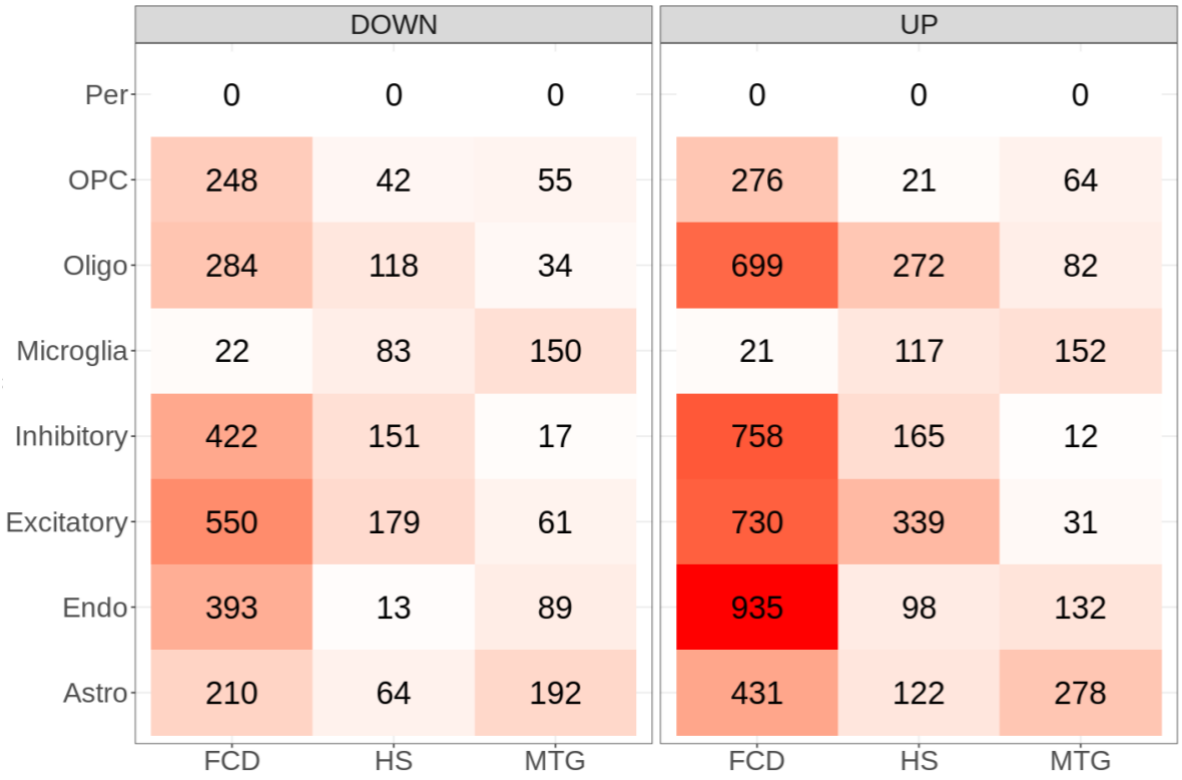


Figure 4.5 – Heatmaps to show numbers of cell-type specific differentially expressed genes across the three comparisons: HS, MTG and FCD. The colour intensity represents the degree of differential expression. The right hand panel represents the up-regulated genes and the left hand panel represents the down-regulated genes. Genes are considered differentially expressed at a threshold of 5% FDR. All results are available in Appendix Table 4.

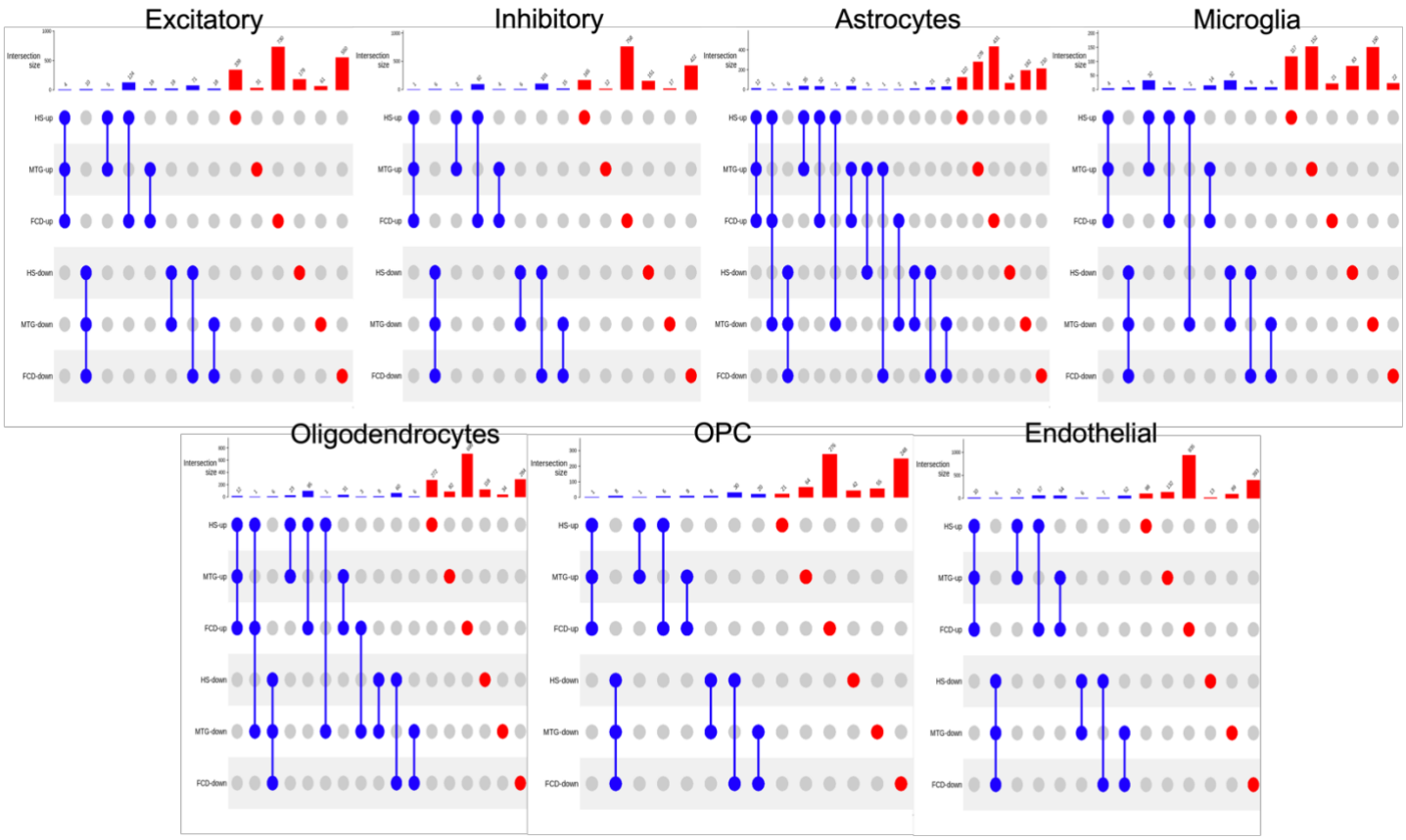


Figure 4.6 – UpSet plots to visualise the overlap of DEGs between the disease comparisons and brain regions in each of the 7 cell-types with significantly differentially expressed genes. Intersections are shown between up- and down- regulated gene sets separately. The blue dots represent gene sets where there are intersecting genes and the red dots represent the numbers of genes with no intersections. The size of the gene set is shown on bar plots across the top of the panels.

To look for converging molecular signals across different epilepsies, overlaps of the differentially expressed gene sets per cell-type are shown in Figure 4.6. Consistent DEGs can be see across all three comparisons in the same direction across all the cell-types, suggesting there is evidence of common genes being dysregulated across various forms of epilepsy and multiple brain regions. The genes that are overlapping across all comparisons in the same

direction are detailed in Table 2. These genes indicate a potential commonality in the transcriptional changes associated with epilepsy regardless of the type of epilepsy or the within the brain. Of these genes, there are several which have been previously connected to epilepsy; *OSGEP, B3GALNT2, WDFY3*, and *TCF4* according to the genes4epilepsy gene panel (223), a comprehensive resource of genes that have been implicated in various aspects of epilepsy. These results indicate a commonality in the transcriptional features of the disease regardless of the type of epilepsy or the location within the brain.

CellType		Intersection - UP		Intersection - DOWN
Excitatory	4	LINC01006 ZNF337 OSGEP AC105411.1	10	ANKH ATXN1 UMAD1 LRRTM3 FRY PRKD1 STAMBPL1 AL133166.1 SYT16 FLRT2
Inhibitory	1	PDE1A	6	ARID5B LRRTM3 SLC2A13 LRCH1 STXBP5 AFF2
Astrocytes	12	SMAP2 AHCTF1 PPM1B ZBTB16 FMNL3 PER1 BCR SREBF2 AL359232.1 B3GALNT2 SLC19A2 OSGEP	6	YAE1 UTP23 ZBTB1 NPAS2 ABCA1 FAM182A
Microglia	4	KLF7 SREBF2 SLC44A2 CSF2RA	7	WDFY3 PTPRE BACH1 ARNTL DNAJC3 HIST1H2AC GNG10
Oligodendrocytes	12	KLHL5 LINC01006 CSGALNACT1 FMNL3 PER1 ZNF337 FOXP2 AL359232.1 PRKCH AC012150.1 CSF2RA OLR1	6	XPO1 SSBP2 APOLD1 TCF4 CD22 SH3RF1
OPC	1	PPP1R13B	8	CLASP1 ZHX2 IGF1R TOX MMP16 PDZRN4 ALK MTSS2
Endothelial	10	PDE1A FNIP2 FOXO3 LDLRAD4 CERS6 BAIAP2 ADAMTSL4 TIPARP TIMP3 NR4A1	6	IFNAR1 ARHGAP29 CFH ABCA1 TGM2 ETS2

Table 4.2 – Table to show the numbers of genes that are consistently differentially expressed in all comparisons in the different cell-types, and the specific genes that are in common. Genes are shown separately for up- and down- regulated separately.

Notably, when considering genes that were convergently DE across two comparisons, more differentially genes can be observed in common across the HS and FCD comparison, than the

HS and MTG comparison (Table 4.2). This is seen in most cell-types other than Astrocytes and Microglia. This could suggest a less active role of the MTG in the manifestation of epilepsy, and perhaps a more homeostatic compensatory role through astrocytes and microglia.

For most cell-types, there are no overlaps between upregulated and downregulated genes between comparisons, except for in astrocytes and oligodendrocytes where some genes are upregulated in the HS and FCD comparisons but downregulated in the MTG. When looking at the genes overlapping in just HS and FCD, there are 48 genes which are present in the genes4epilepsy list, 27 of which are DE in the neuronal cell-types in both comparisons. Conversely, there are only 10 genes in common between HS and MTG which are in the genes4epilepsy list and these are all expressed in Microglia, Astrocytes and Oligodendrocytes. This again suggest the HS and FCD comparisons are more similar in terms of the role they play in seizure generation.

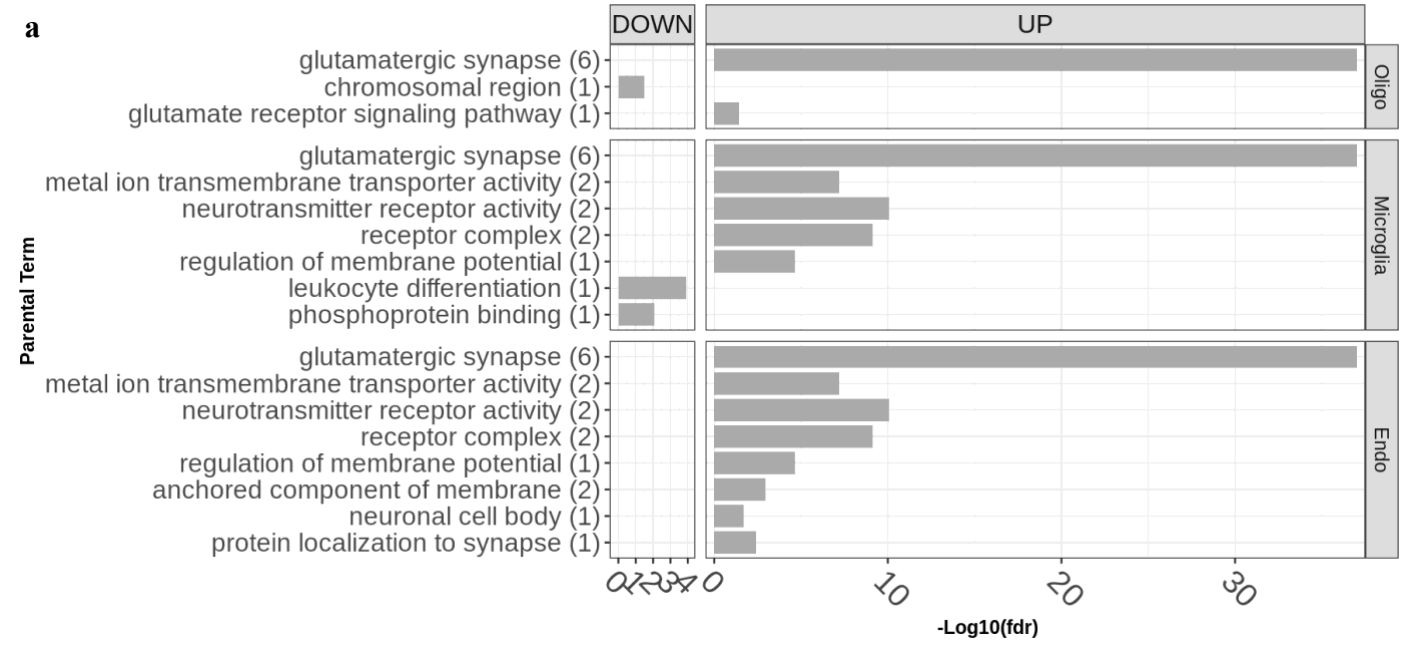
4.3.4 Functional enrichment of cell-type specific DEGs

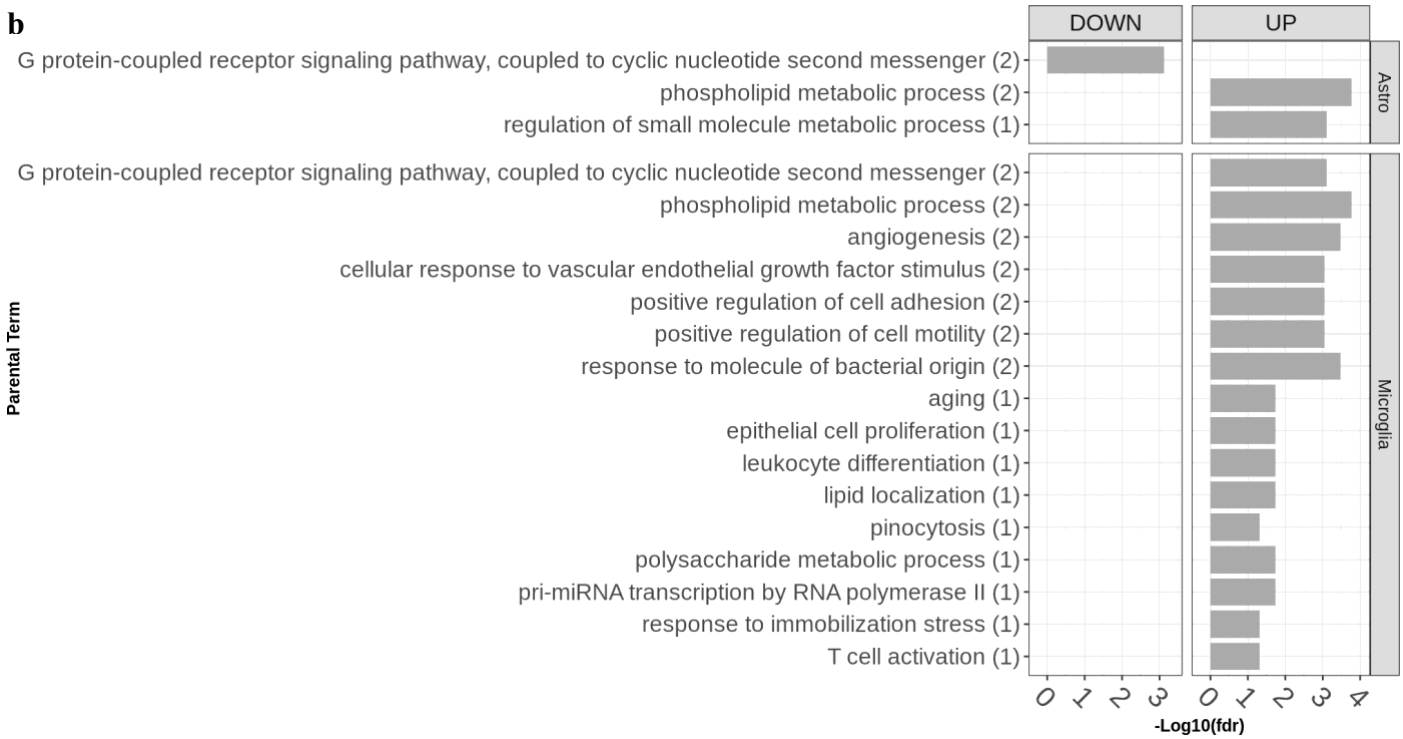
Gene Ontology (GO) enrichment analysis can offer a better understanding of the biological relevance of DEGs in each cell-type by highlighting mechanisms that link DEGs to diseases processes. Over-representation analysis (ORA) was used to annotate the DEGs to biological processes, molecular functions and cellular components. The enriched terms were then categorised by broader parental terms using sematic similarity analysis (130). Significant terms at FDR<0.05 are shown in Figure 4.7.

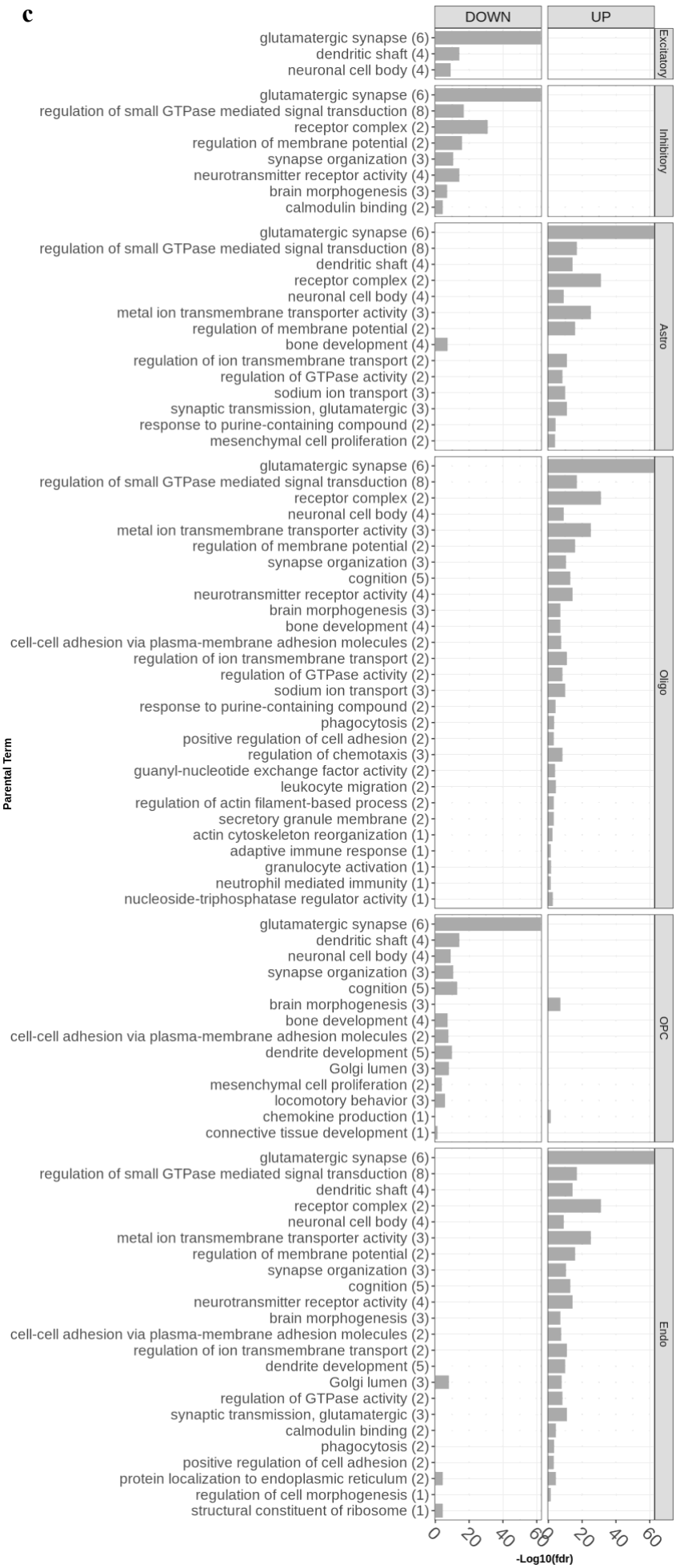
In the HS and MTG comparison, GO terms are only functionally enriched in glial cell-types, and not neurones, and these are predominantly upregulated. In the FCD, functionally enriched terms were seen in all cell-types but microglia where there were very few DEGs seen.

The “glutaminergic synapse” is seen to be enriched in the upregulated DEGs in microglia, oligodendrocytes and endothelial cells in the HS comparison and the FCD comparison, but enriched in the downregulated DEGs in the excitatory and inhibitory neurones, and OPCs in the FCD comparison. Other synapse related terms such as “glutamate receptor signalling pathway”, “neurotransmitter receptor activity” and “regulation of membrane potential” are also seen to be upregulated in glia in the HS comparison and similarly in the FCD comparison. The terms that are significantly enriched in the glial cells in the MTG comparison on the other hand, are more related to inflammatory and metabolic pathways such as “leukocyte differentiation”, “t-cell activation” and “phospholipid metabolic process”, and are primarily downregulated, again suggesting a potential compensatory role in the glia of the MTG. In the FCD comparison, oligodendrocytes and endothelial cells also show upregulation of inflammatory terms such as “phagocytosis” and “adaptive immune response”. Oligodendrocytes and Astrocytes show regulation of common signalling related terms, where as OPCs show more similar profiles to the neurones where enrichment is only seen in the downregulated genes. These findings predominantly point to distinct functions of glial cells across the various epilepsy models, as well as highlighting the involvement of the glutaminergic synapse in HS and FCD but not in the MTG comparison.

Figure 4.7 – Over-representation analysis of Gene Ontology (GO) terms across DEGs per cell-type in the a) Hippocampal Sclerosis in the hippocampus (HS) b) Hippocampal Sclerosis in the MTG (MTG), and Focal Cortical Dysplasia in the PFC. The enrichment of reduced GO terms associated with cell-type specific DEGs is down for up-regulated terms (right) and down- regulated terms(left) with respect to control. Sematic similarity was used to reduce original GO “child” terms, the numbers of which are shown in brackets on the y-axis. The x-axis represents the -log10(FDR) of the most significant GO term associated with that parental term. All results are available in Appendix Table 5.


a
DOWN
UP
Oligo
Microglia
Endo
Parental Term
glutamatergic synapse (6)
chromosomal region (1)
glutamate receptor signaling pathway (1)
glutamatergic synapse (6)
metal ion transmembrane transporter activity (2)
neurotransmitter receptor activity (2)
receptor complex (2)
regulation of membrane potential (1)
leukocyte differentiation (1)
phosphoprotein binding (1)
glutamatergic synapse (6)
metal ion transmembrane transporter activity (2)
neurotransmitter receptor activity (2)
receptor complex (2)
regulation of membrane potential (1)
anchored component of membrane (2)
neuronal cell body (1)
protein localization to synapse (1)
0
10
20
30
-Log10(fdr)


b
DOWN
UP
Astro
Microglia
Parental Term
G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger (2)
phospholipid metabolic process (2)
regulation of small molecule metabolic process (1)
G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger (2)
phospholipid metabolic process (2)
angiogenesis (2)
cellular response to vascular endothelial growth factor stimulus (2)
positive regulation of cell adhesion (2)
positive regulation of cell motility (2)
response to molecule of bacterial origin (2)
aging (1)
epithelial cell proliferation (1)
leukocyte differentiation (1)
lipid localization (1)
pinocytosis (1)
polysaccharide metabolic process (1)
pri-miRNA transcription by RNA polymerase II (1)
response to immobilization stress (1)
T cell activation (1)
0
1
2
3
4
-Log10(fdr)


c
DOWN
UP
Excitatory
glutamatergic synapse (6)
dendritic shaft (4)
neuronal cell body (4)
Inhibitory
glutamatergic synapse (6)
regulation of small GTPase mediated signal transduction (8)
receptor complex (2)
regulation of membrane potential (2)
synapse organization (3)
neurotransmitter receptor activity (4)
brain morphogenesis (3)
calmodulin binding (2)
Astro
glutamatergic synapse (6)
regulation of small GTPase mediated signal transduction (8)
dendritic shaft (4)
receptor complex (2)
neuronal cell body (4)
metal ion transmembrane transporter activity (3)
regulation of membrane potential (2)
bone development (4)
regulation of ion transmembrane transport (2)
regulation of GTPase activity (2)
sodium ion transport (3)
synaptic transmission, glutamatergic (3)
response to purine-containing compound (2)
mesenchymal cell proliferation (2)
Oligo
glutamatergic synapse (6)
regulation of small GTPase mediated signal transduction (8)
receptor complex (2)
neuronal cell body (4)
metal ion transmembrane transporter activity (3)
regulation of membrane potential (2)
synapse organization (3)
cognition (5)
neurotransmitter receptor activity (4)
brain morphogenesis (3)
bone development (4)
cell-cell adhesion via plasma-membrane adhesion molecules (2)
regulation of ion transmembrane transport (2)
regulation of GTPase activity (2)
sodium ion transport (3)
response to purine-containing compound (2)
phagocytosis (2)
positive regulation of cell adhesion (2)
regulation of chemotaxis (3)
guanyl-nucleotide exchange factor activity (2)
leukocyte migration (2)
regulation of actin filament-based process (2)
secretory granule membrane (2)
actin cytoskeleton reorganization (1)
adaptive immune response (1)
granulocyte activation (1)
neutrophil mediated immunity (1)
nucleoside-triphosphatase regulator activity (1)
OPC
glutamatergic synapse (6)
dendritic shaft (4)
neuronal cell body (4)
synapse organization (3)
cognition (5)
brain morphogenesis (3)
bone development (4)
cell-cell adhesion via plasma-membrane adhesion molecules (2)
dendrite development (5)
Golgi lumen (3)
mesenchymal cell proliferation (2)
locomotory behavior (3)
chemokine production (1)
connective tissue development (1)
Endo
glutamatergic synapse (6)
regulation of small GTPase mediated signal transduction (8)
dendritic shaft (4)
receptor complex (2)
neuronal cell body (4)
metal ion transmembrane transporter activity (3)
regulation of membrane potential (2)
synapse organization (3)
cognition (5)
neurotransmitter receptor activity (4)
brain morphogenesis (3)
cell-cell adhesion via plasma-membrane adhesion molecules (2)
regulation of ion transmembrane transport (2)
dendrite development (5)
Golgi lumen (3)
regulation of GTPase activity (2)
synaptic transmission, glutamatergic (3)
calmodulin binding (2)
phagocytosis (2)
positive regulation of cell adhesion (2)
protein localization to endoplasmic reticulum (2)
regulation of cell morphogenesis (1)
structural constituent of ribosome (1)
Parental Term
0
20
40
60
-Log10(fdr)

4.3.4 Gene list enrichment results

To determine whether the cell-type specific DEGs were likely to be disease related, enrichment analysis was performed using a hypergeometric test against two sets of known epilepsy genes. There were; a set of high confidence monogenic epilepsy genes from a Genomics England testing panel (167), and a compiled list of genes with genome-wide significant loci from epilepsy susceptibility GWAS (142). These were tested against a background gene set of all genes expressed in a particular cell-type. An enrichment is considered to be significant at an FDR threshold of 5%, these are shown in Figure 4.8. Monogenic epilepsy are only enriched within one cell-type within the DEGs across all 3 comparisons, and this is downregulated genes in FCD inhibitory neurons. There are more enrichments seen in common epilepsy genes GWAS studies; in the HS comparison, the upregulated microglial genes are enriched, whereas in the MTG comparison, the downregulated excitatory and the upregulated inhibitory DEGs are enriched for common epilepsy genes. It is surprising to see enrichment in the neuronal cell-types in MTG as there were very few DEGs in these cell-types in this comparison, however these FDR values were marginal significant (Inhibitory: FDR=0.049) and (Excitatory: FDR=0.031).

