Exploring the gene regulatory dynamics of the maturing human brain

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

by
Stephanie Fillmore (FLLSTE005)

Supervisor: Dr Dorit Hockman

Department of Health Sciences

University of Cape Town

South Africa

August 2022

Plagiarism Declaration

I, Stephanie Fillmore, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Stephanie Fillmore

2 August 2022

Acknowledgements

I would like to acknowledge my supervisor Dr Dorit Hockman for her constant support and guidance.

Secondly, I would like to acknowledge the neurosurgeons at Red Cross War Memorial Children's hospital and Constantiaberg Mediclinic for providing us with the brain tissue samples that were used in this study.

I would like to thank the National Research Fund for funding this project.

Finally, I would like to acknowledge my fellow lab members, my friends and family for their encouragement and support.

Abstract

The human brain develops gradually overtime where distinct molecular profiles are established in the embryo. These molecular profiles continue to change through aging and in response to environmental factors. The complexity and dynamics of gene expression and regulation at the cell type-specific level are still poorly understood, especially during the process of brain maturation. The overall aim of this project was to obtain a better understanding of how the brain cell atlas changes over time by contributing to the current brain cell atlas with pediatric single cell data. Bio-banked pediatric and adult brain tissue samples, obtained during surgery to treat epilepsy, were used to optimise and generate a nuclei isolation protocol. Single nuclei RNA-seq (snRNA-seq) libraries were generated using the 10x Genomics Platform. snRNA-seq datasets were then sequenced and analysed using bioinformatics tools, including *Cell Ranger* and *Seurat*. The major cell types in the pediatric brain were identified, including the genes being expressed by these cell types. In addition, a pilot differential expression analysis study was conducted between snRNA-seq libraries from the temporal and frontal lobes. Furthermore, Assays for Transposase Accessible Chromatin (ATAC-seq) was performed on pediatric and adult tissue and bulk ATAC-seq libraries were successfully generated. A consensus list of putative enhancers and promoters was generated after testing several bioinformatic pipelines. Differential accessibility analysis was performed on the bulk ATAC-seq datasets and the promoters or enhancers that are being dynamically used over the course of brain development, were also identified. Ultimately, with these findings and with the generation of optimised protocols, this study has contributed to our understanding of gene expression and gene regulation of brain maturation.

Contents

Chapter 4: Discussion

Supplementary file 1

Supplementary file 2

Supplementary file 3

Supplementary file 4

Supplementary file 5

Supplementary file 6

Supplementary file 7

Supplementary file 8

Supplementary file 9

Supplementary file 10

Supplementary file 11

Chapter 1 – Introduction

This chapter provides a literature review about how the brain matures throughout the human lifespan. Most importantly, this review gives a brief outline of how we can study the brain, discover new cell types and ultimately how we can use this information to understand what happens in response to disease. Previous studies conducted thus far have all contributed to the human brain cell atlas, however there is currently very little data from studies on the pediatric brain and this forms the foundation of the research project. Finally, I highlight that this is also a method optimisation project, where we aim to establish snRNA-seq using the 10x Genomics platform and ATAC-seq technology in South Africa. We hope to make these cutting-edge tools accessible to all South African researchers and help boost research outputs coming from South Africa.

1.1 Brain development

1.1.1 Human brain evolution

The human brain is the most complex organ in the body comprising of billions of interconnected neurons which are surrounded by glial or non-neuronal cells (Darmanis *et al.*, 2015; Lake *et al.*, 2018). This complexity provides us with several cognitive and motor capabilities; however, it may also have enhanced our vulnerability to different neurological and psychiatric disorders which will be discussed later in this review. What makes our brain so unique and what distinguishes our brain from that of our ancestors, other mammals, and other primates?

The size of the human brain is the first main difference, with our brains being three times the size of chimpanzees and twice the size of pre-human hominids and the most significant increase seen in the cerebral cortex (Carroll, 2003; Geschwind and Rakic, 2013). This size expansion gives rise to another important difference – a longer period of neural formation resulting in a primate cerebral cortex circuit organization that is distinct from other mammals (Geschwind and Rakic, 2013). The cerebral cortex neurons are arranged in layers and columns to form connected circuits and during cortex evolution, this circuit diagram has increased in size, number and complexity (Kornack and Rakic, 1998; Geschwind and Rakic, 2013). During

human brain development neurons will migrate into the developing cerebral cortex forming six distinct layers (Cooper, 2008), numbered I-VI (**Figure 1.1**). Cortical layer VI is the deepest layer that appears first and the top layers (I, II/III) form later, which has been described as the "inside-out" migration of neurons (Cooper, 2008) (**Figure 1.1**). These distinct layers contain different types of neurons which in turn have different morphologies and functions.

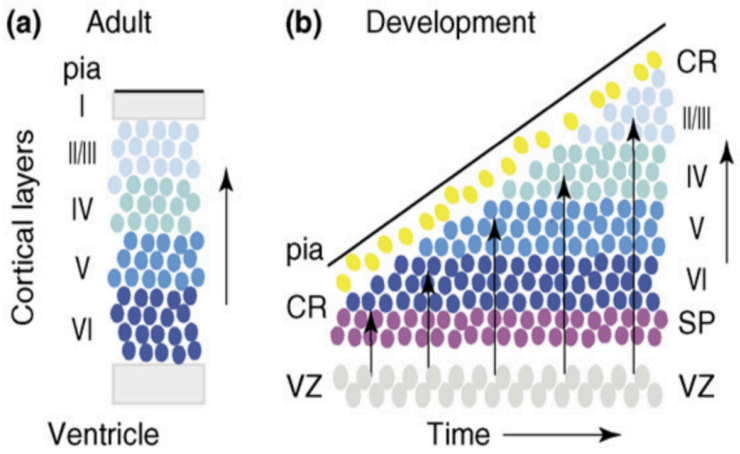


Figure 1.1: ***Development of the human cerebral cortex. A schematic of adult and developing cortical plates are shown.*** **(a)** The adult cortical plate is divided into 6 layers (I-VI) from outside (pia) to inside (ventricle). The layering shows the oldest neurons at the bottom and the youngest