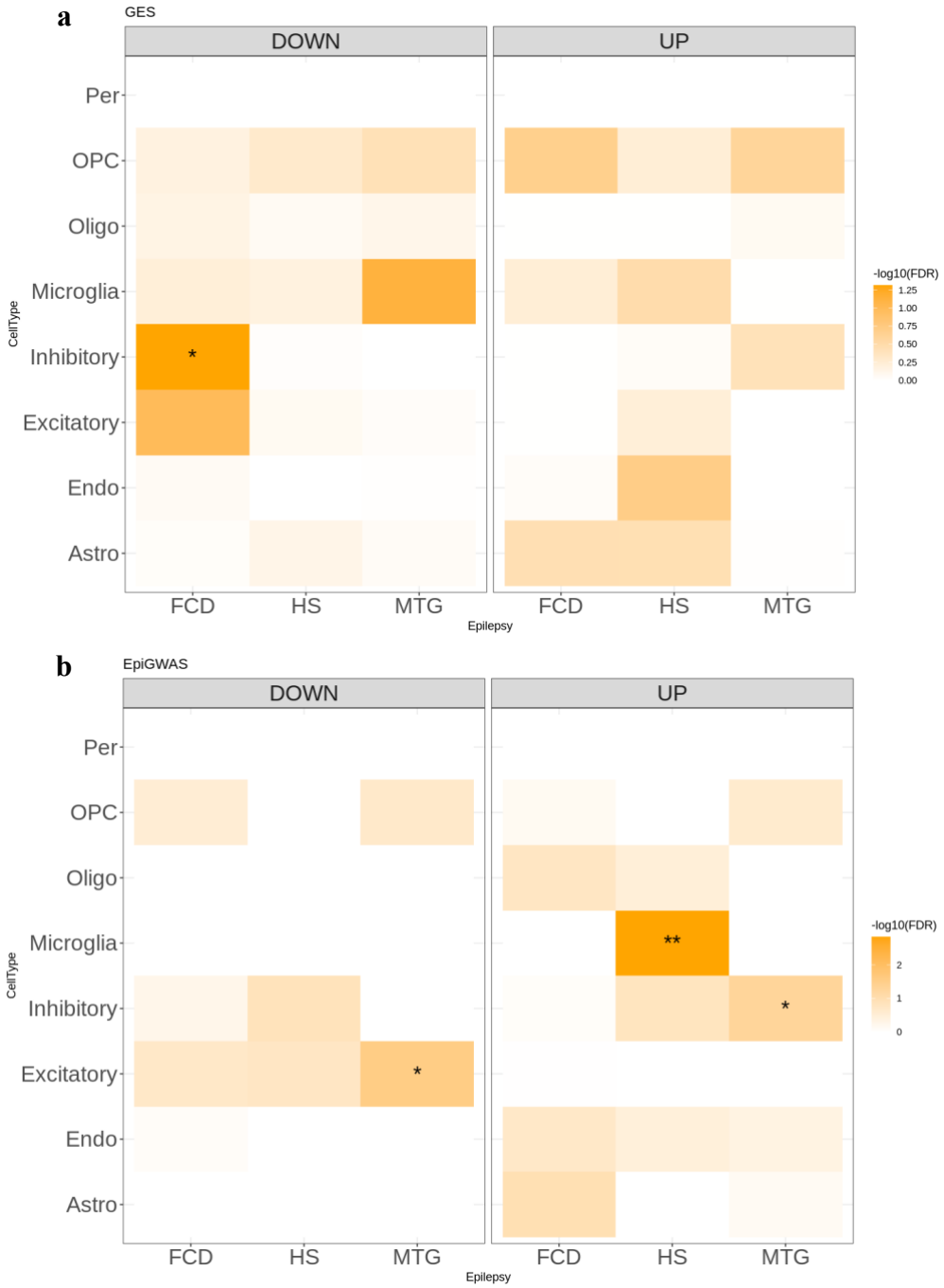


*Figure 4.8 – Heatmaps to show the enrichment of epilepsy-related gene sets within the DEGs across the three DE comparisons. Heatmaps show enrichment FDR values from a fishers exact test using a) a set of monogenic genes from Genomics England, which are categorised as "green" meaning they have a strong level of evidence linking them to epilepsy, and b) genes identified in previous epilepsy GWAS studies. These were tested for enrichment within the cell-type and model specific up-(right) and down-(left) regulated DEGs. Heatmaps are coloured by degree of significance and nominal significance is indicated using: ***FDR < 0.001; **FDR ≤ 0.01; *FDR ≤ 0.05.*

4.3.5 Integration of differentially expressed genes with genetic data

While differential expression analysis serves as a crucial methodology for elucidating the genes and molecular mechanisms that are implicated in disease, observed differentially expressed genes between disease and control patients can observed for several reasons. They can be causing the disease phenotype, they may be consequences of the disease development or they may be simply correlated to the disease. Association of the DEGs with epilepsy heritability can help us better understand the relationship of cell-type specific DEGs to the disease. To address this, the association between cell-type specific DEGs and epilepsy heritability can be assessed through stratified LD score regression (sLDSC) (224). This tool leverages summary statistics from large-scale epilepsy GWAS, to determine the enrichment of common genetic variants that may be mediating epilepsy risk, assuming that genetic variants linked to epilepsy may have a causative role in disease susceptibility. The most recent ILAE epilepsy susceptibility GWAS (33) was used to perform this analyses and performed for both generalized and focal epilepsies separately, as well as all epilepsy types combined. Enrichment was tested for all groups of cell-type specific DEGs (with FDR < 0.05) across each comparison (Figure 4.9).

Significant enrichment was observed across most gene sets, and most significantly in the upregulated genes in the HS comparison, and FCD comparisons. In HS all the upregulated gene sets are significant for generalised epilepsy, and the inhibitory and excitatory for focal epilepsy. For the FCDs, all cell-types are significant in the upregulated gene sets in generalised epilepsy and astrocytes, microglia and excitatory neurones for the focal epilepsy. There is also high significance in the downregulated gene sets in HS and FCD, most significantly in excitatory and inhibitory neurons. In the MTG comparison, the most significant results come from the astrocyte and microglia gene sets, where both up- and down- regulated genes are enriched for variants from both focal and generalised epilepsies.

Neuronal cells did not exhibit significant enrichment in the MTG comparison, emphasising the observation that these are less important in the MTG in comparison to HS and FCD. Conversely, glial cells particularly microglia and astrocytes are repeatedly implicated in the MTG comparison.

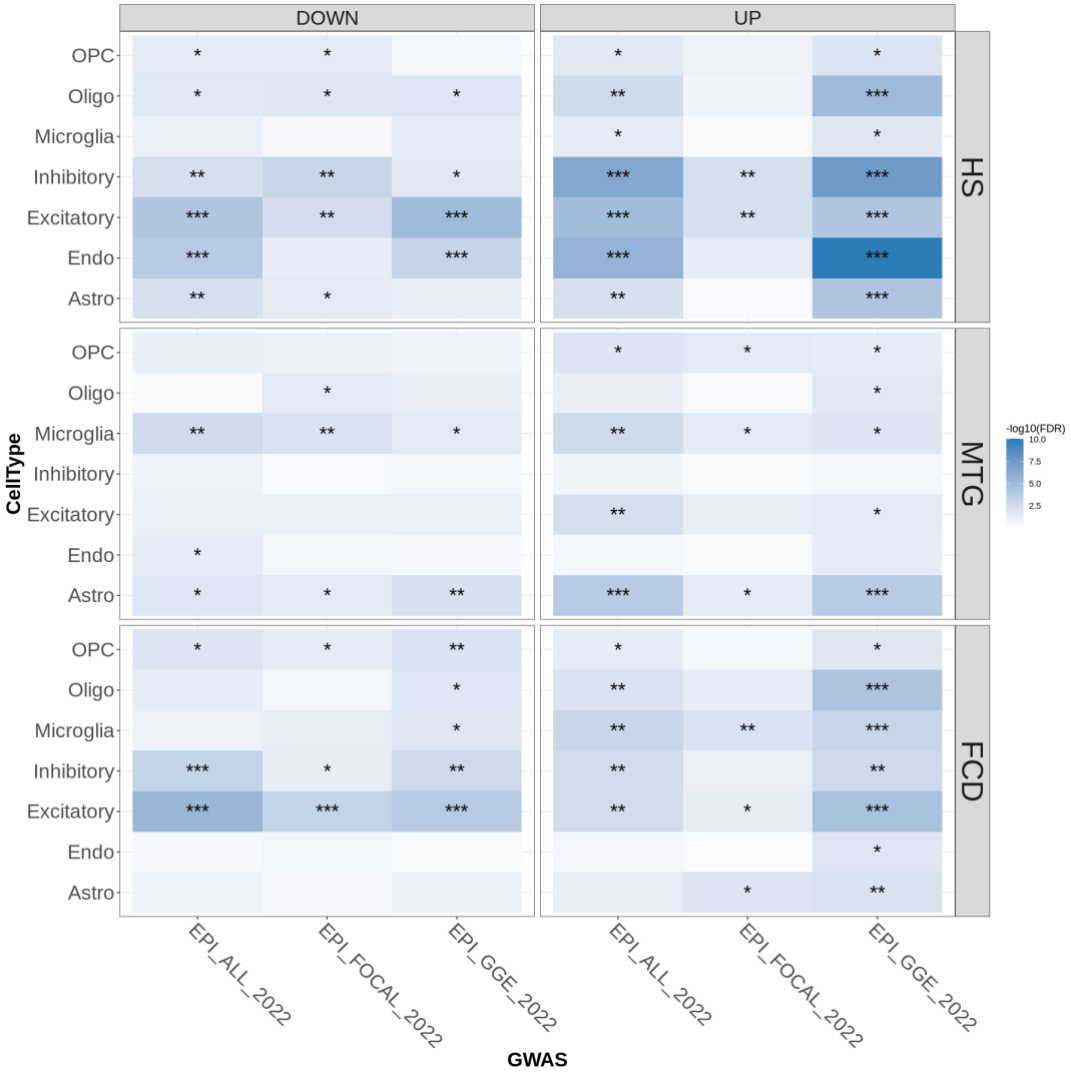


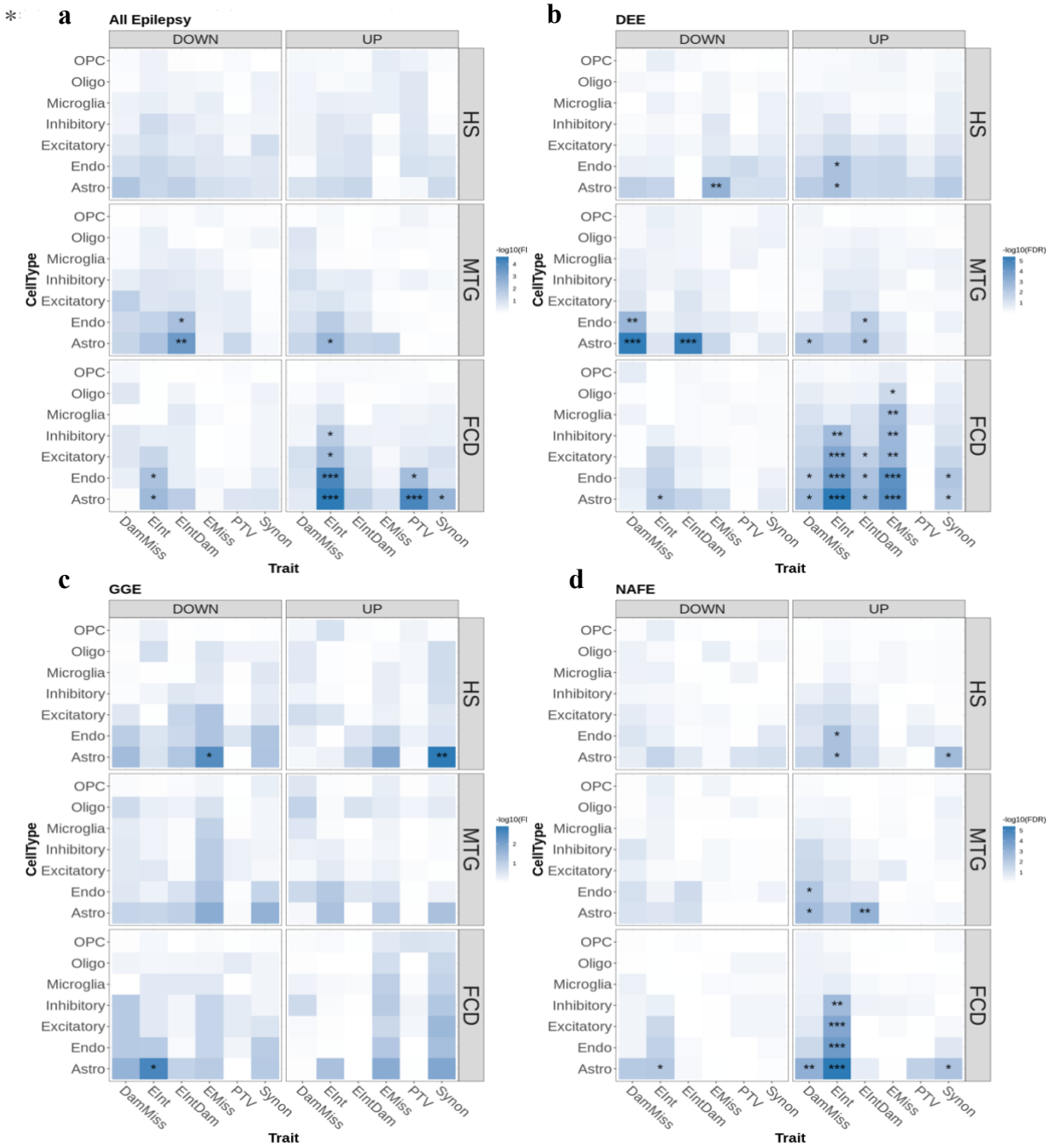
Figure 4.9 – Association of heritability analysis of cell-type specific DEGs using LD Score Regression. Cell-type and model specific DEGs were used to identify genetic association with epilepsy risk using ILAE 2022 GWAS studies. The analysis is run for the Focal Epilepsy, Generalised Epilepsy and All Epilepsies denoted EPI_FOCAL_2022, EPI_GGE_2022, and EPI_ALL_2022. Heatmaps are coloured by degree of significance and nominal significance is indicated using: ***FDR < 0.001; **FDR ≤ 0.01; *FDR ≤ 0.05.

Whilst LDSC analysis can identify whether genes DE in epilepsy are enriched for genetic association to the disease, this analysis (and that based on enrichment of known epilepsy genes described earlier) do not provide information on directionality of the relationship. As well as looking at the enrichment of common variants from GWAS, we can further look at the enrichment of rare variants associated with epilepsy previously identified in Whole Exome Sequencing (WES) studies. EPI25K (225) is a WES dataset from the ILAE Genomics consortium of 30,000 epilepsy patients. This dataset considers epilepsy in four broad categories: severe developmental and epileptic encephalopathies (DEEs), genetic generalized epilepsy (GGE), non-acquired focal epilepsy (NAFE) and All epilepsy. The available data consisted of gene-level association p-values derived from collapsing analysis of deleterious rare variants, for each phenotypic category. The tool GSA-SNP2 (169) was then used, to test the enrichment of (Chapter 3.3.6) rare variants in the DE gene sets using ranked p-values for the following categories of coding variant: Synonymous (Synon), Protein Truncating Variants (PTV), missense (EMiss), damaging missense (DamMiss), intolerant missense (EInt) and damaging and intolerant missense (EIntDam). The results are summarised in Figure 4.10. A convergent significant signature is seen in genes downregulated in astrocytes

There was a significant enrichment for DEGs in astrocytes, in all three comparisons. The highest significance was seen in upregulated genes in FCD for genes that are preferentially impacted by intolerant missense mutations, particularly in DEE and NAFE.
In DEE in particular, the upregulated genes in FCD were strongly enriched in nearly all cell-types but OPCs, predominantly in genes impacted by missense, intolerant, and damaging mutations. HS and MTG also showed enrichment for the upregulated genes particularly in DEE and NAFE. Downregulated genes in HS and MTG also showed significant enrichment for genes with damaging and missense variants in DEE. Downregulated genes in HS and FCD also showed significant enrichment for genes with damaging and missense variants in GGE.

This analysis shows that looking at deleterious high penetrance variants means that we can deduce more direct inference of causality between cell-type specific DEGs and epilepsy. These results alone do not clarify whether the observed DEGs contribute to the development of epilepsy or if they emerge as a result of the body's compensatory, homeostatic responses to the disease. Nevertheless, the observation that genes, which when deleteriously mutated are known to cause epilepsy, are downregulated in conditions like FCD and HS, implies that their reduced expression could play a role in fostering the hyperexcitability characteristic of epilepsy. The deleterious genes enriched in upregulated DEGs are more difficult to interpret but could imply the involvement of these genes in compensatory homeostatic mechanisms which are upregulated as a secondary response to epileptic seizures.

Figure 4.10 – Enrichment of heritability of rare variants using GSA-SNP2. These heatmaps display results of heritability enrichment of ultra-rare variants associated with epilepsy subtypes including (a) All Epilepsy, (b) developmental and epileptic encephalopathies (DEEs), (c) severe genetic generalized epilepsy (GGE), and (d) non-acquired focal epilepsy (NAFE). The following coding variants were included in the analysis: Synonymous (Synon), Protein Truncating Variants (PTV), missense (EMiss), damaging missense (DamMiss), intolerant missense (EInt) and damaging and intolerant missense (EIntDam).Heatmaps are coloured by degree of significance and nominal significance is indicated using:



4.3.7 Comparison to mouse data

In order to identify strongly implicated genes in epilepsy, we looked for signals that are convergent across both human and mouse data. snRNA-sequencing data from 85 samples from three mouse models of epilepsy were previously analysed using the same computational pipeline (Chapter 3). These models were a Pilocarpine induced post-status epileptic model of TLE (PILO), a Self-Sustained Status Epilepticus model (SSSE) of temporal lobe epilepsy, and a genetic model of Focal Cortical Dysplasia (FCD).

Due to low numbers of DE genes in the SSSE model, there are no genes that are convergent across all models and comparisons, however a few genes converge across HS, MTG, FCD human, PILO and FCD mouse in excitatory neurones and inhibitory neurones. The overlaps between these comparisons are shown in Figure 4.11. *STAMBPL1*, *SYT16* are downregulated in excitatory neurones across all 5 comparisons and ARID5B is downregulated in inhibitory neurones in all 5 comparisons. *STAMBPL1* is a paralog of *STAMBP* which is Cytokine-mediated signal transduction in the JAK-STAT, known to cause early-onset refractory epilepsy (226). SYT16 is a prognostic biomarker that has been shown to be correlated with immune infiltrates (227). Differentially methylated regions in generalised epilepsy were shown to be associated with *ARID5B* (228).

Other notable genes that are DE across multiple comparisons include *GABRB2* which is downregulated in in 4 of the comparisons in inhibitory neurones; HS, FCD human, PILO and FCD mouse. Loss of function mutations in *GABRB2* as well as other GABA receptor subunits are well established in the cause of epilepsy and GABAergic medications are an established means of seizure suppression (229).

PIK3CA and *PLCB1* are both genes which have been shown to activate the PI3K/AKT/mTOR pathway (230,231). However, these genes were downregulated in excitatory neurones in both FCD comparisons and HS. The directionality of effect on mTOR

is opposite in effect to what would be expected of a causal mechanism and could be a compensatory response to lower mTOR activation.

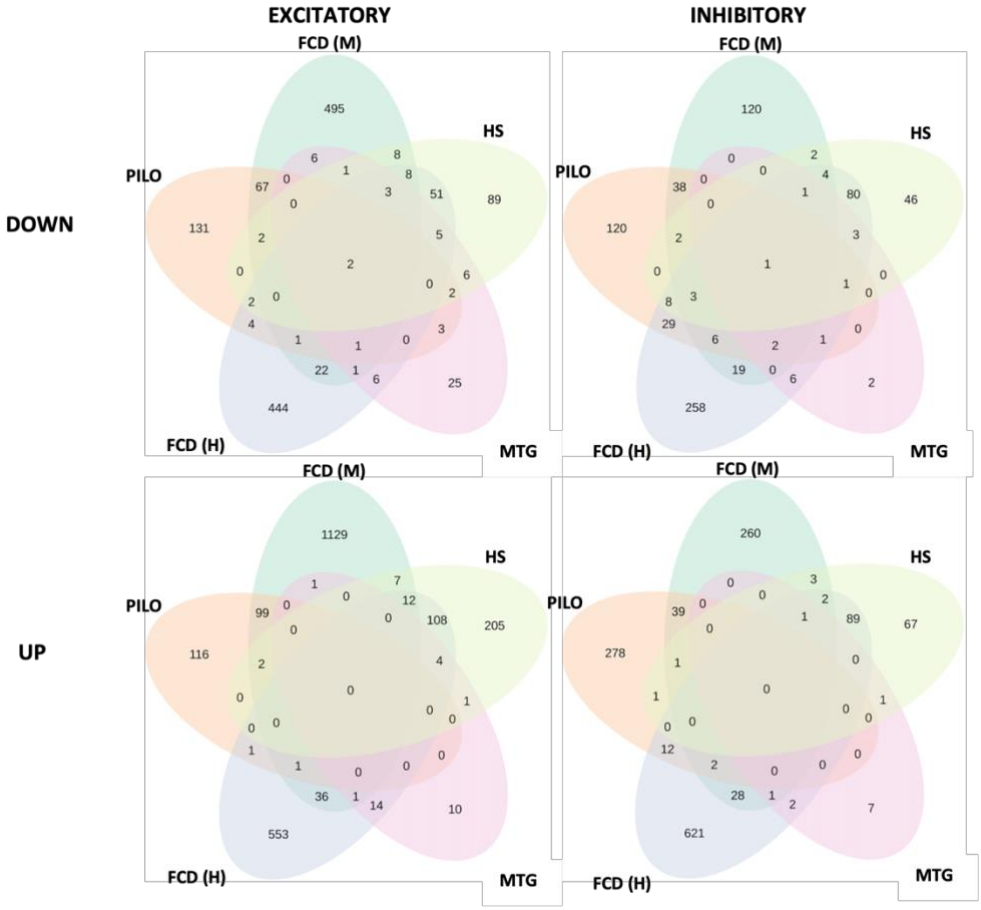


Figure 4.11 – Venn diagrams to show the overlapping differentially expressed genes in up- and down- regulated genes in excitatory and inhibitory neurones.

In terms of functional annotation of GO terms, upregulated genes in mouse models also reveal significant associations with “glutaminergic synapse” within astrocytes and downregulated in excitatory neurones as in the human data. The inhibitory neurones, however, show an upregulation of “glutaminergic synapse” in mice but downregulation in humans. There is a strong implication for inflammatory terms in the excitatory neurones in

mouse FCDs which is not seen the in human data, however, there is indication of inflammatory terms in the oligodendrocytes in the human which is not seen in the mouse. Neither comparison showed any immune function enrichment in microglial DE.

Enrichment of epilepsy gene lists highlighted a consistent pattern in inhibitory neurones; genes associated with monogenic epilepsy were downregulated both comparisons of FCD. Similarly, excitatory neurones displayed significant downregulation of genes linked to common forms of epilepsy, specifically in the PILO comparison and the MTG comparison. Interestingly, upregulated genes in human samples did not show a significant correlation with monogenic epilepsy genes but did exhibit enrichment for genes associated with common epilepsy in certain cell-types. In contrast, mouse data presented a different profile where upregulated genes were significantly enriched for monogenic epilepsy associations but not for common epilepsies. This divergence in gene regulation patterns between human and mouse models and between monogenic and common epilepsies suggests a complex interplay of genetic factors in epilepsy pathogenesis. The findings indicate that while certain genes are consistently dysregulated across different forms of epilepsy and species, the regulatory mechanisms may differ, pointing to the nuanced nature of gene expression in epilepsy.

Association of heritability analysis showed multiple converging cell-types. In the two FCD comparisons, excitatory and inhibitory neurones were significantly enriched in both up- and down- regulated gene sets. This was also seen in the PILO and the HS comparison but not in the MTG. Astrocytes which were significantly enriched in the downregulated genes in the mice comparisons, were also enriched in the downregulated genes in human HS and also the MTG, however not in human FCD. The upregulated genes in astrocytes were enriched for heritability in HS, MTG human FCD, and PILO in GGE.GSA-SNP2 results also converge

across astrocytes in all comparisons; PILO, FCD mouse, HS, MTG and FCD human. This is seen in DEE in all comparison and also in GGE for all comparisons but GGE. In the human epilepsies, there is also convergence in the upregulated genes across all 3 models, particularly in DEE and NAFE, suggesting there is also a compensatory role. These results show a potential converging role for astrocytes in epilepsy however the direction which this is acting may be more complex.

4.3.8 Differential accessibility analysis

Differential accessibility analysis compares the differences in chromatin "openness" between cell-types or conditions. When chromatin is more open, this allows for more transcription factors or other regulatory elements to bind and regulate or influence gene expression. By observing how accessible chromatin is in different cell-types or disease states, we can better understand the regulatory cellular architecture of the genome and identify potential regions of interest for therapeutic target.

	Excitatory	Inhibitory	Astro	Microglia	Oligo	OPC	Per	Endo
HS	9	8	17	2	53	4	0	0
MTG	113	14	34	0	409	13	0	0
FCD	8	7	0	0	4	7	0	0

Table 4.3 – Table to show the numbers of genes differentially accessible genes identified in each cell-type across each of the 3 models. The numbers represent the quantity of genes association with open chromatin regions that were considered to be differentially accessible between disease and control at FDR<0.05.

Table 4.3 describes the numbers of genes associated with significantly (FDR<0.05) differentially accessible regions in each cell-type in each comparison. The highest number of DA genes is seen in the MTG comparison which is not reflective of the DE. There are a large number of open regions in Excitatory neurones and Oligodendrocytes in the MTG particularly. Looking at the regions that are open across multiple cell-types can be more informative of genes regulating epilepsy. Notably, *MTRNR2L1* is the only gene consistently differentially accessible across HS, MTG, and FCD comparisons in inhibitory neurons Figure 4.12. This gene which codes for the Humanin peptide has been shown to have neuroprotective anti-apoptotic effects, suggesting a possible role in neuronal viability in the context of epilepsy(232). This gene which is downregulated in HS and FCD (FC=-1.02 and FC=-0.64) and upregulated in MTG (FC=0.31) could be playing a part in the neuronal loss associated with HS and FCDs, and playing a compensatory role in the MTG.

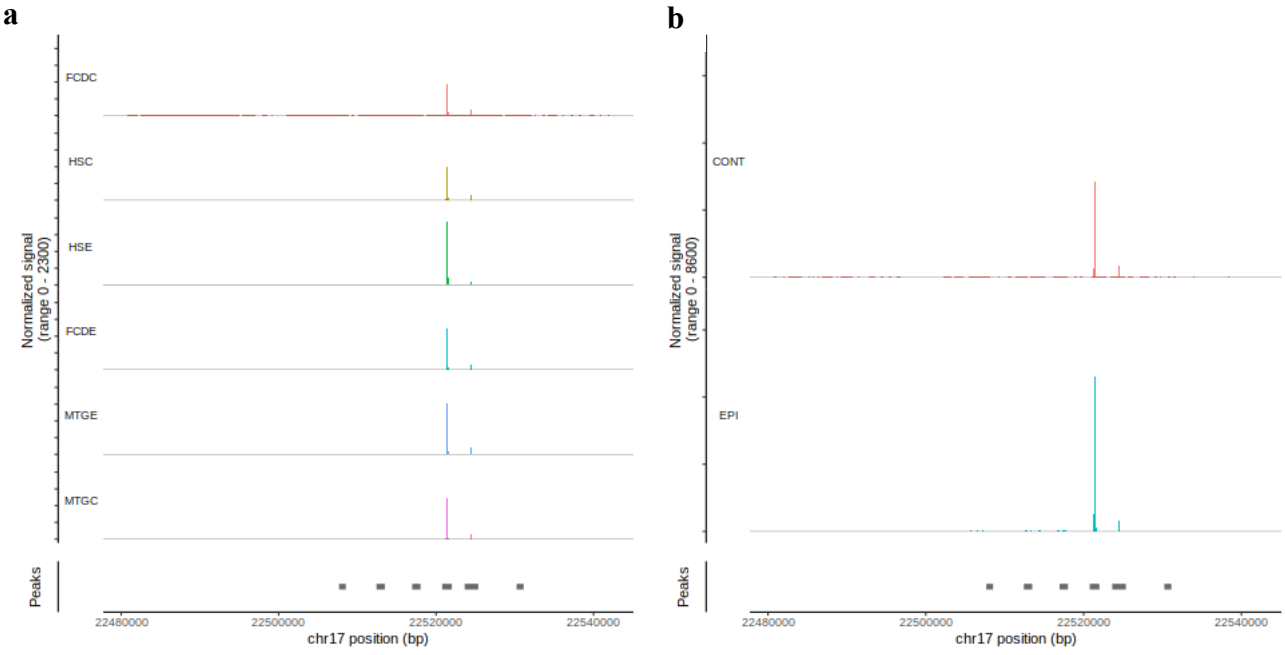


Figure 4.12 – Coverage plots to show the open chromatin of the MTRNR2L1 gene in inhibitory neurons. Less open chromatin is observed a) across all epilepsy groups b) between epilepsy and control.

This analysis also identified several genes that were commonly DA in the HS and also MTG comparisons; 2 in excitatory neurones (*PROS1, FASN*), 1 in inhibitory neurones (*PROS1*), 3 in astrocytes (*NBPF1, ZBED4, HIP1*), 1 in OPCs (*NMT2*) and 50 in Oligodendrocytes. Of these oligodendrocytes, 4 are known epilepsy genes according to genes4epielspy (*KMT2B, TSEN54, UBTF, YWHAG)* which are all involved in severe forms of childhood epilepsies. *YWHAG* and *HIP1* (Figure 4.13) have previously been investigated in connection to Williams-Beuren syndrome, an epileptic syndrome which is characterized by a deletion that includes these two genes in 95% of cases (233). *PROS1*, while differentially accessible in the neuronal cell-types, is a microglial phagocytic signalling molecule that has been shown to be altered in human epileptic tissue which may contribute to the unwanted elimination of cells or synapses (234). Notably this was seen in FCD tissue where *PROS1* is not significantly DA our this analysis.

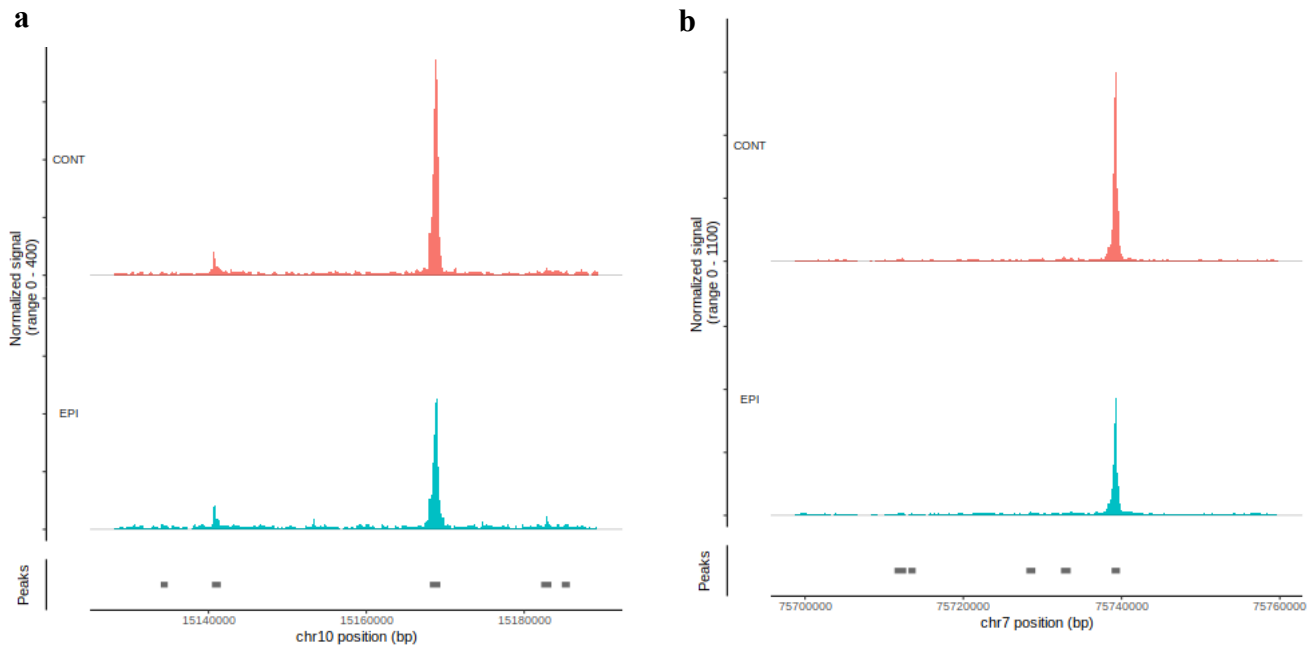


Figure 4.13 – Coverage plots to show the open chromatin of the between epilepsy and control groups in astrocyets of the genes a) YWHAG and b) HIP1.

4.3.9 Intersection of DE and DA

DA data when combined with DE data offers a more intricate insight into the regulatory of genes implicated in epilepsy. Looking at correlations between these datasets may reveal how gene expression patterns can be influenced by chromatin accessibility. Genes which exhibit both DE and DA are particularly interesting as they may be central to disease mechanisms, and may make good candidates for therapeutic targeting. Such integrative approached enhance our understanding of the cell-type specific regulatory mechanism involved in epilepsy and the multi-dimensional regulation.

Cross-analysis of the DA and DE genes was performed and genes were intersected for the corresponding comparison and corresponding cell-type. The most mutual genes were seen in Oligodendrocytes. *ZCCHC2* was DE and DA in HS, *PROS1* was DE and DA in FCD and 4 genes (*SBP2, PFKFB3, CHKA, EGR*) were DE and DA in the MTG comparison. Astrocytes in the MTG also had 3 genes that were both DE and DA (*HP1BP3, RGS12, LONRF1*). Of these genes PROS1 has been previously implicated to be altered in FCD, with a suggested role in the phagocytosis-signalling pathway (234). *CHKA* choline kinase alpha, is an enzyme that has also been shown to be involved in a neurodevelopmental disorder which causes epilepsy (235). These indicate genes are regulators of epilepsy.

Differential accessibility analysis was also applied to all cells across all samples to find genes which are differentially accessible in one cell-type that may be involved in epilepsy, but were not detected in case-control DA due to low power.

4.3.10 CRISPR BRAIN results

To further investigate the cell-type specific regulation of transcriptional shifts in epilepsy, we can look at genes whose perturbation shifts gene expression in a way that impacts the patterns of differentially expressed genes. The CRISPRbrain database (See 3.3.7) is a database of CRISPRi/a functional experiments for 300 gene perturbations. Using a causal reasoning approach developed in the Johnson Lab and detailed in Srivatava et al. (59), the enrichment of the DEGs within the perturbation datasets was tested. When DEGS are significantly enriched among the genes modulated by the gene perturbation from CRISPRbrain, this suggests that the genes can be targeted to impact genes which are DE in epilepsy. Whether these genes are upregulated or downregulated can give insights into whether these genes may be activating or repressing the disease process (Table 4.4).

DE direction	CRISPRbrain direction	Perturbation	Disease effect	Therapeutic strategy
↓	↓	CRISPRa	ACTIVATOR	Block / Inverse agonist
↑	↑	CRISPRa	ACTIVATOR	Block / Inverse agonist
↓	↑	CRISPRi	ACTIVATOR	Block / Inverse agonist
↑	↓	CRISPRi	ACTIVATOR	Block / Inverse agonist
↓	↑	CRISPRa	REPRESSOR	Agonist
↑	↓	CRISPRa	REPRESSOR	Agonist
↓	↓	CRISPRi	REPRESSOR	Agonist
↑	↑	CRISPRi	REPRESSOR	Agonist

Table 4.4 – This table conceptualises how the overlap and directionality between differentially expressed genes in epilepsy and gene perturbations can be interpreted to provide experimentally testable target hypotheses for modulating the disease. This shows how the potential therapeutic strategy can be based on the outcome of the specific perturbation.

Within the analytical framework, we assume (for the sake of the interpretation), that augmenting DE in epilepsy would have a disease activating effect, whereas mitigating DE would be disease repressing. Of course, this interpretation would depend ultimately on the relationship between genes DE in epilepsy and the disease, which could be changes that underpin the disease (and so need repressing) or homeostatic and compensatory (in which case augmenting these expression changes could be disease mitigating. Arrow pointing upwards (↑) represents upregulation, while arrow pointing downwards (↓) signifies downregulation.

The transcriptional readouts from CRISPRbrain perturbations from glutaminergic neurones and microglia were used to perform enrichment test against the perturbations in the matching cell-types. All the gene perturbations were tested and 151 genes were significant at 5% FDR in one or comparison or cell-type. The most significant enrichments came from *NQO1* which was activated using CRISPRa. The downregulated DEGs in the excitatory neurones and as well as inhibitory neurones in all three comparisons but most significantly in FCD (Figure 4.14). This suggests that an inactivation of *NQO1* results in the downregulation of genes also down-regulated in epilepsy. Based on the reasoning of the previously described framework, activating *NQO1* should potentially realign the downregulated genes in epilepsy closer to a healthy state, suggesting *NQO1* as a disease activator. This was also previously seen to be activator in mouse models of epilepsy. *NQO1* knockout mice have been shown to display spontaneous seizures due to alteration in intracellular redox status and altered glucose metabolism (191). Functional annotation of the *NQO1* perturbation showed that cell redox homeostasis was upregulated biological process as well as other metabolic process. A large number of GO terms were downregulated, the majority of which are related to synaptic function. This suggest *NQO1* may be playing a role as an activator of epilepsy throughout metabolic abnormalities leading to seizure initiation. It has been shown to act as a pro-

apoptotic factor and blocking reduces neuronal damage. However, *NQO1* is also a target enzyme of the Nrf2 pathway, which when activated protects neurones from cell death (236). is neuroprotective however and the pathway is suggested to be activated as a response to epilepsy insult (237). *NQO1* is also thought to supresses the mTOR pathway through oxygen AMPK activation. The mechanisms of *NQO1* remain unclear but patients with TLE-HS have been shown to have dysfunctional antioxidant stress response and altered *NQO1* expression due to the Nrf2 pathway (238). Functional annotation of the genes that are downregulated when *NQO1* is activated include "synaptic transmission, GABAergic" and "synaptic transmission, glutamatergic" as well as "regulation of neurotransmitter levels" and "regulation of membrane potential" and other synaptic processes.

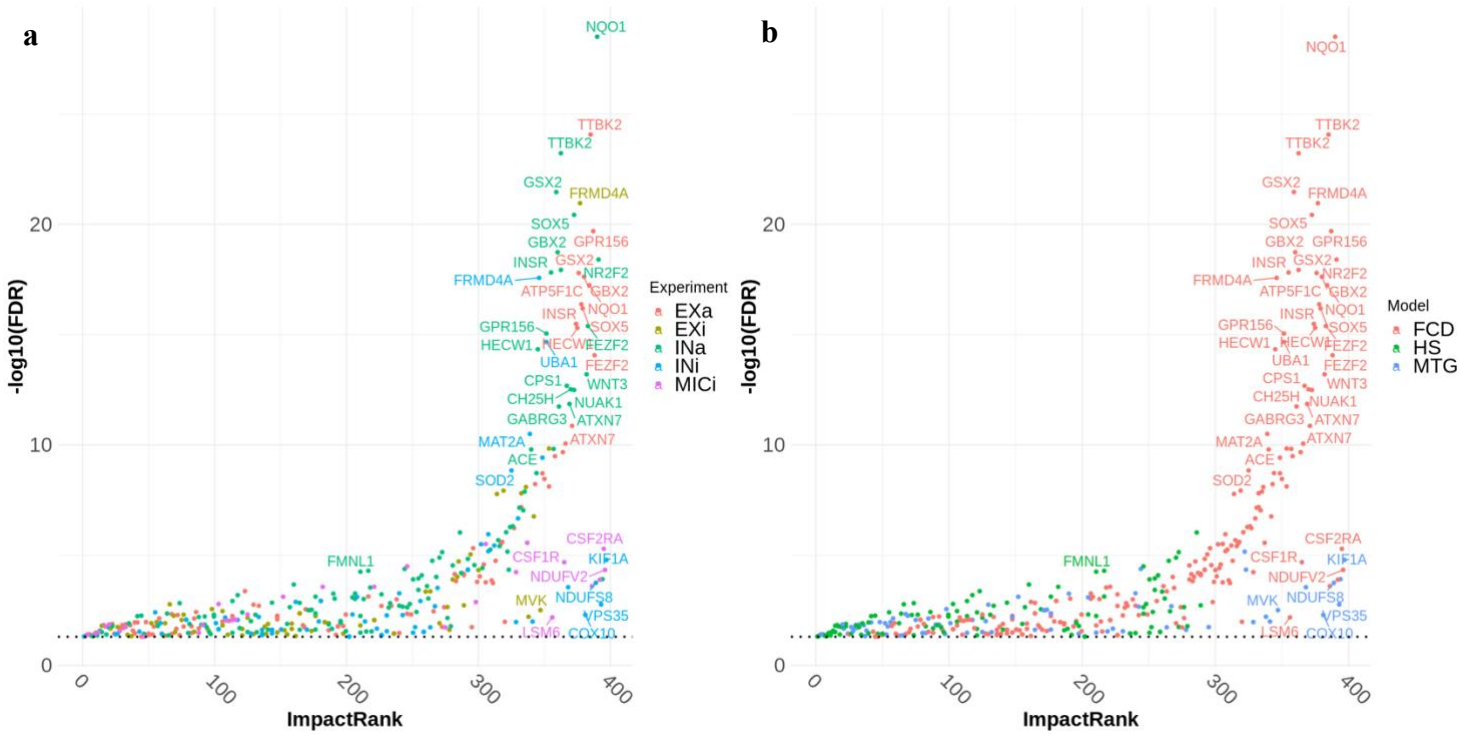


Figure 4.14 – Scatter plot to show the relationship between Impact Rank and -log10(FDR) for all significantly enriched gene perturbations, across different experimental models and conditions. a) The scatter plot is coloured by cell-type and CRISPR perturbation. Excitatory and Inhibitory neuron DEGs were tested against the glutaminergic neuron CRISPRi and CRISPRa. Microglia DEGs were tested against the microglia CRISPRi perturbations. b) The scatter plot is coloured by the mouse model the DEGs were identified in.

Other highly significant genes that have been implicated include *SOX5*, and *UBA5* which are known monogenic epilepsy gene from panels (223,239). Of the significant genes, 23 intersect with the known epilepsy gene list from Genes4Epilepsy, showing that this method has identified many epilepsy genes. *HTT* when inactivated results in the downregulation of genes

which are upregulated in epilepsy also differentially accessible in excitatory neurons and Oligodendrocytes in the MTG comparison. This gene has been shown to promote neuronal survival following seizures (240).

4.3.11 Intersections with interaction databases

Repurposing existing drugs is a strategic approach to speed up the drug development process, given the long and costly process of identifying and testing new drugs. To identify pre-existing drug candidates, significant genes were cross-referenced with a database cataloguing known gene-ligand interactions, the IU/Pharm database.

Using this database, 28 of the significant genes were found to have already established ligand interactions. With the aim of drug repurposing in mind, we focused on genes linked to already approved drugs. Out of the 28 identified genes, 9 genes had interactions with drugs that have already received approval for use in humans (Table 4.5).

Notably, one of these genes is *CSF1R* which was previously identified by the Johnson lab (18) as a potential therapeutic target in epilepsy through a framework known as the Causal Reasoning Analytical Framework for Target discovery (CRAFT), which used causal reasoning to predict cell membrane receptors that could influence gene expression in disease. CRAFT predicted that the blockage of *CSF1R* could have a therapeutic effect in attenuating seizures. CRISPRa perturbation of the gene *IGF1R* which demonstrated an overlap with the downregulated DEGs in the FCD and HS excitatory neurones. This suggests that an inactivation of *IGF1R* results in the downregulation of genes also down-regulated in epilepsy. This suggest inhibiting *IGF1R* should potentially realign the downregulated genes in epilepsy closer to a healthy state, suggesting *IGF1R* as a disease activator. This aligns with what is seen in the literature as the IGF-1 receptor has been shown to enhance hippocampal

excitatory and seizure activity through IGF-1 receptor-mediated mechanisms in the epileptic brain (241).

Gene	Number of Known Ligands	Number of Approved Ligands	Approved Ligands	Known Epilepsy Gene	PDE3 Database	Mouse
ACE	25	21	benazepril, benazeprilat, captopril, cilazapril, enalapril, enalaprilat, fosinopril, imidapril, imidaprilat, lisinopril, moexipril, perindopril, perindoprilat, quinaprilat, ramipril, ramiprilat, spirapril, spiraprilat, trandolapril, trandolaprilat, zofenopril	No	captopril, enalapril	Yes
BCL2L1	6	1	venetoclax	No	No	Yes
GABRG3	2	0		No	No	Yes
INSR	9	3	insulin, brigatinib, ceritinib	No	No	Yes
NTRK1	30	2	gilteritinib, larotrectinib	No	No	Yes
NUAK1	7	0		No	No	Yes
CASP3	8	0		No	No	Yes
SLC2A3	1	0		No	No	Yes
CCR5	37	1	maraviroc	No	No	Yes
HDAC9	14	5	belinostat, belinostat, panobinostat, romidepsin, vorinostat	No	No	Yes
NTRK2	14	0		Yes	No	Yes
MAP3K12	6	0		No	No	Yes
TAF1	1	0		No	No	Yes
BCL2	10	1	venetoclax	No	No	Yes
CTSD	5	0		Yes	No	Yes
CDK12	2	0		No	No	Yes
EPOR	3	2	erythropoietin, peginesatide	No	No	Yes
MAPK14	31	0		No	No	Yes
IGF1R	21	5	insulin-like growth factor 1, teprotumumab, teprotumumab, brigatinib, ceritinib	No	No	No
MVK	9	0		No	No	No
CYP46A1	1	0		No	No	No
ATP1A1	4	4	acetyldigitoxin, deslanoside, digitoxin, digoxin	Yes	digoxin	No
CSF1R	46	3	pazopanib, pexidartinib, G-CSF	Yes	No	No
CSF2RA	2	0		No	No	No
CSNK2A1	3	0		Yes	No	No
MAP2K6	1	0		No	No	No
INPP5D	2	0		No	No	No
TGFBR2	7	0		No	No	No

Table 4.5 – This table shows the significant genes found to have pre-existing ligand interactions, and the number of the ligands from the IU/PHAR database. It further provides details on the specific ligands which are approved for use in humans. Additional information is given regarding whether these gene targets are known epilepsy genes and if they are listed in the PDE3 database.

To further validate these targets and their interactions, these gene targets were intersected with the PDE3 database (Table 4.6). This is a curated database of all drugs that have been used and proven effective in the treatment of epilepsy in either humans or animal models of epilepsy (175). Data from the PDE3 database highlighted three drugs, Captopril and Enalapril, and Digoxin, two of which target the *ACE* (Angiotensin-converting enzyme inhibitor) gene, and the other of which targets *ATP1A1*. An increasing number of studies point to a significant contribution of Na+-K+-ATPase to regulating neuronal excitability in epilepsy (242,243). Digoxin, a drug which has previously been used to treat various heart conditions works by increasing the force of the hearts muscle contractions, and slowing the conductions of electricals signals in the heart. It works by inhibiting the Na+-K+-ATPase in heart cells which can lead to an increase in intracellular sodium which in turn promotes the intracellular calcium (244,245). It also has been shown to display the same effects on the nervous system (242). These intersections act as a way of validating the methods by showing these drugs really can have a therapeutic impact on epilepsy.

Name	Indications	First licensed (year)	Prescription items dispensed in England in 2016 (number)	Number of studies	Number of unique models	Models	AEDs enhanced
Captopril	Hypertension; chronic heart failure; diabetic nephropathy	1981	102,804	3	3	PTZ, AUD, Other	Carbamazepine, Lamotrigine, Topiramate, Valproate, Other
Enalapril	Hypertension; heart failure	1985	1,871,692	1	1	Other	Carbamazepine, Lamotrigine, Topiramate, Valproate, Other
Digoxin	Heart failure; supraventricular arrhythmia	1975	3,297,466	1	2	Electro, Other	

Table 4.6 – This table shows the intersection of the PDE3 database with the significant results from the CRISPRbrain analysis in the human brain. These are drugs that were approved for other diseases but were repurposed and found to be effective in the treatment of epilepsy.

4.3.12 Comparison to mouse

Figure 4.15 shows all the genes (198) that were significant across all 6 comparisons, and which models there were. There are several genes that were significant across 5 of these comparisons. The 33 genes highlighted in red are genes which overlapped with the genes4epilepsy list of genes known to be implicated in epilepsy. Of the significant human targets, 28 genes had known target-ligand interactions, 18 of those interactions were also seen in the mouse, and 10 were unique to the human data. Looking at genes that are significant in both comparisons might give us more confidence in the importance of these genes in epilepsy. Four genes; *EFNA5, GBX2, TFAM, EEF2* were significant in all comparisons 27 of these genes were significant in at least 5 of the comparisons albeit not always in the same direction. *EFNA5* has previously implicated in human TLE (246) and *EEF2* has been shown to regulate excitation in mice (247).

MAT2A which was the most significant gene in the mouse analysis for FCD (P= 5.04E-14) is also highly significant in human FCD (1.75E-07) in excitatory neurones. *NQO1* which was the most significant in human (P=5.74E-18) was also highly significant in mouse (P=5.01E-13)

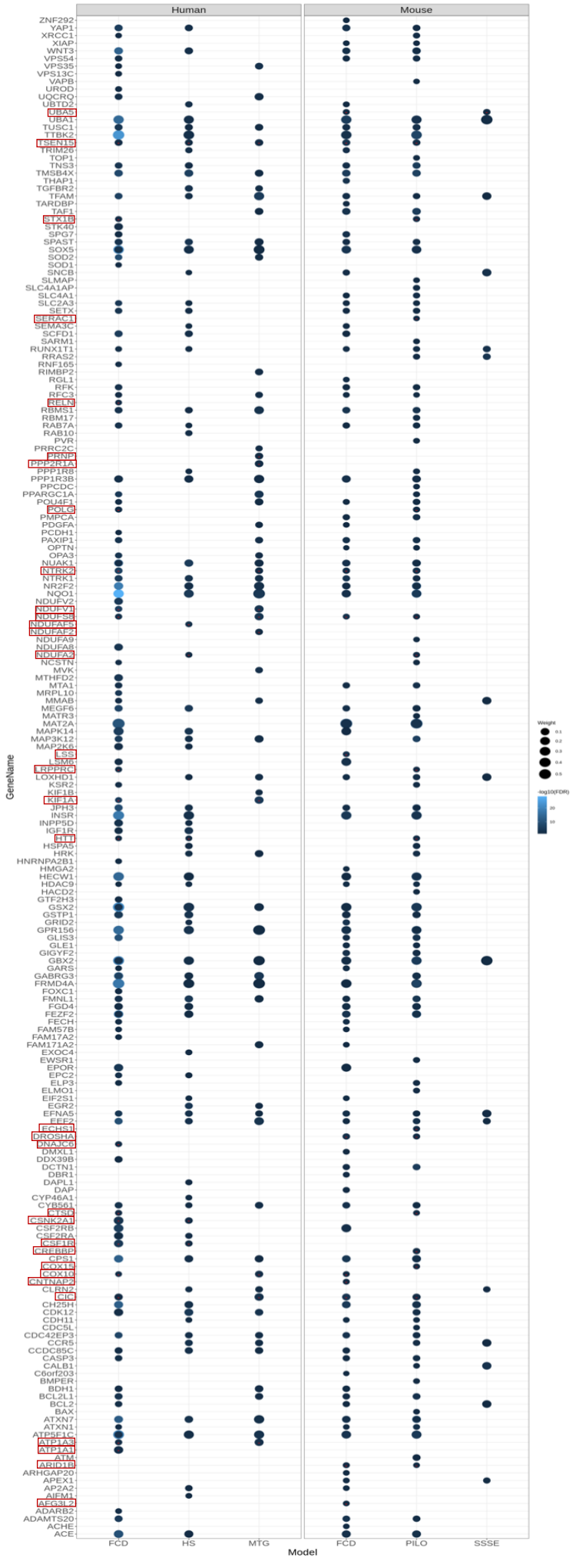


Figure 4.15 – Dot plot to show weights of significant CRISPRbrain enrichments in mouse and human data. Each row represents a significant gene, with selected genes highlighted in red to indicate their presence in the genes4epilepsy gene list. Colour intensity of dot represents -log10(FDR) of enrichment and size of dot represents weight.

4.4 Discussion

This work builds upon the previous Chapter which used mouse models of epilepsy to identify cell-types associated with epilepsy risk. By integrating human and mouse data, we aimed to pinpoint mechanisms that consistently influence disease across species. While animal models are crucial in the understanding of epilepsy, there are inherent challenges with species differences and the replication of human complexity and discrepancies in cellular composition and brain structure may affect seizure generation and treatment response and therefore translatability. However, alignment of results across three mouse models and two human epilepsy types offers compelling evidence of shared pathogenic pathways.

In this study, snRNA-sequencing analyses of human hippocampal sclerosis (HS) and focal cortical dysplasia (FCD) identified genes differentially expressed in disease and subsequent functional annotation and heritability analysis highlighted the role of astrocytes, in many aspects of disease underscoring the possibility of a unified pathogenic mechanism. Heritability analysis was used to associate disease risk with astrocytes which may show that they have a more driving role in pathogenesis particularly when associated with deleterious rare variants.

The involvement of glutamate is another converging signature across the comparisons. Glutamate involvement in epilepsy is well known in epilepsy, the glutamate receptors are involved in calcium transportation which could be a potential mechanism for hyperactivity. Previous single-cell studies have shown that glutamate is one of the strongest dysregulation in epilepsy (184). More similarities were seen between HS and FCD compared to MTG

which pointed to the direct involvement of HS and FCD regions in seizure initiation in the "ictogenic" zones, while MTG may serve a compensatory function to mitigate disease.
In the mouse, we observe both parallels and divergences with human data, indicating potential species-specific roles for glial cells, which could also be a difference in compensatory mechanism. Exploring the genes underpinning astrocytic signatures further could elucidate their functional relevance in epilepsy. Additionally diving deeper into the genes that are driving the convergent signatures in astrocytes could help better elucidate their functional relevance in epilepsy.

The CRISPR brain analysis in the human data identified additional targets not previously identified in mouse and revealed additional possible drug repurposing opportunities such as Digoxin for *ATP1A1*.
A targeted CRISPR analysis in astrocytes would be ideal given their significant role in this analysis, but current perturbation datasets are lacking astrocyte screens. Given their important role of epilepsy, drug targeting of astrocytes may be an important mechanism. CRISPRi data from cytokine-induced inflammatory astrocyte screens are available which may provide valuable insights into the potential mechanism underlying the roles of astrocytes and their roles in maintaining brain homeostasis and their involvement in epilepsy. These functional screens include screens to identify transcriptional regulators of lysosome remodelling induced by IL-1α+TNF+C1q, cell-surface VCAM1 (a marker of inflammatory activation), screen for cell-surface LAMP1 (marker of lysosome exocytosis), and screen for phagocytosis.
However, CRISPR databases and gene analysis provides a robust strategy for identifying potential treatment targets. Experimental validation remains essential, especially given the variability in gene expression directionality across models and species. Targeting key

regulators demands a nuanced approach to ensure treatment efficacy without unintended effects.

Combined with gene expression data, chromatin accessibility data can be used to quantitively associate open chromatin regions to specific gene expression to identify functional relationships between cis-regulatory elements disease differential expression reveal insights into the roles of enhancers and promoters in disease. Chromatin accessibility is correlated to promoters, enhancers and other regulatory regions which provides evidence for regulatory gene interaction.

The ATAC-seq data in the study was not seen to be very concordant with the gene expression data and there are not many genes found to be DE and DA in the same cell type. However it is known that gene expression doesn't have high overlap with chromatin accessibility.

The addition of genotyping data to this analysis may give more information about the regulation as we can do motif analysis and TF binding.

A limitation of this analysis is that we associate open chromatin regions to genes by looking at which genes they are closest to. This is not always the case in gene regulation and regulation elements may be regulating genes further away – particularly in the case of enhancers which can be very distal.

4.4.1 Limitations and Future Work

When performing transcriptomic analysis on diverse groups, it is important to carefully consider covariates in the model. These covariates can introduce systematic biases that may confound gene expression patterns related to epilepsy, if not appropriately addressed. Including covariates in the statistical models, adjusting for batch effects may help to mitigate these biases by controlling for the potential confounders and assessing the true associations

between gene expression levels and experimental conditions or clinical variables (125). In this analysis, we included age and sex as covariates. However, incorporating additional variables such as post-mortem intervals (PMI) and seizure frequency could further enhance the accuracy and reliability of our findings.

Considering functional covariates such as seizure frequency or medication status in the study of epilepsy could helps distinguish disease-specific molecular signatures from treatment effects or seizure consequences. Post-mortem interval (PMI) refers to the time elapsed between death and the collection of tissue samples for analysis. PMI can significantly affect tissue integrity and quality as longer intervals can lead to degradation of RNA which may influence the stability and accuracy of gene expression (248). Additional tissue data was not available in this analysis but should be included in future work to assess other covariates that could impact the data.

Additionally, comparing surgical specimens from epilepsy patients with control post-mortem samples presents several challenges, this lies in the inherent differences between surgical and post-mortem samples. Surgical specimens are typically obtained from living patients undergoing treatment, introducing variability related to treatment effects, disease progression, and medication history. Post-mortem samples lack the physiological processes and treatments affecting gene expression seen in surgical samples, but they are subject to issues like RNA degradation which can influence RNA integrity and gene expression profiles. This makes it challenging to infer accurate conclusions from the differential expression analysis when all the epilepsy samples are surgical and all the control samples are control, we are unable to adjust for these differences. Therefore conclusions from these comparisons should be considered with caution as common genes and pathways that are found to be DE could also be reflecting differences in the tissue collection.

Challenges remain in standardizing sample collection protocols and accounting for additional variables that may influence gene expression, such as brain region specificity and cell-type heterogeneity. This remains a challenge in epilepsy research, as control brain samples must always be post-mortem. Furthermore, we cannot use post-mortem epilepsy samples as patients often die from comorbid conditions, which can confound the analysis of epilepsy-specific factors.

While the ATAC-seq data did not yield substantial information in the context of this thesis, it can be a useful tool when more carefully applied. The rationale for using ATAC-seq lies in its ability to identify open chromatin regions associated with active regulatory elements, such as promoters, enhancers, and transcription factor binding sites. As epilepsy is such a complex and heterogenous disease, it is likely that there is complex and intricate regulation contributing the complexity. This chromatin accessibility data, when combined with gene expression profiles, allows for a deeper understanding of how changes in chromatin structure influence gene activity (249). One of the key benefits of ATAC-seq, particularly relevant for cell-type clustering, is its high sensitivity and resolution, which enables the discrimination of distinct cell-types based on their unique chromatin profiles. This is particularly valuable in tissues where traditional methods of cell-type identification might be challenging due to overlapping markers or subtle differences between cells. It has been shown that in complex tissue, chromatin accessibility can better define cell-types and lineages (250).
However, ATAC-seq has its limitations, particularly when the quality of the data is low, such as in this context. There were significant limitations and technical challenges associated with optimizing protocols which meant the data was very noisy. ATAC-seq data is extremely sparse and working with this data can be computationally demanding. The peak-calling process, a critical step in the analysis, can be highly sensitive and affect the final results

depending on the methods used. Despite these challenges, the ATAC-seq data will be key for future work focused on classifying cell-types at a more refined, subtype-specific level.

Chapter 5

Identification of Dysplastic Neurons

5.1 Introduction

Focal cortical dysplasia (FCD) is a severe disease caused by malformation of cortical development in the cerebral cortex. It is the most common form of refractory epilepsy in children and the second most common form in adults (82). There are several types of FCD with a wide range of characteristics and symptoms, the most notable being the presence of abnormal neurons called dysplastic or dysmorphic neurones (251). These are usually are present in an isolated lesion which is the focus of seizures (252). Seizures are the most common symptom of FCD followed by cognitive disability and language delays or visual impairments, depending on the area of the brain affected. While FCD is highly associated with medically intractable epilepsy, early diagnosis, and treatment for FCD is important and can prevent further development of symptoms and improve outcomes. For example, in the case of cognitive disability, proper care and therapy can significantly impact patient quality of life (199). There are both genetic and acquired factors involved in the development of FCD, however the exact cause is not well understood. It is believed to result from a disruption in the normal migration of neurones during early brain development (48). The mechanisms of how this happens is not well understood and more research is needed to develop more effective treatments and better diagnosis.

5.1.1 FCD pathology

FCD involves a range of brain alterations including cellular and molecular changes of anatomical features. These involve neuronal alternation such as hypertrophy, gliosis, synaptic organisation, inflammation, and structural changes or cell immaturity. The abnormal proliferation and organisation of cells in the brain are usually centred around a lesion where the abnormal brain tissue is present in patients with FCD (207). These lesions are the epileptic zone where seizures originate. The International League Against Epilepsy (ILAE) classification scheme is used for characterizing FCDs (Figure 5.1). It can be classified into three types and nine subtypes based on severity and characteristics (253). These are:

- FCD Type I involves cortical abnormalities which may affect any layer of the cortex. These are usually more subtle and may not always be detectable by MRI.
- FCD Type II is more severe and involves the formation of dysplastic neurones and sometimes balloon cells. The pronounced cortical abnormalities mean it can more easily be detected by MRI. FCD IIa involves only dysplastic neurones whereas FCD IIb involves both dysplastic neurones and balloon cells.
- FCD III and its subtypes were more recently introduced as including the presence of other lesions such as hippocampal sclerosis or tumours (254).

Table 1. The three-tiered ILAE classification system of focal cortical dysplasia (FCD) distinguishes isolated forms (FCD Types I and II) from those associated with another principal lesion (FCD Type III).

FCD Type I (isolated)	Focal cortical dysplasia with abnormal radial cortical lamination (FCD Type Ia)	Focal cortical dysplasia with abnormal tangential cortical lamination (FCD Type Ib)	Focal cortical dysplasia with abnormal radial and tangential cortical lamination (FCD Type Ic)	
FCD Type II (isolated)	Focal cortical dysplasia with dysmorphic neurons (FCD Type IIa)		Focal cortical dysplasia with dysmorphic neurons and balloon cells (FCD Type IIb)	
FCD Type III (associated with principal lesion)	Cortical lamination abnormalities in the temporal lobe associated with hippocampal sclerosis (FCD Type IIIa)	Cortical lamination abnormalities adjacent to a glial or glioneuronal tumor (FCD Type IIIb)	Cortical lamination abnormalities adjacent to vascular malformation (FCD Type IIIc)	Cortical lamination abnormalities adjacent to any other lesion acquired during early life, e.g., trauma, ischemic injury, encephalitis (FCD Type IIId)

FCD Type III (not otherwise specified, NOS): if clinically/radiologically suspected principal lesion is not available for microscopic inspection.
Please note that the rare association between FCD Types IIa and IIb with hippocampal sclerosis, tumors, or vascular malformations should not be classified as FCD Type III variant.

Figure 5.1 – ILAE classifications of FCD types and subtypes. FCD can be classified into 3 types with 9 subtypes based on severity and pathology (Image Credit: ILAE)(253).

FCD diagnosis involves a combination of neuroimaging, histopathological examination through surgical biopsy and clinical symptom evaluation, however it can be difficult to diagnose as it is very variable in appearance on MRI depending on the type. Radiological characteristics include blurring of grey-white matter due to the presence of abnormal cells that extend beyond normal boundaries, cortical thickening due to disorganised and abnormal cells and increased signal intensity in the white matter in FCD lesions in some types of imaging. The cortical surface may appear irregular die to a lack of normal folding however lesions may be subtle and not easily visible on MRI machines (207).

FCD is one cause of intractable epilepsy, and a significant number (50-80%) of patients continue to have seizures even after treatment. Patients often undergo surgery to remove the lesions, however many patients continue to have seizures even after resection which suggests that other parts of the brain can be affected. As with other epilepsies, the available AEDs usually target the symptoms rather than the disease (255). Other than epilepsy medication and surgical resection, there are limited treatment options available.

5.1.2 Dysplastic Neurones

Dysplastic or dysmorphic neurones (DNs) are abnormal neuronal cells that have gone through abnormal proliferation or development and growth. Identification of DNs (and balloons cells (BCs) in FCD Type IIb) usually involves neuropathological observation through immunolabelling. There are a number of markers that are used clinically to identify

DNs however these are not specific to only dysplastic neurones and are usually a combination of developmental markers (Otx1, Pax6, DCX, DCX-like protein, Nestin, MASH1, SOX2, Sox3, Oct-4, CRMP4, c-myc, FoxG1, KLF4), mTOR pathway activation markers (such as phosphorylated ribosomal S6) and channel/synaptic markers (Synaptophysin, VGlut, NKCC1, AQP4, Cx43, NR2B, EEAT3) (256). DN neurones are found to be more intensely labelled with proteins such as Neurofilament, NeuN, Tbr1, Map2, and Map1b however normal neurones can also be positive, so these are not entirely specific. Molecularly, gene expression signatures of FCD have not been reported (83). However it has been shown that failure of neuronal differentiation and migration may be as a result of abnormalities in transcriptional regulators of gene expression (257,258).

5.1.3 mTOR pathway

The mammalian target of rapamycin (mTOR) pathway is a central regulator of cellular metabolism, growth, proliferation, and survival and has been implicated in many neurological disorders including focal cortical dysplasia (FCD). Mutations in the mTOR pathway have been seen as the primary cause of genetic FCD and aberrant mTOR signalling is linked to drug-resistant epilepsy (259). FCD was first linked to mTOR through its association with Tuberous Sclerosis Complex (TSC) or tuberous sclerosis. TSC is a genetic syndrome characterised by brain lesions known as “cortical tubers” as well as benign tumours in other organs. TSC has been attributed to mutations in genes TSC1 and TSC2 which are tumour suppressor genes involved in the mTOR signalling pathway. Similarities between TSC and FCD have been described, particularly in the histopathology and disruption of cortical layers (260).

Mutations in mTOR can be familial or sporadic (261) and often involve the PI3/AKT cascade (262). Dysmorphic neurons and balloon cells have been shown to have abnormal mTOR

activation and mutations in the mTOR pathway that lead to its overactivation have been shown to subsequently lead to FCD pathology. This can be due to mutations in the mTOR gene itself or in other genes in the pathway, such as TSC1,TSC2, DEPDC5, NPRL2 and NPRL3 (252). The mTOR pathway is shown in Figure 5.2.

Medications that supress the mTOR pathway such as rapamycin have demonstrated potential in decreasing seizures and other clinical outcomes (263). Further study of the mTOR pathway could lead into improved diagnostic precision and better understanding and identification of variants in mTOR pathway genes.

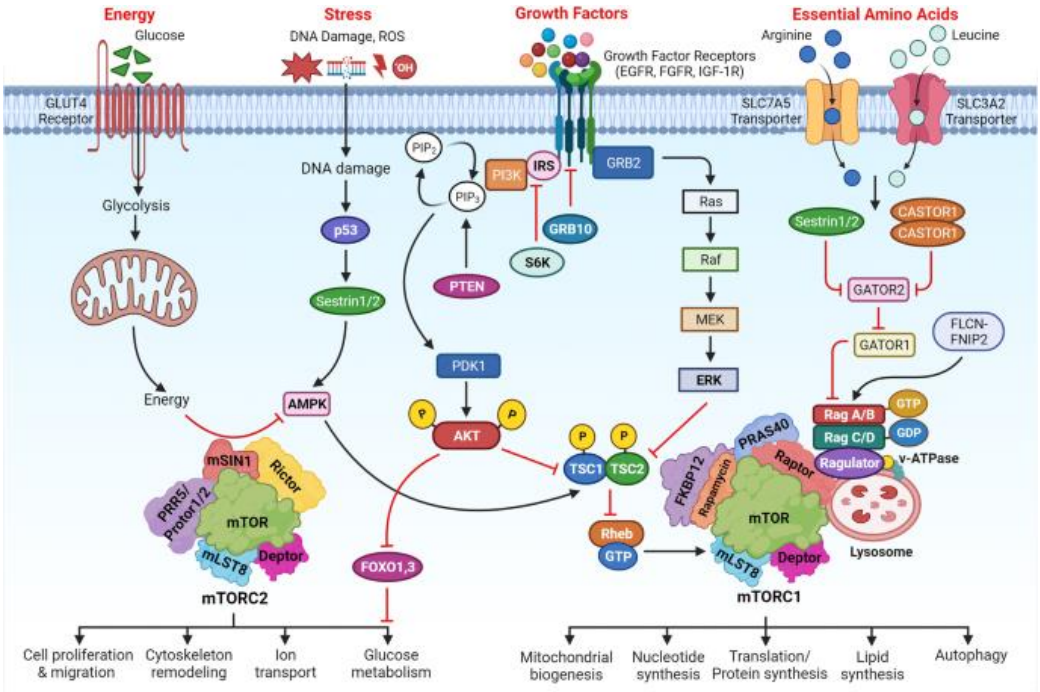


Figure 5.2 – Illustration of the mTOR pathway to demonstrate the many genes and proteins involved. The mTOR pathway plays a role in many important cellular processes such as cell stress and growth (Image credit: Kumari et al. 2020)(264)

5.1.4 Aims

Alongside the development of more effective treatments for interactable epilepsy, there is also a need for better diagnosis of FCD and its subtypes. Visualisation of FCD lesions and abnormalities can be challenging, which makes diagnosing FCD difficult. In this chapter, we aimed to transcriptionally characterise dysplastic neurones by analysing both mouse and human FCD samples. Our goal was to discover biomarkers of dysplastic neurones by identifying more specific markers that could be used for better identification of FCD in clinical settings. Additionally, a better characterisation of dysplastic neurones could potentially provide insights into the mechanisms leading to the abnormalities and therefore the underlying mechanism of FCD. As the mouse model of FCD in this study is genetically induced, we are able to precisely identify which cells in the data are dysplastic. These cells could then be characterised and ideally used to identify dysplastic neurones in the human tissue.

5.2 Methods

5.2.1 Mouse model of FCD

The experimental mouse model of FCD was generated by UCB Pharma using *in utero* electroporation as previously described (265,266). Briefly, embryos of pregnant WT mice are removed at E15 and injected with the DNA of interest. These are plasmids with florescent enhancer constructs, which are then injected into the ventricles of the embryos to target neuronal progenitors. Transfection of sequences is performed using forceps-type electrodes, and embryos are then implanted back in into the mice. Histopathology is then used to observe the expression of the mutated protein in the mouse brain. This showed that the mice recapitulated human pathophysiology of FCD including focal cortical disorganisation, neural migration, cortical dyslamination and dysplastic neurons. Spontaneous and recurrent seizures were also observed using video-EEG.

The plasmids sequences implanted encode a constitutively active form of the *Rheb* gene ($Rheb^{CA}$) which canonically activates the mTOR signalling pathway and increases the activity of mTORC1. mTOR activity is increased in the layer 2/3 neurons of the medial prefrontal cortex which causes tonic-clonic seizures. Samples are then collected when the mice are 3 months old. The electroporation targeted a subset of neural progenitors at day E15 in the cortex which allows preferential targeting of neural progenitor cells that will generate layer 2/3 pyramidal neurones in the Pre-Frontal Cortex (PFC). As the transfection was limited to the neural progenitor stage, only neuronal cells should harbour the introduced plasmid and subsequently become dysplastic. This is therefore a model of FCD Type IIa as we should not expect to see the transfection of any non-neuronal cells and thus the presence of balloon cells. Histopathology was then used to confirm the presence of dysplastic neurones in these mice. These are identified by observing phosphorylated S6 protein levels which is a downstream effector of mTOR.

The mouse samples used in this study are summarised in Table 3. These include; FCD samples from the lesion transfected with the Rheb plasmid (n=15), FCD samples from the non-transfected side of the cortex (n=7) (contralateral), Sham samples which were transfected with a control plasmid (n=7), and control pre-frontal cortex samples not transfected with plasmids (n=8). This data was generated in two batches. Further information on the processing of these samples is described in Chapter 3.

Model	**FCD (Batch 1)**			**FCD (Batch 2)**	
Sample Type	Transfected FCD	Non-transfected FCD	Transfected Sham	Transfected FCD	Non-transfected Control
Brain Region	Prefrontal Cortex	Prefrontal Cortex (contralateral)	Prefrontal Cortex	Prefrontal Cortex	Prefrontal Cortex
Sample Number	7	7	7	8	8

Table 5.1 – Summary of mouse FCD samples used in this analysis. Data was generated in two batches from 37 samples in total.

5.2.2 BBmap

To identify cells that were transfected with the *Rheb* or Control plasmids, the sequences of the plasmids were aligned against the single-nucleus sequencing data. The tool BBmap was used to map the brain tissue snRNA-sequencing reads sequences to the plasmid sequences (Figure 5.3) (267). This is a splice-aware aligning software that can accurately and quickly map reads with high precision. The command BBsplit was used to map snRNA-sequencing reads to multiple plasmid sequences. This takes an input of plasmid sequences in a FASTA file and BAM files from the FCD tissue. BBsplit unambiguously maps reads that only belong

to one the plasmid sequences and quantifies how many reads are in each. Reads were mapped to the following parts of the plasmid sequences.

FCD Plasmids:

- SV40polyA (192bp)
- GFP reporter (720bp)
- Rheb(S16H) (with mutation) (181bp)

Control Plasmids

- SV40polyA (192bp)
- RFP reporter (714bp)

As these sequences are not expected to be found naturally occurring in the genome, if there is a match with the data, then this means that cell has been transfected with a plasmid. We confirmed that there was no other matches to sequences in the genome by using BLAST (Basic Local Alignment Search Tool), a tool used for searching and matching sequences(268). As Rheb is natively expressed in the genome, only sequences that would overlap with the mutation in the Rheb sequence were considered. The base change in the mutated Rheb is highlighted below:

cggtctgt → cggcatgt

As the 10x single-nucleus sequence fragments are only 92bp long, the Rheb sequence was taken with 90bp either side of the mutation. Only 100% matched reads were considered. Once sequences were mapped to plasmids, the reads were attributed to barcodes in the final filtered data (Chapter 3) and quantified per cell.

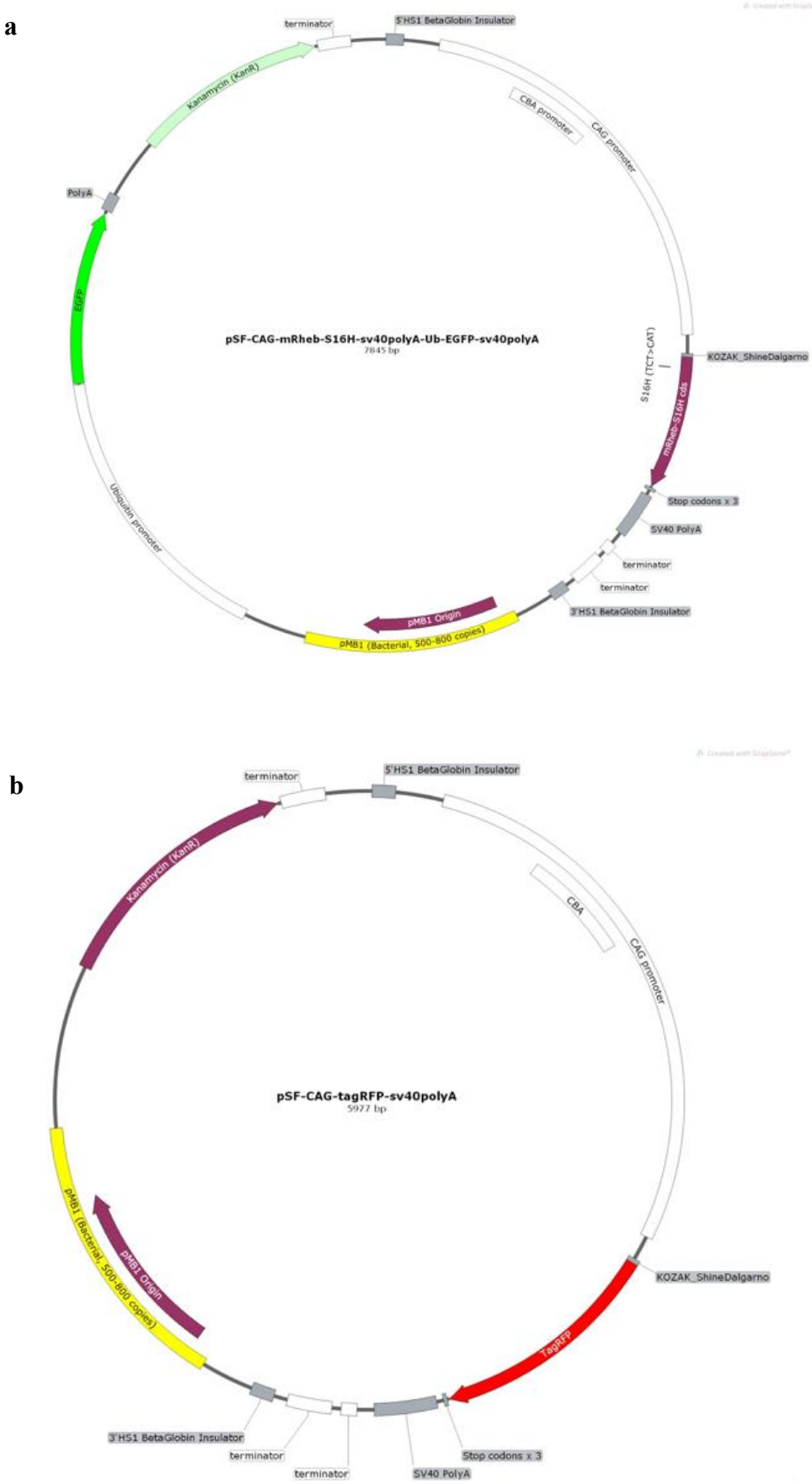


Figure 5.3 – Transfected plasmids used in this study. a) Control plasmid implanted into control mice b) Rheb plasmid implanted into FCD mice which causes mTOR activation.

5.2.3 Human FCD samples

The human FCD samples used in this analysis were provide by the UCL Brain Bank. These are a mixture of FCD Type IIa and Type IIb from patients who have undergone epilepsy surgery. Post-mortem pre-frontal cortex samples from the Imperial brain bank were used as controls. Further information on the processing of these samples is described in Chapter 4.

Group	**Focal Cortical Dysplasia (FCD)**	
Sample Type	Focal Cortical Dysplasia	Control
Brain Region	Prefrontal Cortex	Prefrontal Cortex
Sample Number	6	10

Table 5.2 – Summary of human samples used in this analysis.

5.2.4 Downstream analysis

The pre-processing and quality control steps of the samples used in this analysis was described in Chapter 2 and Chapter 3.2. The downstream analysis including differential expression and functional annotation was also performed as described in Chapters 3 & 4.

5.3 Results

5.3.1 Clustering of dysplastic neurons

To characterise dysplastic neurones (DNs) within the dataset, and attempt to discern transcriptional differences from normal neurons, excitatory and inhibitory neurones in the data were extracted and clustered separately to all other cell populations. Given that the mouse model we use is recapitulating FCD Type IIa, we did not consider balloon cells in this analysis as they are not expected to be present in this context. This was performed on both the human and the mouse datasets Figure 5.4.

This revealed that dysplastic neurones did not cluster separately to normal neurones based on their transcriptomic signatures alone. This showed that more variation within the dataset could be attributed to neuronal subtypes rather than disease state.

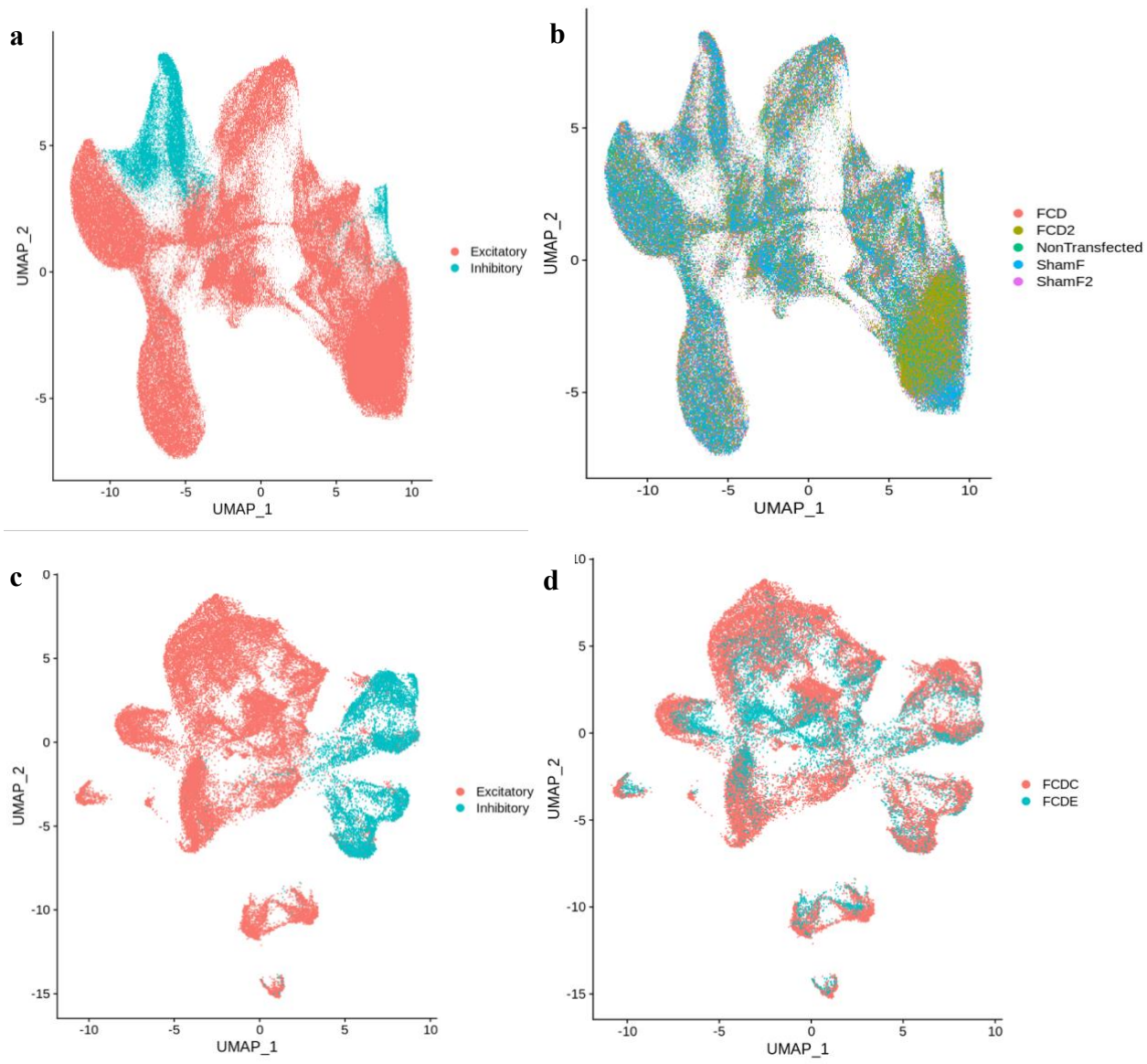
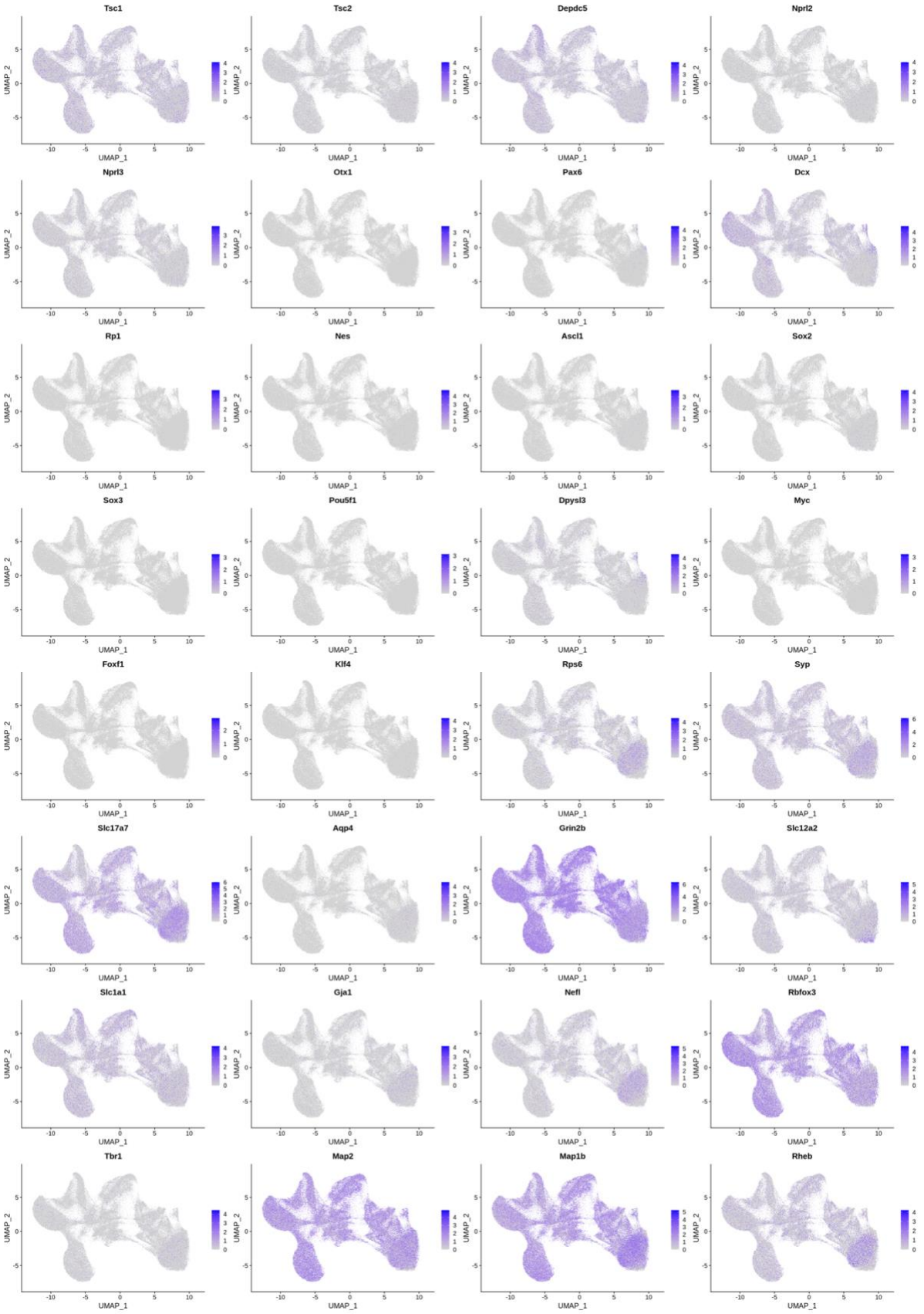
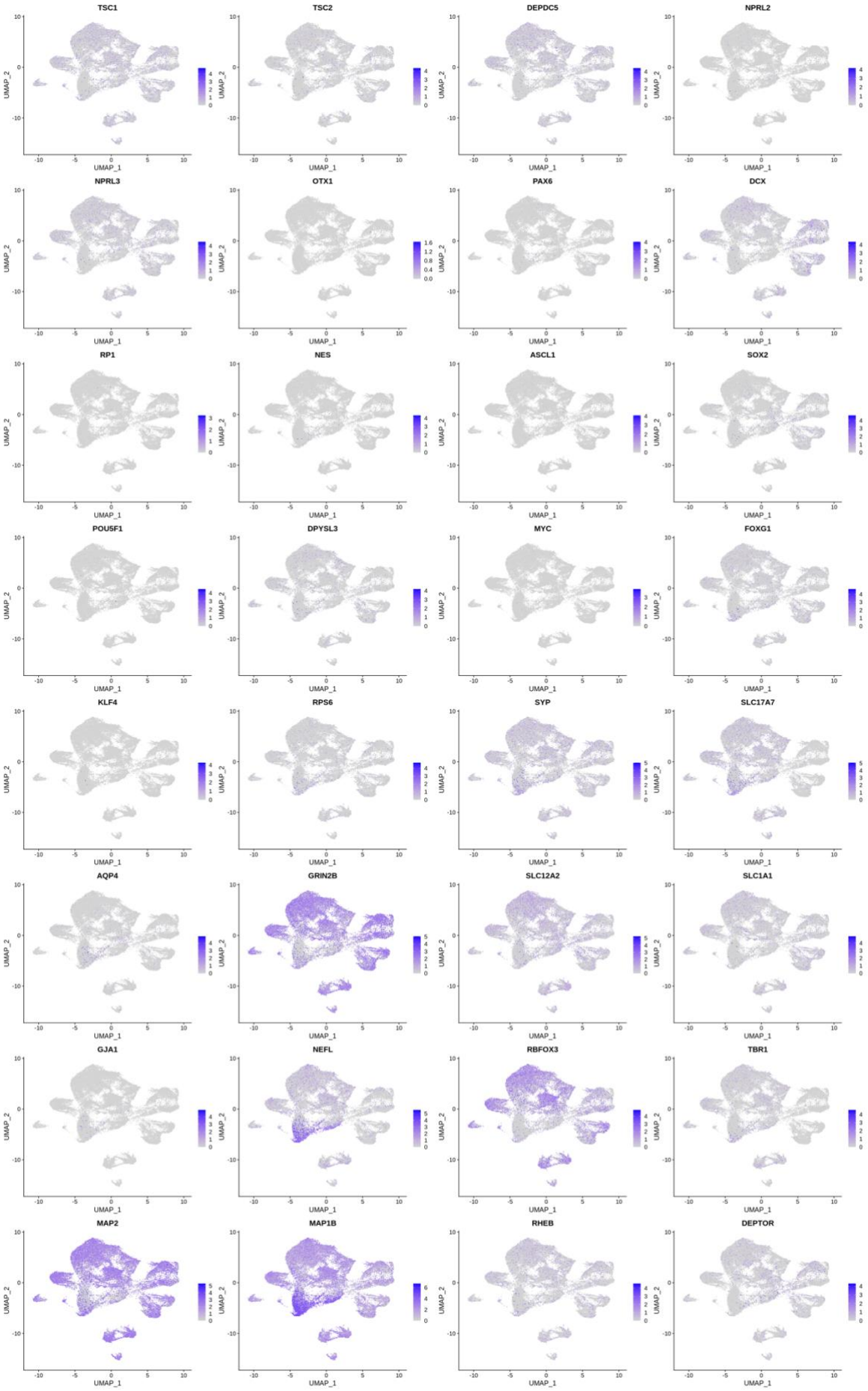


Figure 5.4 – Clustering of pre-frontal cortex excitatory and inhibitory neurones a) Mouse neurones coloured by type a) Mouse neurones coloured by experimental group; FCD and FCD2 are FCD mice form batch 1 and batch 2 respectively, non-transfected are samples from the contralateral non-transfected brain, ShamF are control mice transfected with control plasmids and Cont are non-transfected control mice c) Mouse neurones coloured by experimental group d) Human neurones coloured by disease state; FCDE are patients with FCD and FCDC are control patients.

Next, only neurones from FCD samples were clustered and known protein markers of DNs were mapped on to the data to see whether DNs expressed these markers transcriptionally in different patterns to non-dysplastic neurones (Figure 5.5 & 5.6). We used markers of dysplastic neurones that are used clinically in neuropathological annotation. These markers are not specific to DNs and are usually used in combination in a clinical setting as they are also expressed in normal tissue. This again could not identify a distinct signature that would help distinguish DNs, and most of these markers were either expressed across all the neurones or were not expressed highly at all.


Tsc1
Tsc2
Depdc5
Nprl2
Nprl3
Otx1
Pax6
Dcx
Rp1
Nes
Ascl1
Sox2
Sox3
Pou5f1
Dpysl3
Myc
Foxf1
Klf4
Rps6
Syp
Slc17a7
Aqp4
Grin2b
Slc12a2
Slc1a1
Gja1
Nefl
Rbfox3
Tbr1
Map2
Map1b
Rheb
UMAP_1
UMAP_2


TSC1
TSC2
DEPDC5
NPRL2
NPRL3
OTX1
PAX6
DCX
RP1
NES
ASCL1
SOX2
POU5F1
DPYSL3
MYC
FOXG1
KLF4
RPS6
SYP
SLC17A7
AQP4
GRIN2B
SLC12A2
SLC1A1
GJA1
NEFL
RBFOX3
TBR1
MAP2
MAP1B
RHEB
DEPTOR
UMAP_1
UMAP_2

As we could not identify DNs through marker expression alone, we next sought to identify dysplastic neurones through the genetic sequencing data. As the mouse FCD samples are generated through genetic manipulation by transfection of plasmid sequences, we should be able to identify which cells are dysplastic by identifying which cells contain plasmid sequences. Considering Rheb is endogenously expressed, we cannot use its expression alone to identify transfected cells. Therefore, only sequences overlapping with the introduced mutation were considered. To observe that mutated *Rheb* was only present in the transfected FCD samples, we used IGV viewer (269) to confirm the presence of the mutation. Figure 5.7 shows that this was only present in the FCD samples. The other sample have only native Rhebs and should therefore not express the plasmid.

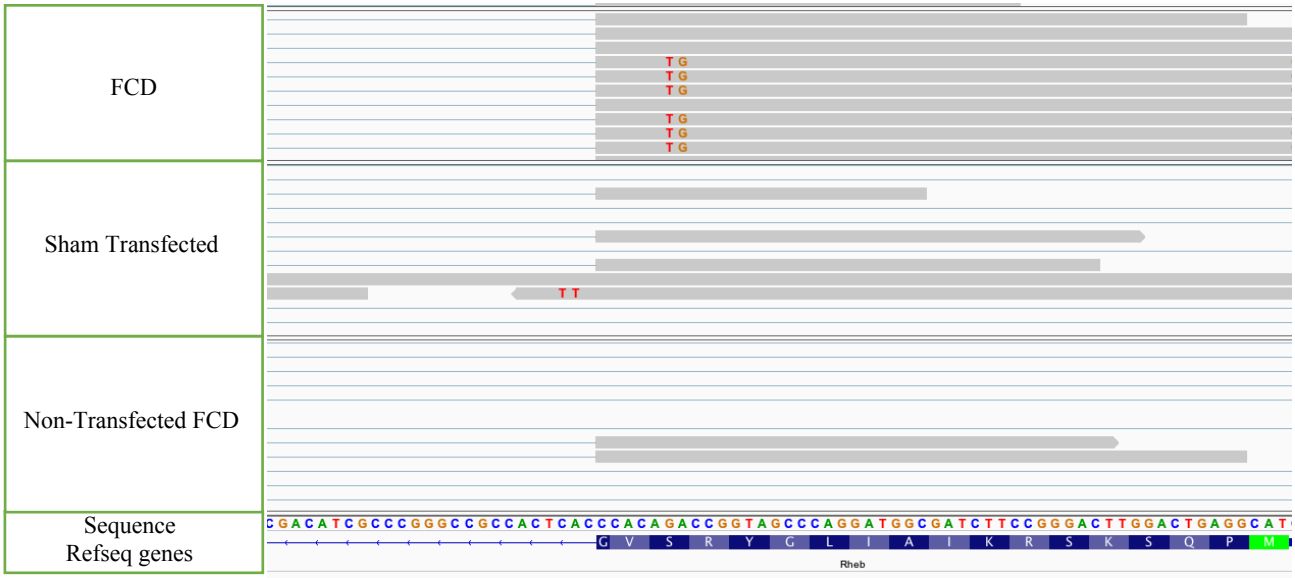


Figure 5.7 – IGV viewer showing FCD samples transfected with mutated Rheb.

BBmap (267) was then used to map the sequencing reads onto the plasmids considering only exact matches. Cells with sequences that matched with the plasmid sequences were identified and counted to identify how many dysplastic neurons there were per sample. We assume that

cells containing the plasmids are DNs as they will be overexpressing Rheb. The numbers of transfected neurones are shown in Table 5.3.

	FCD (Batch 1)	FCD (Batch 2)	Transfected Control (Batch 1)	Non-Transfected Control (Batch 2)
Non-Transfected Neurons	42,907	50,165	34,064	48,510
Transfected Neurons	2,709	4,539	945	251
% Dysplastic	5.94	8.30	2.70	0.51

Table 5.3 – Numbers of transfected neurones identified in the mouse data.

The percentage of neurones that are transfected are between 2.7-8.3% in this data. Experimentally, ~10% percent of neurones are found to be transfected (160). It is important to note however that if a plasmid sequence has not been detected in a cell, this doesn't mean the cell is not dysplastic as single-cell data is sparse with a high dropout rate. This mapping also does not account for sequences that may not match 100% due to point mutation or sequencing errors. We also see a small number of neurones in the non-transfected control group that contain plasmid sequences. This may indicate that ambient RNA has not fully been accounted for in the data and that some plasmid sequences have contaminated other cells. We also see a small number of glial cells which contain plasmids. It is unclear whether this is due to ambient RNA which has not fully been accounted for by correction, or whether this is due to mislabelling of neurons that have clustered with other cell-types.

Next, we visualised which cells of the FCD neurones contained plasmid sequences. Figure 7 shows where the plasmid sequences are mapped in the data. While it appears, there is a higher concentration of dysplastic neurones in cluster 0, all other clusters also express plasmid sequences and the percentages per cluster are comparable. This did not reveal any clusters with transcriptionally distinct signatures.

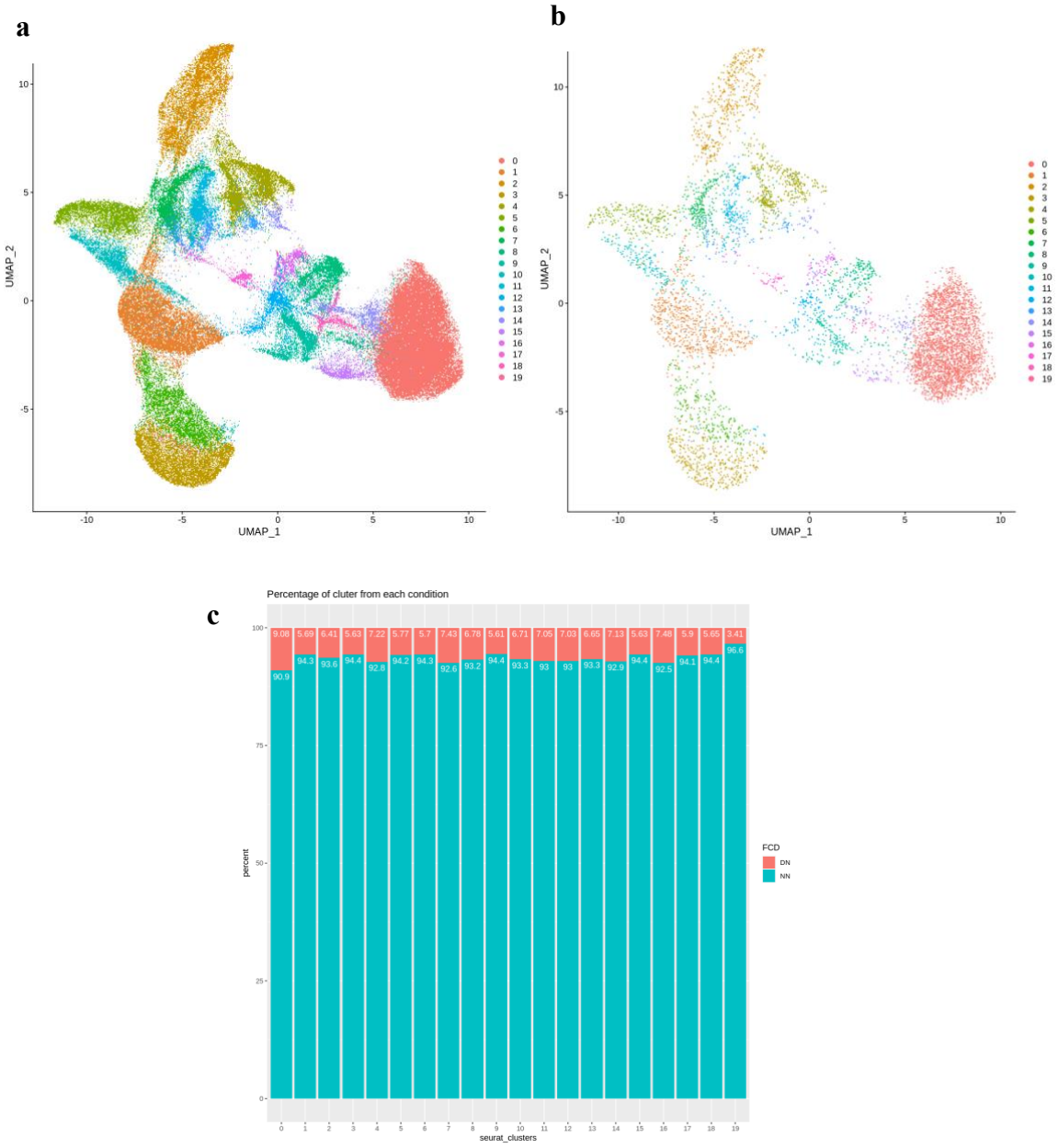


Figure 5.8 – Distribution of dysplastic neurones within data. a) Clustered FCD neurones b) Distribution of Dysplastic neurones within clusters c) Percentage of dysplastic neurons (DN) per cluster (NN- normal neurons).

5.3.2 Differential Expression of dysplastic neurons

Next, to attempt to identify markers of dysplastic neurons. Differential expression analysis (DEA) between dysplastic and normal neurons in the FCD samples was performed within Seurat using the FindMarkers() function which uses a Wilcoxon test to perform DEA. There were 89 genes found to be differentially expressed between Dysplastic neurones and Non-dysplastic FCD neurons. Two of these (*Rheb, Nefl*) are known markers of DNs mentioned above. The top 15 most DE genes are shown in Table 5.4 below.

	P Value	Log2(FC)	% Expressed (DNs)	% Expressed (NNs)	FDR
Rheb	7.83e-275	0.6589685	0.582	0.398	2.35e-270
Npy	3.00e-144	0.4908736	0.658	0.527	9.03e-140
Cck	9.41e-92	0.3519914	0.766	0.688	2.82e-87
Scg2	2.25e-89	0.3867652	0.584	0.474	6.77e-85
Nefm	6.14e-89	0.3634933	0.395	0.280	1.84e-84
Mif	3.01e-84	0.3105374	0.513	0.393	9.06e-80
Cplx1	3.08e-83	0.3129284	0.446	0.334	9.27e-79
Dbi	3.28e-83	0.3056780	0.401	0.289	9.88e-79
Ubc	1.15e-78	0.3051860	0.431	0.320	3.47e-74
Atp6v1e1	3.78e-73	0.2835312	0.554	0.441	1.13e-68
Uchl1	5.97e-73	0.3140628	0.622	0.531	1.79e-68
Efhd2	4.82e-71	0.2872668	0.449	0.346	1.44e-66
Aldoa	1.11e-70	0.2967180	0.623	0.521	3.36e-66
Prdx5	5.87e-69	0.2692057	0.422	0.314	1.76e-64
Ywhag	5.28e-68	0.2940283	0.585	0.494	1.58e-63

Table 5.4 – Top 15 differentially expressed markers between dysplastic neurons and non-dysplastic neurons in FCD samples.

Functional enrichment the markers was also performed using GO terms (Figure 5.9). Biological processes are shown in Figure 5.8. Many processes associated with normal neuronal function are downregulated in FCD such as 'regulation of synapse structure or activity', 'regulation of synapse structure or activity', 'synaptic transmission, glutamatergic', 'synaptic transmission, GABAergic', and 'membrane biogenesis, synapse organization'. Ion and transmembrane transport and metabolism of nucleosides are upregulated, processes which are associated with tumourgenesis and brain function (270,271).

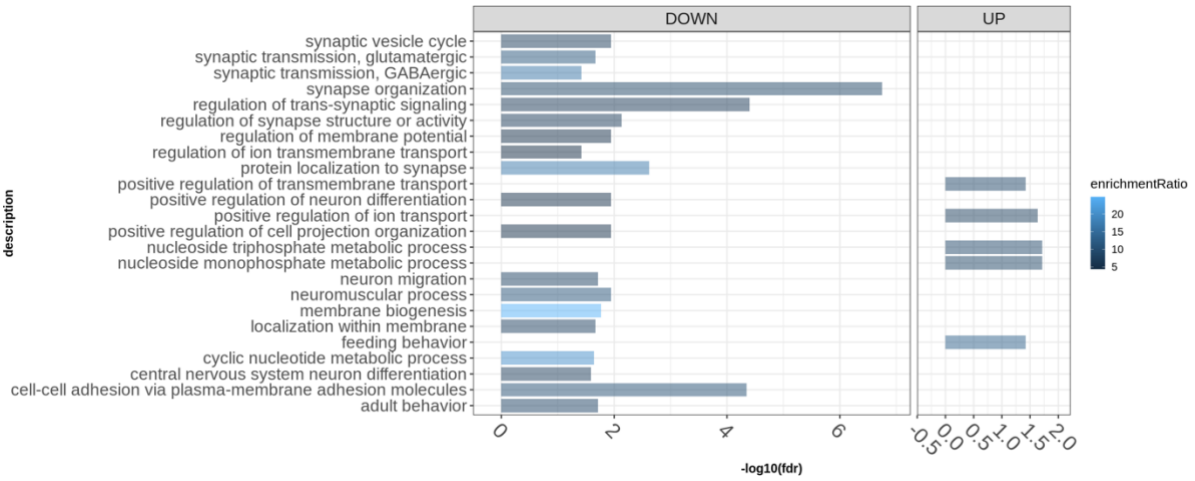


Figure 5.9 – Functional annotation of GO terms in markers of dysplastic neurones in FCD neurones. Significant biological processes at FDR >5% are shown separately for up and down regulate genes. The colours of the bars represent enrichment ratio.

DE was also performed between DNs and all pre-frontal cortex neurones (including sham and control), as some FCD neurones may be dysplastic but plasmids were not identified (Table 5.5). 899 genes were identified as DE (Table 5.4), seven of which are known DN markers; *Rheb, Slc17a7, Map1b, Rps6, Nefl, Syp, and Dcx.*

Functional annotation of these genes again showed additional important neuronal processes such as ‘dendrite development’ and ‘cognition’ as being downregulated in dysplastic neurons (Figure 5.10). Many additional metabolic and redox related terms were upregulated, as well as terms related to oxidative stress such as ‘response to oxidative stress’ and ‘reactive oxygen species metabolic process’. Alterations in metabolic pathways and oxidative stress have been observed in epilepsy both as initiating events and as a response to seizures (79,272). Many terms related to protein function including protein export, folding, importing and localisation are implicated. Protein aggregation and is associated with many neurodegenerative diseases such as Alzheimer’s and Parkinson’s disease which involve the accumulation of misfolded proteins which generates toxicity in the cells, ‘toxin transport’ is also seen to be upregulated in DNs (273). Interestingly, some terms such as ‘synapse organisation’ and ‘signal release’ were both up- and down- regulated, suggesting different genes in the pathways are acting differently.

	P Value	Log2(FC)	% Expressed (DNs)	% Expressed (NNs)	FDR
Scg2	0	0.9676665	0.584	0.317	0
Dbi	0	0.6667104	0.401	0.189	0
Rgs2	0	0.6164991	0.281	0.125	0
Sccpdh	0	0.6070840	0.356	0.164	0
Hpcal4	0	0.6810813	0.456	0.262	0
Trnp1	0	0.6480707	0.374	0.188	0
Efhd2	0	0.6412497	0.449	0.242	0
Rheb	0	0.9981856	0.582	0.301	0
Uchl1	0	0.8032992	0.622	0.408	0
Nptx2	0	0.5928551	0.304	0.143	0
Gpnmb	0	0.4972882	0.122	0.034	0
Npy	0	1.2947318	0.658	0.321	0
Tmsb10	0	0.5943306	0.460	0.252	0
Slc6a5	0	0.4155609	0.107	0.026	0
Thrsp	0	0.4571298	0.170	0.057	0

Table 5.5 – Top 15 differentially expressed markers between dysplastic neurons and non-dysplastic neurons in all pre-frontal cortex samples.

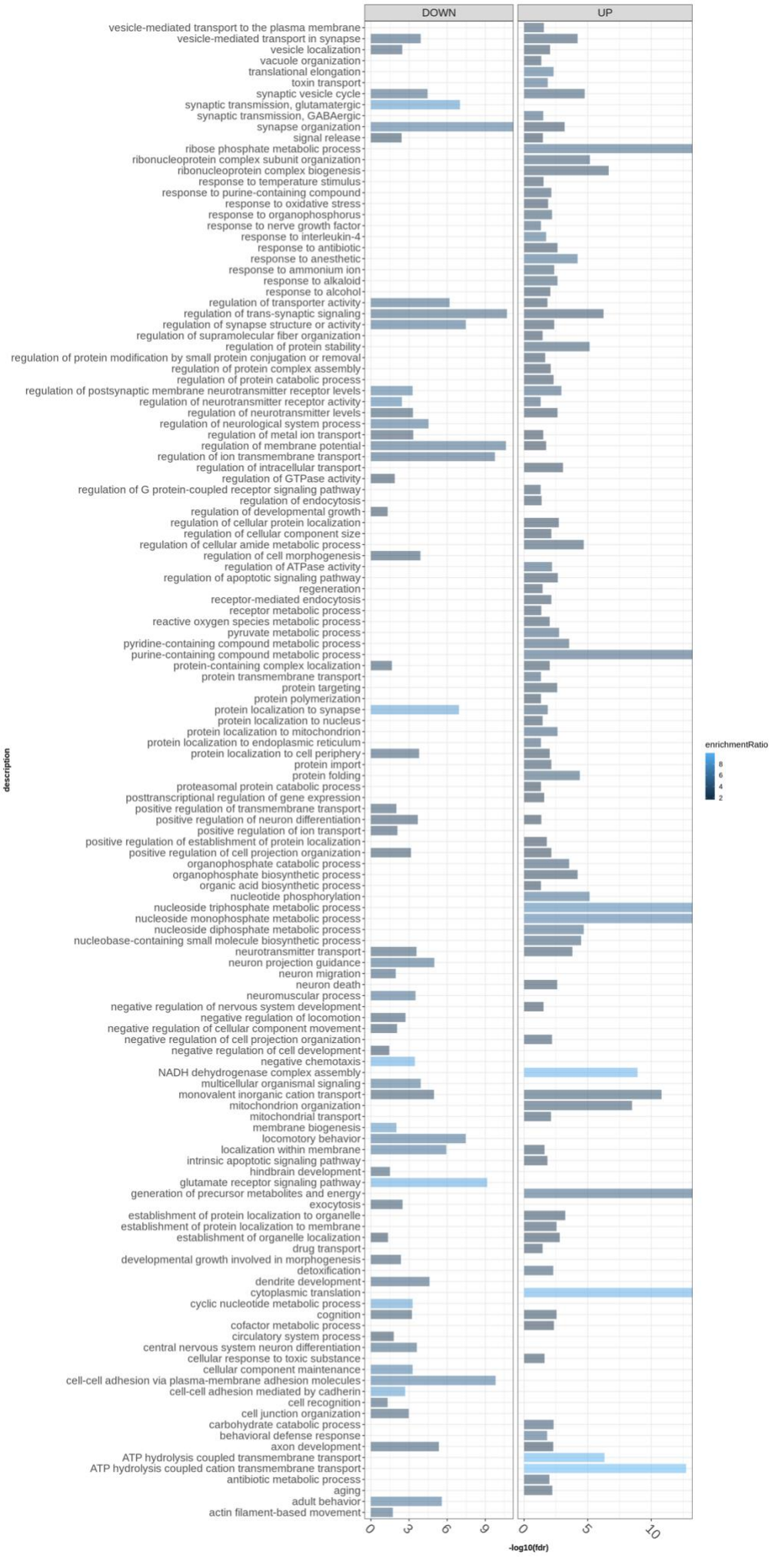

DOWN
UP
vesicle-mediated transport to the plasma membrane
vesicle-mediated transport in synapse
vesicle localization
vacuole organization
translational elongation
toxin transport
synaptic vesicle cycle
synaptic transmission, glutamatergic
synaptic transmission, GABAergic
synapse organization
signal release
ribose phosphate metabolic process
ribonucleoprotein complex subunit organization
ribonucleoprotein complex biogenesis
response to temperature stimulus
response to purine-containing compound
response to oxidative stress
response to organophosphorus
response to nerve growth factor
response to interleukin-4
response to antibiotic
response to anesthetic
response to ammonium ion
response to alkaloid
response to alcohol
regulation of transporter activity
regulation of trans-synaptic signaling
regulation of synapse structure or activity
regulation of supramolecular fiber organization
regulation of protein stability
regulation of protein modification by small protein conjugation or removal
regulation of protein complex assembly
regulation of protein catabolic process
regulation of postsynaptic membrane neurotransmitter receptor levels
regulation of neurotransmitter receptor activity
regulation of neurotransmitter levels
regulation of neurological system process
regulation of metal ion transport
regulation of membrane potential
regulation of ion transmembrane transport
regulation of intracellular transport
regulation of GTPase activity
regulation of G protein-coupled receptor signaling pathway
regulation of endocytosis
regulation of developmental growth
regulation of cellular protein localization
regulation of cellular component size
regulation of cellular amide metabolic process
regulation of cell morphogenesis
regulation of ATPase activity
regulation of apoptotic signaling pathway
regeneration
receptor-mediated endocytosis
receptor metabolic process
reactive oxygen species metabolic process
pyruvate metabolic process
pyridine-containing compound metabolic process
purine-containing compound metabolic process
protein-containing complex localization
protein transmembrane transport
protein targeting
protein polymerization
protein localization to synapse
protein localization to nucleus
protein localization to mitochondrion
protein localization to endoplasmic reticulum
protein localization to cell periphery
protein import
protein folding
proteasomal protein catabolic process
posttranscriptional regulation of gene expression
positive regulation of transmembrane transport
positive regulation of neuron differentiation
positive regulation of ion transport
positive regulation of establishment of protein localization
positive regulation of cell projection organization
organophosphate catabolic process
organophosphate biosynthetic process
organic acid biosynthetic process
nucleotide phosphorylation
nucleoside triphosphate metabolic process
nucleoside monophosphate metabolic process
nucleoside diphosphate metabolic process
nucleobase-containing small molecule biosynthetic process
neurotransmitter transport
neuron projection guidance
neuron migration
neuron death
neuromuscular process
negative regulation of nervous system development
negative regulation of locomotion
negative regulation of cellular component movement
negative regulation of cell projection organization
negative regulation of cell development
negative chemotaxis
NADH dehydrogenase complex assembly
multicellular organismal signaling
monovalent inorganic cation transport
mitochondrion organization
mitochondrial transport
membrane biogenesis
locomotory behavior
localization within membrane
intrinsic apoptotic signaling pathway
hindbrain development
glutamate receptor signaling pathway
generation of precursor metabolites and energy
exocytosis
establishment of protein localization to organelle
establishment of protein localization to membrane
establishment of organelle localization
drug transport
developmental growth involved in morphogenesis
detoxification
dendrite development
cytoplasmic translation
cyclic nucleotide metabolic process
cognition
cofactor metabolic process
circulatory system process
central nervous system neuron differentiation
cellular response to toxic substance
cellular component maintenance
cell-cell adhesion via plasma-membrane adhesion molecules
cell-cell adhesion mediated by cadherin
cell recognition
cell junction organization
carbohydrate catabolic process
behavioral defense response
axon development
ATP hydrolysis coupled transmembrane transport
ATP hydrolysis coupled cation transmembrane transport
antibiotic metabolic process
aging
adult behavior
actin filament-based movement
description
0
3
6
9
0
5
10
-log10(fdr)
enrichmentRatio
8
6
4
2

Figure 5.10 - Functional annotation of GO terms in markers of dysplastic neurones in both FCD and control pre-frontal cortex neurones. Significant biological processes at FDR >5% are shown separately for up and down regulate genes. The colours of the bars represent enrichment ratio.

Finally, enrichment of rare variant from Epi2k WES was performed to see whether markers of dysplastic neurons could be associated with deleterious mutations in epilepsy which could imply a functional role of DNs in the development of epilepsy. Down-regulated genes were enriched for Intolerant mutation in DEE and NAFE (Figure 5.11), suggesting loss-of-function mutations in these marker genes may be driving epilepsy.

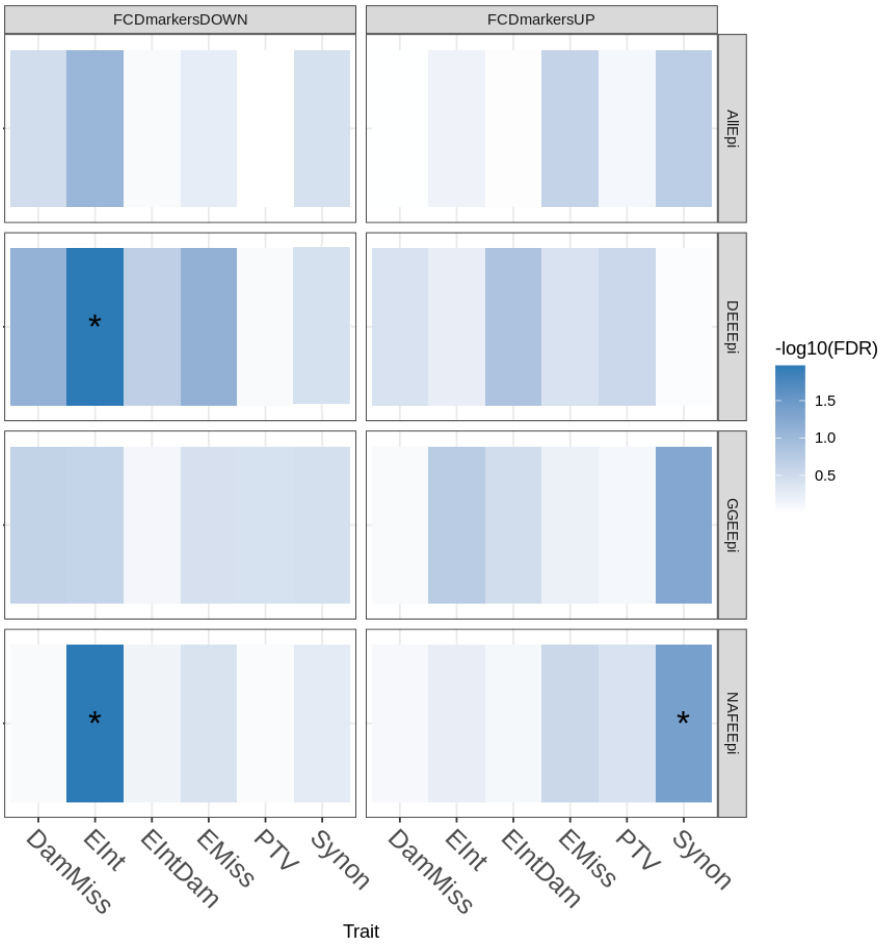


Figure 5.11 – Enrichment of heritability of rare variants using GSA-SNP2 for ultra-rare variants associated with epilepsy subtypes for various categories of coding variants.

5.4 Discussion

Biomarkers play a crucial role in research and can serve as indicators of disease, progression, risk, or treatment response. Diagnosis of FCD can be difficult, particularly in Type I which does not show clear indicators in MRI. Having clear markers of disease is particularly important when early detection can influence disease management and improve patient outcome. FCD has been shown to develop into a more severe disease and early detection and resection could help mitigate this (274).

FCD is characterised by mutations in the mTOR pathway which are often somatic and result in mosaicism (275). This makes studying FCD in humans challenging due to the genetic variably, as it complicates the understanding of the underlying mechanisms. Mouse models here provide a controlled environment to study the mechanisms and the neuropathological features.

This analysis highlighted that the transcriptional landscape of FCD is complex and intricate. Using currently known markers alone may was not sufficient for identifying dysplastic neurons (DNs) as they do not display distinct enough transcriptional signatures to segregate independently from normal neurons. Nonetheless, once the DNs were identified in the mouse data, they did reveal differentially expressed genes to normal neurones, indicating that there are some transcriptional differences in these neurons. These however are not specific to only DNs and may have more subtle and temporal expression patterns. Additional factors may also be contributing to the manifestation of dysplastic neurones beyond gene expression. Studies have suggested epigenetic regulators, such as microRNAs play an important role in FCD and their dysregulation could lead to the pathogenesis (276).

The lack of specificity and widespread nature of these DE genes calls for more sophisticated techniques to discern these neurons in human data. Future work involves incorporating machine learning methods to classify DNs. As the presence of a plasmid sequences in the mouse data serves as a "ground truth" for indicating DNs, this provides a good basis for developing a machine learning model. Methods like Support Vector Machines (SVM) (277)and Gradient Boosting (XGBoost), (278)have the potential to discern subtler patterns that traditional clustering algorithms might miss. Once optimised, this model could then be applied to human datasets which lack clear identifiers of DNs, and allow the detection of these cells. The steps required to build a machine learning models are shown in Figure 5.12. An additional factor to consider in this analysis is that the quantity of reads associated with a cell could influence its dysplasticity, given that increased expression of Rheb may lead to greater mTOR activation. This factor should also be considered in future analysis as more dysplastic cells will display more abnormal properties.

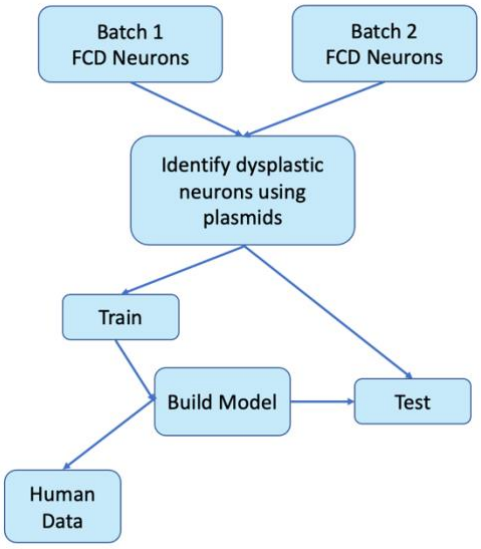


Figure 5.12 – Flow diagram to show how ML models will be used to help classify DNs in future analysis.

5.4.1 Limitations and Future Work

A significant challenge in this analysis was the ability to distinguish dysplastic from non-dysplastic neurons, and identify more specific markers that could be used for characterisation of dysplastic neurones. Single-cell sequencing lacks the spatial context necessary to accurately differentiate between these neuron types, and alternative approaches such as spatial transcriptomics could be used to address this limitation (279). This technique can help identify dysplastic neurons based on their location within the tissue map the gene expression within the anatomical context of the tissue providing a more precise understanding of the cellular and molecular alterations in dysplastic regions (280). Dysplastic neurons, which are morphologically distinct and typically originate from a single focus, can be identified in tissue samples. Integrating this identification into our analysis would enhance our ability to distinguish dysplastic neurons, improving the ability to perform differential expression analysis. This would better allow us to localize individual cells and their transcriptional signatures and elucidating the specific roles of dysplastic cells and cell subpopulations. Targeted pull-down assays also offer complementary strategies to extract the cells expressing the plasmid in order to identify more dysplastic neurons in the mouse tissue. Targeted pull-down assays involve using specific probes or antibodies to selectively enrich cells of interest from a complex mixture. This could allow for a more focused analysis of these cells. Furthermore, integrating other omics techniques, such as proteomics and epigenomics, may also help elucidate the differences that are not apparent through transcriptomics alone.

Chapter 6

Discussion

This chapter provides a summary of the key findings in this thesis and discusses their implications within the field of epilepsy research. It will also address some limitations of the study and suggest potential directions and validations for future work in relation to this analysis. The overarching aim of the work presented in this thesis was to increase the understanding of cell-types and pathways involved in epilepsy as well as identifying potential candidates for drug targeting. By using brain tissue from different types and models of epilepsy, we aimed to find convergent signatures that are perturbed in epilepsy that might point to important mechanisms for effective therapeutic intervention. This was achieved through leveraging single-nucleus datasets as well as several publicly available datasets, in combination with optimised bioinformatics techniques. This integrative approach aims to contribute a novel perspective to the understanding and treatment of epilepsy.

6.1 Cell-types in Epilepsy

Epilepsy is a complex and heterogenous disease. To develop better treatments, we require a better knowledge of the cell-types, genes and biological pathways that contribute to the risk and progression of the disease. The primary focus of this work was to use gene expression to address gaps in the knowledge of cell-types involved in epilepsy and to address whether specific cell-types could be consistently associated with epilepsy. In chapters 3 and 4, cell-type specific DEGs between case and control comparisons are identified and used to link common and rare variants to specific brain cell-types. This revealed extensive dysregulation across all cell-types and epilepsy heritability was found to be significantly enriched in many

of the disease related cell-type specific DEGs, painting a complex picture of epilepsy risk. For the acquired epilepsies, this suggests that the disease state cannot be attributed to a single cell-type or faulty gene expression, and is likely to be the result of a complex interplay between multiple dysregulated genes in several cell-types.

A key aim of the work was to try and identify a convergent signature that could be targeted across diverse types of epilepsy. Despite the clear heterogeneity in gene expression changes astrocytes were highlighted as a potential converging contributor across multiple epilepsy models in both mouse and human tissue, with astrocytic DEG enriched for epilepsy heritability using data from both rare and common forms of epilepsy. Inhibitory and excitatory neurones were also significantly implicated across some comparisons particularly human Hippocampal Sclerosis and Focal Cortical Dysplasia. The glutaminergic synapse was a highly implicated term across several cell-types including glial cells despite this being a pathway associated more commonly with neurones. This was largely seen to be downregulated in neurones and upregulated in glia suggesting a close synergistic neuronal-glial interaction around the glutaminergic synapse, modulating hyperexcitability. Neurones are known to be surrounded by various glia such as astrocytes and microglia which modulate neuronal activity for various neurotransmitters, such as glutamate and GABA, at synapses (74). These dynamic interactions are important in normal synaptic function and any disruption in these interactions could lead to a disruption in the control of excitability. The observed upregulation of gene expression in glial cells may indicate a compensatory mechanism that aims to counterbalance the activity in neurons, maintaining synaptic stability and excitatory equilibrium. Astrocytes have been shown to be involved in the reuptake of glutamate, helping to maintain a balance in the synaptic environment and maintaining homeostasis of neuronal circuits (281–283). Additionally, reactive gliosis of astrocytes which

involves astrocyte proliferation and the release of signalling molecules, has been shown to be both detrimental or protective in epilepsy (284–286). The question of whether gliosis is causal or consequential of disease has been debated, but there is evidence suggesting that astrogliosis leads to seizure generation in mice (287), and reactive astrocytes in human refractory TLE can contribute to the progression of the disease (288). Studies have also confirmed the presence of a subpopulation of specialised "glutaminergic astrocytes" with a discrete molecular signature resembling the glutaminergic synapse. These are involved in exocytosis of glutamate which may help control synaptic strength and transmission and may be a mechanism through which astrocytes are contributing to seizure generation when disturbed (289). Astrocytes have also been shown to release pro-inflammatory molecules and cytokines in response to brain injury and inflammation which is associated with epilepsy, and may also be contributing to neuronal hyperexcitability and seizure activity (171). This great diversity of function may explain the strong converging disease burden around astrocytes and why dysregulated genes in both directions show a strong association with genetic susceptibility.

As well as in astrocytes, and microglia, inflammatory pathways were also implicated in neurons which highlight links to immune function across different forms of epilepsy. Inflammation is increasingly recognised as a significant factor in the development and progression of epilepsy, and has been shown to exacerbate hyperexcitability and lead to alterations in ion channels. Seizures themselves can also cause the release pro-inflammatory cytokines which can in turn alter neuronal excitability. This could also be a consequential response to seizures which can initiate reactive gliosis due to cell death (80,81).

These analyses highlight the variety of cell-types implicated in epilepsy where the focus has previously been centred around neurons. As epilepsy involves a complex network disruption

beyond induvial neurons, where disrupted synchronization and coordination leads to seizures, and dysregulation can arise from alterations in many aspects of synapse regulation, epilepsy is likely to be a complex interplay between cell-types rather than a focus on just one. To show that a cell-type is causal in epilepsy, it needs to be shown that function of this cell-type is perturbed before the onset of epilepsy, and that this change in isolation would lead to the development of epilepsy. Astrocytes have been recently shown to be contribute to early network hyperactivity in Alzheimer's disease through disruptions in calcium signalling, when the calcium signals were restored, seizure susceptibility was recovered in mice. This study implicated astrocytes are key mediators in early AD pathology through disruption of excitability, which could be evidence for an early functional role in seizure generation (290). A better understanding of astrocyte role in epilepsy requires a better understanding of the specific mechanism and pathways that are dysfunctional in order to be able to validate these findings experimentally.

6.2 Genetics of Epilepsy

Epilepsy is a highly heterogenous disease and there is ongoing debate about whether it represents many diseases arising from rare monogenic variation or a large number of discrete polygenic contributions (32). It is increasingly believed that complex diseases can be caused by both rare and common variants, and in the exploration of the underpinnings of epilepsy it is important to consider both types of variants and their implications in gene dysregulation (291). Prior research suggests a convergence of mechanisms between rare and common forms of epilepsy, particularly related to synaptic function (36). In this study we showed that rare and common variants are both enriched in astrocytic dysregulation which could indicate additional shared mechanisms. There is mounting evidence to suggest epilepsy common and rare variants coalesce around the same or at least overlapping sets of genes, (292). Even in

the case of rare epilepsies, small effect common variants could influence the clinical manifestation. The genetic architecture of epilepsy is complex and while over 400 genes have been identified as playing a pathogenic role in epilepsy, these have been concentrated around ion channels and synaptic vulnerability. New mechanism of diseases are needed to find better treatments that maybe better able to target the refractory epilepsies (293). Finding out how genetic variants contribute to common forms of epilepsy can help translate genetic discoveries into relevant information for patients and their families.

6.3 Alternative mechanisms

A recurring theme in this analysis has been the involvement of the mTOR pathway converging across comparisons and species. mTOR is a protein kinase involved in intracellular signalling, and the mTOR pathway is a crucial pathway in the regulation of cell growth, proliferation, protein synthesis and survival (294,295). Dysregulation of both upstream and downstream regulators of mTOR have been shown to cause several genetic types of epilepsies such as FCD and Tubular Sclerosis (TS), which lead to drug refractory seizures termed mTOR-pathies. While this pathway has mainly been implicated in these specific types of epilepsy, it may have implication in other epilepsies too (160).

In this analysis several genes were highlighted to be involved in activating the mTOR pathway. *STAMPB1*, *PIK3CA*, and *PLCB1* were genes which were differentially expressed across 4 or more comparisons in the DE analysis across both species, showing a convergent dysregulation in multiple forms of epilepsy. *NQO1* and *MAT2A* were two of the most significant genes identified as being regulators of disease-related gene expression intersections with CRISPRbrain perturbation data. All of these genes have been shown to be involved in activation of the mTOR pathway (189,191,230,231,296). Additionally, two genes in the mTOR pathway were also found to be significant regulators in this analysis:

PPARGC1A and *EEF2*. The mTOR pathway has more recently gained attention as a therapeutic target for intervention in epilepsy. mTOR inhibitors such as Everolimus and Sirolimus which are approved molecules for certain types of tumours, have also been shown to decrease seizure frequency in TSC (263). It has been shown that modulation of this pathway may have anti-epilepsy effects, however the relationship between mTOR signalling and epilepsy is complex and may involve several interactions with other cellular pathways and environmental factors (297). As well as being associated with mTOR activation, NQO1 was also shown to be involved in oxidative stress pathways. Characterisation of dysplastic neurones in Chapter 5 – cells in which mTOR is overactivated – showed that several oxidative stress pathways were upregulated in these neurons. Oxidative stress has been associated with neuronal hyperexcitability and is an underlying mechanisms in the initiation of seizures and contributes to neuronal degeneration. It has recently been evaluated as a therapeutic target in epilepsy through Nrf2 (298).

An additional mechanism potentially contributing to epilepsy's complexity is the role of homeostatic plasticity. In a healthy brain, excitability is carefully regulated by homeostatic mechanisms and any perturbations in neuronal function should be brought back to normal by compensatory mechanism. As epilepsy involves the abnormal firing of neurones that lead to seizures, epilepsy has been previously described as a failure of homeostatic plasticity or a maladaptive response to perturbations in activity as this perturbation can’t be regulated (185). A convergent mechanism between drug resistant epilepsies or between rare and common epilepsy could be a failure of homeostatic plasticity. Mutations which cause a loss of function in the genes involved in restoring balance may mean the homeostatic mechanism cannot compensate for normal cellular processes. Whether this occurs through astrocytes or other cell-types, it is still in debate exactly how homeostatic plasticity is involved in the

epileptogenic process, and complexity arises from the compensatory responses that need to distinguished from those that drive disease pathogenesis (299,300).

6.4 Drug discovery

Despite a wide range of anti-seizure drugs, one third of patients still have seizures which cannot be treated, and currently, there are no medication that are able to prevent the onset or halt progression of the disease. Past research has primarily focused on drugs targeting synaptic and ion channel proteins, but targeting new mechanisms that contribute to hyperexcitability, such as astrocyte dysfunction of disease may result in the development of better medication (301). In this study we use gene perturbation data to identify upstream regulators that may be causing changes in gene expression leading to epilepsy in specific cell-types. This approach may allow for the identification of better disease-modifying targets, which can realign the disease profile to those of healthy cells. Hypothesis-free, data-driven approaches such as the ones used in this study allow us to study large number of targets unbiasedly, as opposed to a small handful of targets that is often seen with preclinical studies. Having genetic evidence has been shown to be important in the rate of success of approval of a drug, however for this information to be used effectively drug development, we need a mechanistic link between the variant and the disease, as the highest failure rates for drugs occur when targeting proteins with new mechanisms of action or diseases with poorly understood pathogenesis (44,302).

6.5 Limitations

Epilepsy is complex and has an intricate transcriptomics and temporal profile. It is increasingly recognised as a group of syndromes and not just one disease and the issue of drug resistance in epilepsy is recognized as being multifactorial and patient-specific, rather than being caused by a single mechanism. While we see some evidence of convergent signals

in this analysis, the inherent variability necessitates a personalized approach to understanding and treating epilepsy. This adds to the challenge of drug discovery as drug resistance is likely not driven by singular mechanism or variant and is likely diverse and personalised.

It is important to note that while differential expression analysis is a core part of analysis of transcriptional datasets, a key issue is how can we distinguish causal from compensatory changes. The genes identified here can give us a useful insight into the mechanisms being impacted in disease which can help provide information into the pathogenesis. However, looking at DEGs alone cannot tell us whether the differences in gene expression we see, are the differences that are driving the disease, and incorporating additional analyses and methodologies can help uncover the complex molecular basis of epilepsy and provide more comprehensive insights into the disease. Heritability analysis is very useful in determining whether there is a genetic basis to the changes we see in the specific cell-types involved, however establishing causality requires more than just evidence of heritability. It involves demonstrating that changes in the gene or genes directly lead to changes in the trait, which often requires experimental intervention or observation of the effects of natural mutations.

Additionally, is important to look at cells within the context of a complex multidimensional space in which they can be affected by many perturbations. Shifts in gene expression can be caused by many other factors such as changes in cell proportions, or cell sub-types, or cell-cell interactions. A limitation of the CRISPRbrain perturbation analysis comes with the use of induced pluripotent stem cells (iPSCs). The use of iPSCs provides a valuable tool, but static time points may not capture the dynamic nature of the disease, and studying cells in isolation may not fully capture the interactions and network effects that occur in the brain. Due to the complex nature of neuronal networks and homeostasis, even a single change can have

widespread effects. Identifying molecules that can precisely target specific cell-types remains a challenge due to the complexity of cell interactions in the brain. This interconnectivity means that influence on one cell-type will likely lead to secondary effects. Even within a cell-type there can be different responses and a drug may not be as effective against one cell-type. Furthermore, when pursuing disease-modifying drug targets, reliance on mouse models may be a limitation as they also may not accurately model disease modifying effects. Many animal models of neurological disease are known to be isomorphic, meaning they duplicate the disease but not necessarily the underlying aetiology.

In conclusion, integrating genomic approaches and drug repurposing strategies can provide valuable insights into identifying disease-associated targets and potential treatments for epilepsy, however it is essential to consider the complexities and limitations of each approach, as well as the need for comprehensive understanding of disease mechanisms and careful evaluation of downstream effects.

6.6 Future Work

While the above analyses focus on the major cell-type classes, other studies have looked at subtypes or subsets of cells. These "subsets" reflect transcriptionally district subpopulations of the major cell classes. We have chosen not to focus on these in this study, due to the ambiguity associated with this type of analysis. It is unclear whether annotated subtypes are referring to transcriptionally distinct types of cells or whether these differences in gene expression are reflecting different states that the cells are in and so preliminary work has focused on major classes. Future work includes looking at brain layers and regions individually to identify cell-types that belong to particular areas of the brain e.g. the dentate

gyrus which is often where seizures are initiated. Looking at ictogenic subregions again may again help to identify transcriptional differences that may influence epilepsy risk.

Finally, as the analyses in this study were largely centered around genes that were differentially expressed between case and control transcriptomic data, we cannot confidently infer the disease is causing the changes we see and it likely the changes we see are as a result of the disease. This is referred to reverse causality where the cause-and-effect relationship between two variable is unclear. Future work includes further exploring the causal relationship between gene expression and diseases by using additional methodologies such as expression quantitative trait loci (eQTLs) analysis, Mendelian Randomization (MR), and colocalization. These are statistical methods that use genetic variants to determine a causal link between gene expression and disease. A further 70 human epilepsy samples will be added to this analysis to perform a large single-cell epilepsy mendelian randomization as described previously described (303). These methods can better assess whether gene expression is causally linked to epilepsy risk.

Moreover, recent advances in machine learning methods has lead to the emergence of novel methodologies capable of predicting gene expression outcomes following genetic perturbations. Foundation models are pre-trained computational frameworks that have been developed to understand the function and relationships of genes based on expression data from single-cell transcriptomics (304,305). These models may facilitate drug discovery by simulating the cellular impacts of gene perturbations, allowing for the predictions of effects of genes across various conditions, deepening out understanding of the genetic underpinnings of epilepsy. Future work involves applying these models using the snRNA-sequencing epilepsy datasets to enable predictions of candidate therapeutic targets.

The analysis and results presented take the first steps in identifying candidate genes for drug targeting. While bioinformatics serves as a great starting point for research and provides an unbiased exploratory and data-driven methods for finding drug targets, experimental validation in model systems is still necessary. Future work involves experimentally validating some of the drug targets identified in this study to confirm whether their effect and directionality is as predicted. Experimental validation can also help us understand the functional context in which these genes are operating. Conditional knockout models could be used to confirm whether the predicted perturbation effect is indeed causing epilepsy or seizures, or similarly monitoring of synaptic potential in cell cultures could also be used to validate the modulation of genes. While targeting upstream regulators that affect multiple genes can have a disease-modifying impact, it is important to consider potential downstream effects, underscoring the need for rigorous validation. Additionally, as we explore alternative mechanisms in epilepsy treatment, traditional drug testing paradigms may require adaptation to accommodate these novel targets. So while disease-modifying approaches and drug repurposing approaches may speed up the drug discovery process, practical applications need to make significant considerations of various biological, technical, and regulatory factors that add complexity to the process.

References

1. Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE official report: a practical clinical definition of epilepsy. Epilepsia. 2014;55(4):475–82.
2. Epilepsy [Internet]. [cited 2019 May 7]. Available from: https://www.who.int/news-room/fact-sheets/detail/epilepsy
3. Ngugi AK, Bottomley C, Kleinschmidt I, Sander JW, Newton CR. Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. Epilepsia. 2010 May;51(5):883–90.
4. Fisher RS. Redefining epilepsy. Curr Opin Neurol. 2015 Apr;28(2):130–5.
5. Pack BAM. Epilepsy Overview and Revised Classification of Seizures and Epilepsies. Contin Lifelong Learn Neurol. 2019 Apr 1;25(April):306–21.
6. Devinsky O, Vezzani A, O'Brien TJ, Jette N, Scheffer IE, De Curtis M, et al. Epilepsy. Nat Rev Dis Prim 2018 41. 2018 May 3;4(1):1–24.
7. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017 Apr 1;58(4):512–21.
8. Bell GS, Neligan A, Sander JW. An unknown quantity--the worldwide prevalence of epilepsy. Epilepsia. 2014;55(7):958–62.
9. Devinsky O, Spruill T, Thurman D, Friedman D. Recognizing and preventing epilepsy-related mortality. Neurology. 2016 Feb 23;86(8):779–86.
10. Duncan JS, Sander JW, Sisodiya SM, Walker MC. Adult epilepsy. Lancet. 2006 Apr 1;367(9516):1087–100.
11. Berkovic SF, Mulley JC, Scheffer IE, Petrou S. Human epilepsies: interaction of genetic and acquired factors. Trends Neurosci. 2006 Jul 1;29(7):391–7.

12. Dalby NO, Mody I. The process of epileptogenesis: a pathophysiological approach. Curr Opin Neurol. 2001;14(2):187–92.

13. Shao LR, Habela CW, Stafstrom CE. Pediatric Epilepsy Mechanisms: Expanding the Paradigm of Excitation/Inhibition Imbalance. Children. 2019 Feb 1;6(2).

14. Delaney S, Fitzsimons M, White M, Power K, O' Donoghue S, Kilbride R, et al. Analysis of the aetiology of epilepsy in 3,216 adult patients attending a tertiary referral center enabled by an electronic patient record. Seizure. 2020 Oct 1;81:332–7.

15. Pitkänen A, Virtanen AI, Lukasiuk K. Mechanisms of epileptogenesis and potential treatment targets. Vol. 10, Lancet Neurol. 2011.

16. Goldberg EM, Coulter DA. Mechanisms of epileptogenesis: a convergence on neural circuit dysfunction. Nat Rev Neurosci. 2013 May;14(5):337.

17. Vezzani A, Fujinami RS, White HS, Preux PM, Blümcke I, Sander JW, et al. Infections, inflammation and epilepsy. Acta Neuropathol. 2016 Feb 1;131(2):211.

18. Srivastava PK, van Eyll J, Godard P, Mazzuferi M, Delahaye-Duriez A, Steenwinckel J Van, et al. A systems-level framework for drug discovery identifies Csf1R as an anti-epileptic drug target. Nat Commun. 2018 Dec 3;9(1):3561.

19. Vezzani A, Friedman A, Dingledine RJ. The role of inflammation in epileptogenesis. Neuropharmacology. 2013 Jun;69:16–24.

20. Speed D, O'Brien TJ, Palotie A, Shkura K, Marson AG, Balding DJ, et al. Describing the genetic architecture of epilepsy through heritability analysis. Brain. 2014 Oct 1;137(10):2680–9.

21. Berkovic SF, Howell RA, Hay DA, Hopper JL. Epilepsies in twins: genetics of the major epilepsy syndromes. Ann Neurol. 1998 Apr;43(4):435–45.

22. Kjeldsen MJ, Corey LA, Christensen K, Friis ML. Epileptic seizures and syndromes in twins: the importance of genetic factors. Epilepsy Res. 2003 Jun 1;55(1–2):137–46.

23. Steinlein OK, Mulley JC, Propping P, Wallace RH, Phillips HA, Sutherland GR, et al. A missense mutation in the neuronal nicotinic acetylcholine receptor alpha 4 subunit is associated with autosomal dominant nocturnal frontal lobe epilepsy. Nat Genet. 1995;11(2):201–3.

24. Robinson R, Gardiner M. Molecular basis of Mendelian idiopathic epilepsies. 2009;

25. Ding J, Li X, Tian H, Wang L, Guo B, Wang Y, et al. SCN1A Mutation—Beyond Dravet Syndrome: A Systematic Review and Narrative Synthesis. Front Neurol. 2021 Dec 24;12:743726.

26. Mefford HC. Clinical Genetic Testing in Epilepsy. Epilepsy Curr. 2015;15(4):197.

27. McMullin D. Epi4K: Gene discovery in 4,000 genomes. Epilepsia. 2012 Aug 1;53(8):1457.

28. Oliver KL, Ellis CA, Scheffer IE, Ganesan S, Leu C, Sadleir LG, et al. Common risk variants for epilepsy are enriched in families previously targeted for rare monogenic variant discovery. eBioMedicine. 2022 Jul 1;81.

29. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010 Jul;42(7):570–5.

30. McInnes G, Yee SW, Pershad Y, Altman RB. Genomewide Association Studies in Pharmacogenomics. Clin Pharmacol Ther. 2021 Sep 1;110(3):637–48.

31. Anney RJL, Avbersek A, Balding D, Baum L, Becker F, Berkovic SF, et al. Genetic determinants of common epilepsies: A meta-analysis of genome-wide association studies. Lancet Neurol. 2014 Sep 1;13(9):893–903.

32. Abou-Khalil B, Auce P, Avbersek A, Bahlo M, Balding DJ, Bast T, et al. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun. 2018 Dec 1;9(1).

33. Epilepsies ILAEC on C, Berkovic SF, Cavalleri GL, Koeleman BP. Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture. medRxiv. 2022 Jun 14;2022.06.08.22276120.

34. Perucca P, Bahlo M, Berkovic SF. The Genetics of Epilepsy. 2020;

35. McTague A, Howell KB, Cross JH, Kurian MA, Scheffer IE. The genetic landscape of the epileptic encephalopathies of infancy and childhood. Lancet Neurol. 2016 Mar 1;15(3):304–16.

36. Delahaye-Duriez A, Srivastava P, Shkura K, Langley SR, Laaniste L, Moreno-Moral A, et al. Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. Genome Biol 2016 171. 2016 Dec 13;17(1):1–18.

37. Clossen BL, Reddy DS. Novel therapeutic approaches for disease-modification of epileptogenesis for curing epilepsy. Biochim Biophys Acta - Mol Basis Dis. 2017 Jun 1;1863(6):1519–38.

38. Glauser T, Ben-Menachem E, Bourgeois B, Cnaan A, Guerreiro C, Kälviäinen R, et al. Updated ILAE evidence review of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. Epilepsia. 2013 Mar 1;54(3):551–63.

39. Goldenberg MM. Overview of Drugs Used For Epilepsy and Seizures: Etiology, Diagnosis, and Treatment. Pharm Ther. 2010 Jul;35(7):392.

40. Sills GJ, Rogawski MA. Mechanisms of action of currently used antiseizure drugs. Neuropharmacology. 2020 May 15;168:107966.

41. Rho JM, White HS. Brief history of anti-seizure drug development. Epilepsia Open. 2018 Dec 1;3(Suppl Suppl 2):114.

42. Löscher W, Klitgaard H, Twyman RE, Schmidt D. New avenues for anti-epileptic drug

discovery and development. Nat Rev Drug Discov 2013 1210. 2013 Sep 20;12(10):757–76.

43. Borlongan C V. Recent preclinical evidence advancing cell therapy for Alzheimer's disease. Exp Neurol. 2012;237(1):142–6.

44. Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, et al. The druggable genome and support for target identification and validation in drug development. Sci Transl Med. 2017 Mar 29;9(383).

45. Hughes JP, Rees SS, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011 Mar;162(6):1239.

46. Shukla G, Prasad AN. Natural History of Temporal Lobe Epilepsy: Antecedents and Progression. Epilepsy Res Treat. 2012 Mar 26;2012:1–8.

47. De Paola L, Cendes F. Mesial temporal lobe epilepsy syndrome: an updated overview. J Epilepsy Clin Neurophysiol. 2005;11(3):141–4.

48. Lee WS, Baldassari S, Stephenson SEM, Lockhart PJ, Baulac S, Leventer RJ. Cortical Dysplasia and the mTOR Pathway: How the Study of Human Brain Tissue Has Led to Insights into Epileptogenesis. Int J Mol Sci. 2022 Feb 1;23(3).

49. Scerif F. Identification of Gene Networks in Childhood Epilepsy: Evidence from Focal Cortical Dysplasia.

50. Pembroke WG, Hartl CL, Geschwind DH. Evolutionary conservation and divergence of the human brain transcriptome. Genome Biol. 2021 Dec 1;22(1).

51. Kandratavicius L, Alves Balista P, Lopes-Aguiar C, Ruggiero RN, Umeoka EH, Garcia-Cairasco N, et al. Animal models of epilepsy: Use and limitations. Vol. 10, Neuropsychiatric Disease and Treatment. Dove Medical Press Ltd.; 2014. p. 1693–705.

52. Wang Y, Wei P, Yan F, Luo Y, Zhao G. Animal Models of Epilepsy: A Phenotype-

oriented Review. Aging Dis. 2022 Feb 1;13(1):215.

53. Reddy DS, Kuruba R. Experimental Models of Status Epilepticus and Neuronal Injury for Evaluation of Therapeutic Interventions. Int J Mol Sci. 2013 Sep 5;14(9):18284.

54. Nadeau JH, Auwerx J. The virtuous cycle of human genetics and mouse models in drug discovery. Vol. 18, Nature Reviews Drug Discovery. Nature Publishing Group; 2019. p. 255–72.

55. de Klerk E, 't Hoen PAC. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet. 2015 Mar 1;31(3):128–39.

56. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57.

57. Johnson MR, Behmoaras J, Bottolo L, Krishnan ML, Pernhorst K, Santoscoy PLM, et al. Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. Nat Commun 2015 61. 2015 Jan 23;6(1):1–11.

58. Srivastava PK, Bagnati M, Delahaye-Duriez A, Ko J-H, Rotival M, Langley SR, et al. Genome-wide analysis of differential RNA editing in epilepsy. Genome Res. 2017;27(3):440–50.

59. Srivastava PK, Eyll J van, Godard P, Mazzuferi M, Delahaye-Duriez A, Steenwinckel J Van, et al. A systems-level framework for drug discovery identifies Csf1R as an anti-epileptic drug target. Nat Commun 2018 91. 2018 Sep 3;9(1):1–15.

60. McCall MN. Estimation of Gene Regulatory Networks. Postdoc J a J Postdr Res Postdr Aff. 2013 Jan 21;1(1):60.

61. Mu Q, Chen Y, Wang J. Deciphering Brain Complexity Using Single-cell Sequencing. Genomics Proteomics Bioinformatics. 2019 Aug 1;17(4):344–66.

62. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009 65. 2009 Apr 6;6(5):377–82.

63. Grindberg R V., Yee-Greenbaum JL, McConnell MJ, Novotny M, O'Shaughnessy AL, Lambert GM, et al. RNA-sequencing from single nuclei. Proc Natl Acad Sci U S A. 2013 Dec 3;110(49):19802–7.

64. Ragupathy H, Vukku M, Barodia SK. Cell-Type-Specific Mitochondrial Quality Control in the Brain: A Plausible Mechanism of Neurodegeneration. Int J Mol Sci 2023, Vol 24, Page 14421. 2023 Sep 22;24(19):14421.

65. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun 1;15(6):e8746.

66. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017 Aug 18;9(1):1–12.

67. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018 508. 2018 Aug 7;50(8):1–14.

68. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 2017 1410. 2017 Aug 28;14(10):955–8.

69. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019 Jun 20;570(7761):332–7.

70. Zeisel A, M͡oz-Manchado AB, Codeluppi S, Lönnerberg P, Manno G La, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138–42.

71. Szabo PA, Levitin HM, Miron M, Snyder ME, Senda T, Yuan J, et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nat Commun 2019 101. 2019 Oct 17;10(1):1–16.

72. Gorter JA, van Vliet EA, Lopes da Silva FH. Which insights have we gained from the kindling and post-status epilepticus models? J Neurosci Methods. 2016 Feb 15;260:96–108.

73. Coulter DA, Steinhäuser C. Role of Astrocytes in Epilepsy. Cold Spring Harb Perspect Med. 2015;5(3).

74. Patel DC, Tewari BP, Chaunsali L, Sontheimer H. Neuron–glia interactions in the pathophysiology of epilepsy. Nat Rev Neurosci 2019 205. 2019 Feb 21;20(5):282–97.

75. Legido A, Katsetos CD. Experimental Studies in Epilepsy: Immunologic and Inflammatory Mechanisms. Semin Pediatr Neurol. 2014 Sep 1;21(3):197–206.

76. Wong M. The role of glia in epilepsy, intellectual disability, and other neurodevelopmental disorders in tuberous sclerosis complex. J Neurodev Disord. 2019 Dec 16;11(1):1–9.

77. Kwan P, Arzimanoglou A, Berg AT, Brodie MJ, Hauser WA, Mathern G, et al. Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia. 2010;51(6):1069–77.

78. Bromfield EB, Cavazos JE, Sirven JI. Basic Mechanisms Underlying Seizures and Epilepsy. 2006;

79. Aguiar CCT, Almeida AB, Arajo PVP, Abreu RNDC De, Chaves EMC, Vale OC Do, et al. Oxidative Stress and Epilepsy: Literature Review. Oxid Med Cell Longev. 2012;2012.

80. Rana A, Musto AE. The role of inflammation in the development of epilepsy. J Neuroinflammation 2018 151. 2018 May 15;15(1):1–12.

81. Marchi N, Granata T, Janigro D. Inflammatory pathways of seizure disorders. Trends Neurosci. 2014 Feb;37(2):55.

82. Bast T, Ramantani G, Seitz A, Rating D. Focal cortical dysplasia: Prevalence, clinical

presentation and epilepsy in children and adults. Acta Neurol Scand. 2006 Feb;113(2):72–81.

83. Jesus-Ribeiro J, Pires LM, Melo JD, Ribeiro IP, Rebelo O, Sales F, et al. Genomic and Epigenetic Advances in Focal Cortical Dysplasia Types I and II: A Scoping Review. Front Neurosci. 2021 Jan 22;14:580357.

84. Nguyen QH, Pervolarakis N, Nee K, Kessenbrock K. Experimental Considerations for Single-Cell RNA Sequencing Approaches. Front Cell Dev Biol. 2018 Sep 4;6(SEP).

85. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods 2017 144. 2017 Mar 6;14(4):381–7.

86. Armand EJ, Li J, Xie F, Luo C, Mukamel EA. Single cell sequencing of brain cell transcriptomes and epigenomes. Neuron. 2021 Jan 1;109(1):11.

87. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017 Feb 16;65(4):631-643.e4.

88. Zsurka G, Kunz WS. Mitochondrial dysfunction and seizures: the neuronal energy crisis. Lancet Neurol. 2015 Sep 1;14(9):956–66.

89. Staba RJ, Stead M, Worrell GA. Electrophysiological Biomarkers of Epilepsy. Neurotherapeutics. 2014 Apr 1;11(2):334.

90. Svensson V, da Veiga Beltrame E, Pachter L. A curated database reveals trends in single-cell transcriptomics. Database J Biol Databases Curation. 2020;2020.

91. Wen L, Tang F. Recent advances in single-cell sequencing technologies. Precis Clin Med. 2022 Mar 22;5(1).

92. Zappia L, Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biol. 2021 Dec 1;22(1):1–18.

93. Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLOS Comput Biol. 2018 Jun 1;14(6):e1006245.

94. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020 Jan 16;21(1):1–32.

95. Yu L, Cao Y, Yang JYH, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol. 2022 Dec 1;23(1):1–21.

96. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods 2021 191. 2021 Dec 23;19(1):41–50.

97. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet. 2019 Apr 5;10:317.

98. Caglayan E, Liu Y, Konopka G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. Neuron. 2022 Dec 21;110(24):4043-4056.e5.

99. Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. Nat Commun 2021 121. 2021 Feb 2;12(1):1–9.

100. Krishnaswami SR, Grindberg R V., Novotny M, Venepally P, Lacar B, Bhutani K, et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat Protoc. 2016 Mar 1;11(3):499–524.

101. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2023 Nov 9]. Available from:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

102. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017 81. 2017 Jan 16;8(1):1–12.

103. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinforma. 2015 Sep 3;51:11.14.1-19.

104. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D933–41.

105. Eghbalnia HR, Wilfinger WW, Mackey K, Chomczynski P. Coordinated analysis of exon and intron data reveals novel differential gene expression changes. Sci Rep. 2020 Dec 1;10(1):15669.

106. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis.

107. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol. 2020 Aug 27;21(1):1–30.

108. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, Marioni JC. EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019 Mar 22;20(1).

109. Fleming SJ, Chaffin MD, Arduini A, Akkad AD, Banks E, Marioni JC, et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. Nat Methods 2023 209. 2023 Aug 7;20(9):1323–35.

110. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019 Apr 24;8(4):329-337.e4.

111. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016

Feb 17;17(1):1–15.

112. Feleke R, Reynolds RH, Smith AM, Tilley B, Taliun SAG, Hardy J, et al. Cross-platform transcriptional profiling identifies common and distinct molecular pathologies in Lewy body diseases. Acta Neuropathol. 2021 Sep 1;142(3):449.

113. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019 Dec 23;20(1):1–15.

114. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods 2017 146. 2017 May 15;14(6):565–71.

115. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat Protoc 2020 1511. 2020 Oct 12;15(11):3632–62.

116. Chen W, Zhao Y, Chen X, Yang Z, Xu X, Bi Y, et al. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. Nat Biotechnol 2020 399. 2020 Dec 21;39(9):1103–14.

117. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13;177(7):1888-1902.e21.

118. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008 Mar 4;2008(10).

119. Pearson K. LIII. On lines and planes of closest fit to systems of points in space . London, Edinburgh, Dublin Philos Mag J Sci. 1901 Nov;2(11):559–72.

120. Cai TT, Ma R. Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data. J Mach Learn Res. 2021 May 16;23:1–54.

121. Mcinnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020;

122. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2018 371. 2018 Dec 3;37(1):38–44.

123. BB L, R A, GE K, NS S, YC Y, R L, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016 Jun 24;352(6293):1586–90.

124. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018 Dec 12;362(6420).

125. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):1–21.

126. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995 Jan 1;57(1):289–300.

127. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019 Jul 1;47(W1):W199–205.

128. Newton MA, Quintana FA, Boon JA den, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. https://doi.org/101214/07-AOAS104. 2007 Jun 1;1(1):85–106.

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

130. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010 Feb 23;26(7):976–8.

131. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. 2020 Nov 30;9(12):1–10.

132. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018 Oct 1;19(4):562–78.

133. Kharchenko P V., Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740.

134. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127.

135. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015 Dec 10;16(1).

136. Bi R, Liu P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. BMC Bioinformatics. 2016 Mar 31;17(1).

137. Kharchenko P V. The triumphs and limitations of computational methods for scRNA-seq. Nat Methods 2021 187. 2021 Jun 21;18(7):723–32.

138. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004 510. 2004 Sep 15;5(10):1–16.

139. DI Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017 354. 2017 Apr 11;35(4):316–9.

140. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet. 2019 Apr 5;10(APR):441123.

141. McKenzie AT, Wang M, Hauberg ME, Fullard JF, Kozlenkov A, Keenan A, et al. Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. Sci Reports 2018 81. 2018 Jun 11;8(1):1–19.

142. Pfisterer U, Petukhov V, Demharter S, Meichsner J, Thompson JJ, Batiuk MY, et al. Identification of epilepsy-associated neuronal subtypes and gene expression underlying epileptogenesis. Nat Commun. 2020 Dec 1;11(1).

143. Chauhan P, Philip SE, Chauhan G, Mehra S. The Anatomical Basis of Seizures. Epilepsy. 2022 Apr 2;15–24.

144. Righes Marafiga J, Vendramin Pasquetti M, Calcagnotto ME. GABAergic interneurons in epilepsy: More than a simple change in inhibition. Epilepsy Behav. 2021 Aug 1;121:106935.

145. Powell EM. Interneuron Development and Epilepsy: Early Genetic Defects Cause Long-Term Consequences in Seizures and Susceptibility. Epilepsy Curr. 2013;13(4):172.

146. Berg M, Petoukhov I, van den Ende I, Meyer KB, Guryev V, Vonk JM, et al. FastCAR: fast correction for ambient RNA to facilitate differential gene expression analysis in single-cell RNA-sequencing datasets. BMC Genomics. 2023 Dec 1;24(1):1–10.

147. Fleming SJ, Chaffin MD, Arduini A, Akkad A-D, Banks E, Marioni JC, et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. bioRxiv. 2022 Jul 12;791699.

148. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, et al. A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New

Resource for Understanding Brain Development and Function. J Neurosci. 2008 Jan 2;28(1):264–78.

149. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec 29;9(1):1–13.

150. Skene NG, Grant SGN. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. Front Neurosci. 2016 Jan 27;10(JAN):179460.

151. Pai B, Tome-Garcia J, Cheng WS, Nudelman G, Beaumont KG, Ghatan S, et al. High-resolution transcriptomics informs glial pathology in human temporal lobe epilepsy. Acta Neuropathol Commun. 2022 Dec 1;10(1):1–19.

152. Kumar P, Lim A, Hazirah SN, Chua CJH, Ngoh A, Poh SL, et al. Single-cell transcriptomics and surface epitope detection in human brain epileptic lesions identifies pro-inflammatory signaling. Nat Neurosci 2022 257. 2022 Jun 23;25(7):956–66.

153. Chen T, Giri M, Xia Z, Subedi YN, Li Y. Genetic and epigenetic mechanisms of epilepsy: a review. Neuropsychiatr Dis Treat. 2017 Jul 13;13:1841.

154. Iorio F, Shrestha RL, Levin N, Boilot V, Garnett MJ, Saez-Rodriguez J, et al. A Semi-Supervised Approach for Refining Transcriptional Signatures of Drug Response and Repositioning Predictions. PLoS One. 2015 Oct 9;10(10).

155. Yin YH, Ahmad N, Makmor-Bakry M. Pathogenesis of Epilepsy: Challenges in Animal Models. Iran J Basic Med Sci. 2013 Nov;16(11):1119.

156. Becker AJ. Review: Animal models of acquired epilepsy: insights into mechanisms of human epileptogenesis. Neuropathol Appl Neurobiol. 2018 Feb 1;44(1):112–29.

157. Kandratavicius L, Alves Balista P, Lopes-Aguiar C, Ruggiero RN, Umeoka EH, Garcia-Cairasco N, et al. Animal models of epilepsy: use and limitations.

Neuropsychiatr Dis Treat. 2014;10:1693.

158. Curia G, Longo D, Biagini G, Jones RSG, Avoli M. The pilocarpine model of temporal lobe epilepsy. J Neurosci Methods. 2008 Jul 30;172(2):143–57.

159. Thergarajan P, Hudson MR, Carmichael I, Clasadonte J, Dedeurwaerdere S, O'Brien TJ, et al. Characterising seizure development, behavioural comorbidities and neuroinflammation in a self-sustained electrical status epilepticus model of mesial temporal lobe epilepsy in C57BL/6J mice. Neurobiol Dis. 2022 Jun 15;168:105688.

160. Hsieh LS, Wen JH, Claycomb K, Huang Y, Harrsch FA, Naegele JR, et al. Convulsive seizures from experimental focal cortical dysplasia occur independently of cell misplacement. Nat Commun. 2016 Jun 1;7.

161. Mazzuferi M, Kumar G, Rospo C, Kaminski RM. Rapid epileptogenesis in the mouse pilocarpine model: video-EEG, pharmacokinetic and histopathological characterization. Exp Neurol. 2012 Dec;238(2):156–67.

162. Mazzuferi M, Kumar G, Van Eyll J, Danis B, Foerch P, Kaminski RM. Nrf2 defense pathway: Experimental evidence for its protective role in epilepsy. Ann Neurol. 2013 Oct;74(4):560–8.

163. Nguyen LH, Mahadeo T, Bordey A. mTOR Hyperactivity Levels Influence the Severity of Epilepsy and Associated Neuropathology in an Experimental Model of Tuberous Sclerosis Complex and Focal Cortical Dysplasia. J Neurosci. 2019 Apr 4;39(14):2762.

164. Library Construction - Official 10x Genomics Support [Internet]. [cited 2023 Feb 19]. Available from: https://www.10xgenomics.com/support/single-cell-immune-profiling/documentation/steps/library-prep/chromium-single-cell-v-d-j-reagent-kits-user-guide-v-1-1-chemistry-with-feature-barcoding-technology-for-cell-surface-protein

165. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 2019 1612. 2019 Nov 18;16(12):1289–96.

166. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun 2021 121. 2021 Sep 28;12(1):1–15.

167. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet 2019 5111. 2019 Nov 1;51(11):1560–5.

168. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct 10;99(4):877.

169. Yoon S, Nguyen HCT, Yoo YJ, Kim J, Baik B, Kim S, et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Res. 2018;46(10):60.

170. Dräger NM, Sattler SM, Huang CTL, Teter OM, Leng K, Hashemi SH, et al. A CRISPRi/a platform in human iPSC-derived microglia uncovers regulators of disease states. Nat Neurosci 2022 259. 2022 Aug 11;25(9):1149–62.

171. Leng K, Rose IVL, Kim H, Xia W, Romero-Fernandez W, Rooney B, et al. CRISPRi screens in human iPSC-derived astrocytes elucidate regulators of distinct inflammatory reactive states. Nat Neurosci 2022 2511. 2022 Oct 27;25(11):1528–42.

172. Tian R, Abarientos A, Hong J, Hashemi SH, Yan R, Dräger N, et al. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. Nat Neurosci. 2021 Jul 1;24(7):1020.

173. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration

of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 2009 48. 2009 Jul 23;4(8):1184–91.

174. Harding SD, Armstrong JF, Faccenda E, Southan C, Alexander SPH, Davenport AP, et al. The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. Nucleic Acids Res. 2022 Jan 1;50(D1):D1282.

175. Sivapalarajah S, Krishnakumar M, Bickerstaffe H, Chan YY, Clarkson J, Hampden-Martin A, et al. The prescribable drugs with efficacy in experimental epilepsies (PDE3) database for drug repurposing research in epilepsy. Epilepsia. 2018 Feb 1;59(2):492–501.

176. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018 Dec 14;362(6420).

177. Börner K, Teichmann SA, Quardokus EM, Gee JC, Browne K, Osumi-Sutherland D, et al. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. Nat Cell Biol 2021 2311. 2021 Nov 8;23(11):1117–28.

178. Zeisel A, M͡oz-Manchado AB, Codeluppi S, Lönnerberg P, Manno G La, Juréus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-). 2015 Mar 6;347(6226):1138–42.

179. Murphy AE, Skene NG, Arising F, Zimmerman KD, Langefeld MA&, Nature CD. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat Commun 2022 131. 2022 Dec 22;13(1):1–4.

180. Sáinz-Jaspeado M, Smith RO, Plunde O, Pawelzik SC, Jin Y, Nordling S, et al. Palmdelphin Regulates Nuclear Resilience to Mechanical Stress in the Endothelium. Circulation. 2021 Nov 16;144(20):1629–45.

181. Srivastava Y, Donta M, Mireles LL, Paulucci-Holthauzen A, Waxham MN, McCrea PD. Role of a Pdlim5:PalmD complex in directing dendrite morphology. bioRxiv Prepr Serv Biol. 2023 Sep 1;

182. Jean G, Carton J, Haq K, Musto AE. The role of dendritic spines in epileptogenesis. Front Cell Neurosci. 2023 Aug 2;17:1173694.

183. Paz JT, Huguenard JR. Microcircuits and their interactions in epilepsy: Is the focus out of focus? Nat Neurosci. 2015 Mar 27;18(3):351.

184. Pfisterer U, Petukhov V, Demharter S, Meichsner J, Thompson JJ, Batiuk MY, et al. Identification of epilepsy-associated neuronal subtypes and gene expression underlying epileptogenesis. Nat Commun 2020 111. 2020 Oct 7;11(1):1–19.

185. Lignani G, Baldelli P, Marra V. Homeostatic Plasticity in Epilepsy. Front Cell Neurosci. 2020 Jun 26;14:551830.

186. Motelow JE, Povysil G, Dhindsa RS, Stanley KE, Allen AS, Feng YCA, et al. Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. Am J Hum Genet. 2021 Jun 3;108(6):965–82.

187. Nam D, Kim J, Kim S-Y, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms.

188. Golbourn BJ, Halbert ME, Halligan K, Varadharajan S, Krug B, Mbah NE, et al. Loss of MAT2A compromises methionine metabolism and represents a vulnerability in H3K27M mutant glioma by modulating the epigenome. Nat Cancer 2022 35. 2022 Apr 14;3(5):629–48.

189. Cai S, Ye Q, Zeng X, Yang G, Ye C, Chen M, et al. CBS and MAT2A improve methionine-mediated DNA synthesis through SAMTOR/mTORC1/S6K1/CAD pathway during embryo implantation. Cell Prolif. 2021 Jan 1;54(1).

190. Dhediya RM, Joshi SS, Gajbhiye S V., Jalgaonkar S V., Biswas M. Evaluation of

antiepileptic effect of S-adenosyl methionine and its role in memory impairment in pentylenetetrazole-induced kindling model in rats. Epilepsy Behav. 2016 Aug 1;61:153–7.

191. Beaver SK, Mesa-Torres N, Pey AL, Timson DJ. NQO1: a target for the treatment of cancer and neurological diseases, and a model to understand loss of function disease mechanisms ACCEPTED MANUSCRIPT. BBA - Proteins Proteomics. 2019;

192. De Sarro G, Paola ED Di, Gratteri S, Gareri P, Rispoli V, Siniscalchi A, et al. Fosinopril and zofenopril, two angiotensin-converting enzyme (ACE) inhibitors, potentiate the anticonvulsant activity of antiepileptic drugs against audiogenic seizures in DBA/2 mice. Pharmacol Res. 2012 Mar 1;65(3):285–96.

193. Łukawski K, Jakubus T, Raszewski G, Czuczwar SJ. Captopril potentiates the anticonvulsant activity of carbamazepine and lamotrigine in the mouse maximal electroshock seizure model. J Neural Transm. 2010 Oct;117(10):1161–6.

194. Dong X, Fan J, Lin D, Wang X, Kuang H, Gong L, et al. Captopril alleviates epilepsy and cognitive impairment by attenuation of C3-mediated inflammation and synaptic phagocytosis. J Neuroinflammation. 2022 Dec 1;19(1):1–22.

195. Krasniqi S, Daci A. Role of the Angiotensin Pathway and its Target Therapy in Epilepsy Management. Int J Mol Sci. 2019 Feb 1;20(3).

196. Zhang Z, Li Y, Jiang S, Shi FD, Shi K, Jin WN. Targeting CCL5 signaling attenuates neuroinflammation after seizure. CNS Neurosci Ther. 2023 Jan 1;29(1):317–30.

197. Xue K, Gu JJ, Zhang Q, Mavis C, Hernandez-Ilizaliturri FJ, Czuczman MS, et al. Vorinostat, a histone deacetylase (HDAC) inhibitor, promotes cell cycle arrest and re-sensitizes rituximab- and chemo-resistant lymphoma cells to chemotherapy agents. J Cancer Res Clin Oncol. 2016 Feb 1;142(2):379–87.

198. Vezzani A, French J, Bartfai T, Baram TZ. The role of inflammation in epilepsy. Nat

Rev Neurol. 2011 Jan;7(1):31.

199. Tahta A, Turgut M. Focal cortical dysplasia: etiology, epileptogenesis, classification, clinical presentation, imaging, and management. Child's Nerv Syst. 2020 Dec 1;36(12):2939–47.

200. Thergarajan P, Hudson MR, Carmichael I, Clasadonte J, Dedeurwaerdere S, O'Brien TJ, et al. Characterising seizure development, behavioural comorbidities and neuroinflammation in a self-sustained electrical status epilepticus model of mesial temporal lobe epilepsy in C57BL/6J mice. Neurobiol Dis. 2022 Jun 15;168.

201. Kitano H. Systems biology: A brief overview. Science (80-). 2002 Mar 1;295(5560):1662–4.

202. Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019 May 1;20(5):257–72.

203. Caskey CT. The Drug Development Crisis: Efficiency and Safety. 2006;

204. Ferrer I, Vidal N. Neuropathology of cerebrovascular diseases. Handb Clin Neurol. 2018;145:79–114.

205. Walker MC. Hippocampal Sclerosis: Causes and Prevention. Semin Neurol. 2015 Jun 12;35(3):193–200.

206. Boling WW. Surgical Considerations of Intractable Mesial Temporal Lobe Epilepsy. Brain Sci. 2018 Feb 20;8(2).

207. Kabat J, Król P. Focal cortical dysplasia - review. Polish J Radiol. 2012;77(2):35–43.

208. Choi SA, Kim KJ. The Surgical and Cognitive Outcomes of Focal Cortical Dysplasia. J Korean Neurosurg Soc. 2019 May 1;62(3):321.

209. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013 1012. 2013 Oct 6;10(12):1213–

8.

210. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell. 2008 Jan 1;132(2):311.

211. Miraldi ER, Pokrovskii M, Watters A, Castro DM, De Veaux N, Hall JA, et al. Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. Genome Res. 2019 Mar 1;29(3):449–63.

212. Jansen C, Ramirez RN, El-Ali NC, Gomez-Cabrero D, Tegner J, Merkenschlager M, et al. Building gene regulatory networks from scATAC-seq and scRNA-seq using Linked Self Organizing Maps. PLOS Comput Biol. 2019;15(11):e1006555.

213. Muto Y, Wilson PC, Ledru N, Wu H, Dimke H, Waikar SS, et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. Nat Commun 2021 121. 2021 Apr 13;12(1):1–17.

214. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81(3):559.

215. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 8;25(16):2078.

216. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016 4810. 2016 Aug 22;48(10):1279–83.

217. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016 Nov 1;48(11):1443–8.

218. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al.

Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol 2017 361. 2017 Dec 11;36(1):89–94.

219. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021 Nov 1;18(11):1333–41.

220. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Reports 2019 91. 2019 Jun 27;9(1):1–5.

221. Zandigohar M, Dai Y. Information retrieval in single cell chromatin analysis using TF-IDF transformation methods.

222. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 6;184(13):3573.

223. Oliver KL, Scheffer IE, Bennett MF, Grinton BE, Bahlo M, Berkovic SF. Genes4Epilepsy: An epilepsy gene resource. Epilepsia. 2023 May 1;64(5):1368–75.

224. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 2015 4711. 2015 Sep 28;47(11):1228–35.

225. Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. Am J Hum Genet. 2021 Jun 3;108(6):965–82.

226. Wu F, Dai Y, Wang J, Cheng M, Wang Y, Li X, et al. Early-onset epilepsy and microcephaly-capillary malformation syndrome caused by a novel STAMBP mutation in a Chinese boy. Mol Med Rep. 2019;20(6):5145–51.

227. Huang C, Walker EM, Dadi PK, Hu R, Xu Y, Zhang W, et al. Synaptotagmin 4 regulates pancreatic β-cell maturation by modulating the Ca2+ sensitivity of insulin secretion vesicles. Dev Cell. 2018 May 5;45(3):347.

228. Boison D, Rho JM. Epigenetics and epilepsy prevention: the therapeutic potential of adenosine and metabolic therapies. Neuropharmacology. 2020 May 5;167:107741.

229. el Achkar CM, Harrer M, Smith L, Kelly M, Iqbal S, Maljevic S, et al. Characterization of the GABRB2 associated neurodevelopmental disorders. Ann Neurol. 2021 Mar 1;89(3):573.

230. Schoonjans AS, Meuwissen M, Reyniers E, Kooy F, Ceulemans B. PLCB1 epileptic encephalopathies; Review and expansion of the phenotypic spectrum. Eur J Paediatr Neurol. 2016 May 1;20(3):474–9.

231. Tobochnik S, Pisano W, Lapinskas E, Ligon KL, Lee JW. Effect of PIK3CA variants on glioma-related epilepsy and response to treatment. Epilepsy Res. 2021 Sep 1;175.

232. Cavalcante GC, Brito LM, Schaan AP, Ribeiro-Dos-santos Â, de Araújo GS. Mitochondrial Genetics Reinforces Multiple Layers of Interaction in Alzheimer's Disease. Biomedicines. 2022 Apr 1;10(4).

233. Ramocki MB, Bartnik M, Szafranski P, Kołodziejska KE, Xia Z, Bravo J, et al. Recurrent Distal 7q11.23 Deletion Including HIP1 and YWHAG Identified in Patients with Intellectual Disabilities, Epilepsy, and Neurobehavioral Problems. Am J Hum Genet. 2010 Dec 12;87(6):857.

234. Wyatt SK, Witt T, Barbaro NM, Cohen-Gadol AA, Brewster AL. Enhanced classical complement pathway activation and altered phagocytosis signaling molecules in human epilepsy. Exp Neurol. 2017 Sep 1;295:184–93.

235. Klöckner C, Fernández-Murray JP, Tavasoli M, Sticht H, Stoltenburg-Didinger G, Scholle LM, et al. Bi-allelic variants in CHKA cause a neurodevelopmental disorder with epilepsy and microcephaly. Brain. 2022 Jun 1;145(6):1916.

236. Lee H, Oh ET, Choi BH, Park MT, Lee JK, Lee JS, et al. NQO1-induced activation of AMPK contributes to cancer cell death by oxygen-glucose deprivation. Sci Reports 2015 51. 2015 Jan 14;5(1):1–11.

237. Sandouka S, Saadi A, Singh PK, Olowe R, Shekh-Ahmad T. Nrf2 is predominantly

expressed in hippocampal neurons in a rat model of temporal lobe epilepsy. Cell Biosci. 2023 Dec 1;13(1).

238. Kishore M, Pradeep M, Narne P, Jayalakshmi S, Panigrahi M, Patil A, et al. Regulation of Keap1-Nrf2 axis in temporal lobe epilepsy—hippocampal sclerosis patients may limit the seizure outcomes. Neurol Sci. 2023 Dec 1;44(12):4441–50.

239. Tumienė B, Maver A, Writzl K, Hodžić A, Čuturilo G, Kuzmanić-Šamija R, et al. Diagnostic exome sequencing of syndromic epilepsy patients in clinical practice. Clin Genet. 2018 May 1;93(5):1057–62.

240. Rodrigues FB, Wild EJ. Psychogenic non-epileptic seizures in early Huntington's disease. Pract Neurol. 2016 Dec 1;16(6):452–4.

241. Jiang G, Wang W, Cao Q, Gu J, Mi X, Wang K, et al. Insulin growth factor-1 (IGF-1) enhances hippocampal excitatory and seizure activity through IGF-1 receptor-mediated mechanisms in the epileptic brain. Clin Sci (Lond). 2015;129(12):1047–60.

242. Lin Z, Li J, Ji T, Wu Y, Gao K, Jiang Y. ATP1A1 de novo Mutation-Related Disorders: Clinical and Genetic Features. Front Pediatr. 2021 Apr 21;9:657256.

243. Sun J, Zheng Y, Chen Z, Wang Y. The role of Na+-K+-ATPase in the epileptic brain. CNS Neurosci Ther. 2022 Sep 1;28(9):1294–302.

244. Dostanic I, Schultz JEJ, Lorenz JN, Lingrel JB. The α1 isoform of Na,K-ATPase regulates cardiac contractility and functionally interacts and co-localizes with the Na/Ca exchanger in heart. J Biol Chem. 2004 Dec 24;279(52):54053–61.

245. Gherghiceanu M, Barad L, Novak A, Reiter I, Itskovitz-Eldor J, Binah O, et al. Cardiomyocytes derived from human embryonic and induced pluripotent stem cells: Comparative ultrastructure. J Cell Mol Med. 2011;15(11):2539–51.

246. Buchin A, de Frates R, Nandi A, Mann R, Chong P, Ng L, et al. Multi-modal characterization and simulation of human epileptic circuitry. Cell Rep. 2022 Dec

27;41(13):111873.

247. Heise C, Taha E, Murru L, Ponzoni L, Cattaneo A, Guarnieri FC, et al. eEF2K/eEF2 Pathway Controls the Excitation/Inhibition Balance and Susceptibility to Epileptic Seizures. Cereb Cortex. 2017 Mar 1;27(3):2226–48.

248. Highet B, Parker R, Faull RLM, Curtis MA, Ryan B. RNA Quality in Post-mortem Human Brain Tissue Is Affected by Alzheimer's Disease. Front Mol Neurosci. 2021 Dec 21;14:780352.

249. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol 2020 211. 2020 Feb 3;21(1):1–16.

250. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci U S A. 2018 Jul 24;115(30):7723–8.

251. Barkovich AJ, Guerrini R, Kuzniecky RI, Jackson GD, Dobyns WB. A developmental and genetic classification for malformations of cortical development: update 2012. Brain. 2012;135(5):1348.

252. Rossini L, Medici V, Tassi L, Cardinale F, Tringali G, Bramerio M, et al. Layer-specific gene expression in epileptogenic type II focal cortical dysplasia: Normal-looking neurons reveal the presence of a hidden laminar organization. Acta Neuropathol Commun. 2014 Jan 27;2(1):1–15.

253. Najm I, Lal D, Alonso Vanegas M, Cendes F, Lopes-Cendes I, Palmini A, et al. The ILAE consensus classification of focal cortical dysplasia: An update proposed by an ad hoc task force of the ILAE diagnostic methods commission. Epilepsia. 2022 Aug 1;63(8):1899–919.

254. Najm IM, Sarnat HB, Blümcke I. Review: The international consensus classification of Focal Cortical Dysplasia – a critical update 2018. Neuropathol Appl Neurobiol. 2018

Feb 1;44(1):18–31.

255. Mihara T, Matsuda K, Tottori T, Otsubo T, Kubota Y, Inoue Y, et al. Focal Cortical Dysplasia and Epilepsy Surgery. J Epilepsy Res. 2013;3(2):43.

256. Aronica E, Boer K, Baybis M, Yu J, Crino P. Co-expression of cyclin D1 and phosphorylated ribosomal S6 proteins in hemimegalencephaly. Acta Neuropathol. 2007 Sep 5;114(3):287–93.

257. Li L, Liu CQ, Li TF, Guan YG, Zhou J, Qi XL, et al. Analysis of Altered Micro RNA Expression Profiles in Focal Cortical Dysplasia IIB. J Child Neurol. 2016 Apr 1;31(5):613–20.

258. Lee JY, Park AK, Lee ES, Park WY, Park SH, Choi JW, et al. miRNA expression analysis in cortical dysplasia: regulation of mTOR and LIS1 pathway. Epilepsy Res. 2014;108(3):433–41.

259. Löscher W, Potschka H, Sisodiya SM, Vezzani A. Drug Resistance in Epilepsy: Clinical Impact, Potential Mechanisms, and New Innovative Treatment Options. Pharmacol Rev. 2020 Jul 1;72(3):606.

260. Kumari K, Sharma MC, Kakkar A, Malgulwar PB, Pathak P, Suri V, et al. mTOR pathway activation in focal cortical dysplasia. Ann Diagn Pathol. 2020 Jun 1;46:151523.

261. Baldassari S, Ribierre T, Marsan E, Adle-Biassette H, Ferrand-Sorbets S, Bulteau C, et al. Dissecting the genetic basis of focal cortical dysplasia: a large cohort study. Acta Neuropathol. 2019 Dec 1;138(6):885–900.

262. Jansen LA, Mirzaa GM, Ishak GE, O'Roak BJ, Hiatt JB, Roden WH, et al. PI3K/AKT pathway mutations cause a spectrum of brain malformations from megalencephaly to focal cortical dysplasia. Brain. 2015 Jun 1;138(Pt 6):1613–28.

263. Griffith JL, Wong M. The mTOR pathway in treatment of epilepsy: a clinical update.

Future Neurol. 2018 May 1;13(2):49.

264. Panwar V, Singh A, Bhatt M, Tonk RK, Azizov S, Raza AS, et al. Multifaceted role of mTOR (mammalian target of rapamycin) signaling pathway in human health and disease. Signal Transduct Target Ther 2023 81. 2023 Oct 2;8(1):1–25.

265. Hsieh LS, Wen JH, Claycomb K, Huang Y, Harrsch FA, Naegele JR, et al. Convulsive seizures from experimental focal cortical dysplasia occur independently of cell misplacement. Nat Commun 2016 71. 2016 Jun 1;7(1):1–12.

266. Szczurkowska J, Cwetsch AW, Dal Maschio M, Ghezzi D, Ratto GM, Cancedda L. Targeted in vivo genetic manipulation of the mouse or rat brain by in utero electroporation with a triple-electrode probe. Nat Protoc. 2016 Mar 1;11(3):399–412.

267. Bushnell A, Bushnell B, Egan R, Copeland A, Foster B, Clum A, et al. Lawrence Berkeley National Laboratory LBL Publications Title BBMap: A Fast, Accurate, Splice-Aware Aligner Permalink https://escholarship.org/uc/item/1h3515gn BBMap: A Fast, Accurate, Splice-Aware Aligner. 2014.

268. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009 Dec 15;10(1):1–9.

269. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

270. Mullen NJ, Singh PK. Nucleotide metabolism: a pan-cancer metabolic dependency. Nat Rev Cancer 2023 235. 2023 Mar 27;23(5):275–94.

271. Garcia-Gil M, Camici M, Allegrini S, Pesi R, Petrotto E, Tozzi MG. Emerging Role of Purine Metabolizing Enzymes in Brain Function and Tumors. Int J Mol Sci. 2018 Nov 14;19(11).

272. Patel M. A Metabolic Paradigm for Epilepsy. Epilepsy Curr. 2018 Sep 1;18(5):318.

273. Da Cruz S, Cleveland DW. Disrupted nuclear import/export in neurodegeneration. Science. 2016 Jan 1;351(6269):125.

274. Achiriloaie A, Deisch J, Boling W, Bannout F. Striking MRI Changes of Focal Cortical Dysplasia Over Time: A Case Series and Literature Review. Neurol Clin Pract. 2021 Oct 1;11(5):445.

275. Mirzaa GM, Campbell CD, Solovieff N, Goold CP, Jansen LA, Menon S, et al. Wide spectrum of developmental brain disorders from megalencephaly to focal cortical dysplasia and pigmentary mosaicism caused by mutations of MTOR. JAMA Neurol. 2016 Jul 7;73(7):836.

276. Avansini SH, Torres FR, Vieira AS, Dogini DB, Rogerio F, Coan AC, et al. Dysregulation of NEUROG2 plays a key role in focal cortical dysplasia. Ann Neurol. 2018 Mar 1;83(3):623.

277. Cristianini N, Ricci E. Support Vector Machines. Encycl Algorithms. 2008;928–32.

278. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min.

279. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016 Jul 1;353(6294):78–82.

280. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for in situ tissue profiling. Nat Methods. 2019 Oct 1;16(10):987–90.

281. Coulter DA, Eid T. Astrocytic Regulation of Glutamate Homeostasis in Epilepsy. Glia. 2012 Aug;60(8):1215.

282. Rose CR, Felix L, Zeug A, Dietrich D, Reiner A, Henneberger C. Astroglial Glutamate Signaling and Uptake in the Hippocampus. Front Mol Neurosci. 2017 Jan 17;10.

283. Khakh BS, Deneen B. The Emerging Nature of Astrocyte Diversity. https://doi.org/101146/annurev-neuro-070918-050443. 2019 Jul 8;42:187–207.

284. Verhoog QP, Holtman L, Aronica E, van Vliet EA. Astrocytes as Guardians of Neuronal Excitability: Mechanisms Underlying Epileptogenesis. Front Neurol. 2020 Nov 26;11:591690.

285. Vezzani A, Ravizza T, Bedner P, Aronica E, Steinhäuser C, Boison D. Astrocytes in the initiation and progression of epilepsy. Nat Rev Neurol 2022 1812. 2022 Oct 24;18(12):707–22.

286. Lentini C, d'Orange M, Marichal N, Trottmann MM, Vignoles R, Foucault L, et al. Reprogramming reactive glia into interneurons reduces chronic seizure activity in a mouse model of mesial temporal lobe epilepsy. Cell Stem Cell. 2021 Dec 2;28(12):2104-2121.e10.

287. Robel S, Buckingham SC, Boni JL, Campbell SL, Danbolt NC, Riedemann T, et al. Reactive astrogliosis causes the development of spontaneous seizures. J Neurosci. 2015;35(8):3330–45.

288. Chen ZP, Wang S, Zhao X, Fang W, Wang Z, Ye H, et al. Lipid-accumulated reactive astrocytes promote disease progression in epilepsy. Nat Neurosci 2023 264. 2023 Mar 20;26(4):542–54.

289. de Ceglia R, Ledonne A, Litvin DG, Lind BL, Carriero G, Latagliata EC, et al. Specialized astrocytes mediate glutamatergic gliotransmission in the CNS. Nat 2023 6227981. 2023 Sep 6;622(7981):120–9.

290. Shah D, Gsell W, Wahis J, Luckett ES, Jamoulle T, Vermaercke B, et al. Astrocyte calcium dysfunction causes early network hyperactivity in Alzheimer's disease. Cell Rep. 2022 Aug 23;40(8):111280.

291. Akingbuwa WA, Hammerschlag AR, Bartels M, Nivard MG, Middeldorp CM. Ultra-

rare and common genetic variant analysis converge to implicate negative selection and neuronal processes in the aetiology of schizophrenia. Mol Psychiatry. 2022 Sep 1;27(9):3699.

292. Allen AS, Bellows ST, Berkovic SF, Bridgers J, Burgess R, Cavalleri G, et al. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. Lancet Neurol. 2017 Feb 1;16(2):135–43.

293. Oyrer J, Maljevic S, Scheffer IE, Berkovic SF, Petrou S, Reid CA. Ion Channels in Genetic Epilepsy: From Genes and Mechanisms to Disease-Targeted Therapies. Pharmacol Rev. 2018 Jan 1;70(1):142.

294. Ben-Sahra I, Manning BD. mTORC1 signaling and the metabolic control of cell growth. Curr Opin Cell Biol. 2017 Apr 1;45:72–82.

295. Saxton RA, Sabatini DM. mTOR Signaling in Growth, Metabolism, and Disease. Cell. 2017 Mar 9;168(6):960–76.

296. McDonell LM, Mirzaa GM, Alcantara D, Schwartzentruber J, Carter MT, Lee LJ, et al. Mutations in STAMBP, encoding a deubiquitinating enzyme, cause microcephaly-capillary malformation syndrome. Nat Genet. 2013 May;45(5):556–62.

297. Moloney PB, Cavalleri GL, Delanty N. Epilepsy in the mTORopathies: opportunities for precision medicine. Brain Commun. 2021 Oct 1;3(4).

298. Carmona-Aparicio L, Pérez-Cruz C, Zavala-Tecuapetla C, Granados-Rojas L, Rivera-Espinosa L, Montesinos-Correa H, et al. Overview of Nrf2 as Therapeutic Target in Epilepsy. Int J Mol Sci. 2015 Aug 7;16(8):18348.

299. Chaligne R, Gaiti F, Silverbush D, Schiffman JS, Weisman HR, Kluegel L, et al. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. Nat Genet 2021 5310. 2021 Sep 30;53(10):1469–79.

300. Curia G. Hebbian and homeostatic synaptic plasticity of AMPA receptors in

epileptogenesis. Cell Reports Med. 2023;4:101047.

301. Talevi A. Antiseizure medication discovery: Recent and future paradigm shifts. Epilepsia Open. 2022 Aug 1;7(Suppl 1):S133.

302. Ochoa D, Karim M, Ghoussaini Maya, Hulcoop DG, McDonagh EM, Dunham I. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov. 2022 Aug 1;21(8):551.

303. Haglund A, Zuber V, Yang Y, Abouzeid M, Feleke R, Ko JH, et al. Single-cell Mendelian randomisation identifies cell-type specific genetic effects on human brain disease and behaviour. bioRxiv. 2022 Nov 29;2022.11.28.517913.

304. Theodoris C V., Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nat 2023 6187965. 2023 May 31;618(7965):616–24.

305. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell 2022 410. 2022 Sep 26;4(10):852–66.

Appendix

Tables in .xls format can be found at https://github.com/Mybzd/Appendices:

1. Appendix Table 1: Cell-type specific differential expression analyses results for all mouse comparisons

2. Appendix Table 2: Functional pathway enrichments for all mouse comparisons

4. Appendix Table 3: All results of CRISPR brain perturbation analysis in mouse

5. Appendix Table 4: Functional pathway enrichments of CRISPRbrain perturbations

6. Appendix Table 5: Cell-type specific differential expression analyses results for all human comparisons

7. Appendix Table 6: All results of CRISPR brain perturbation analysis in human

7. Appendix Table 7: Functional pathway enrichments across all human comparisons

8. Appendix Table 8: All marker genes of dysplastic neurons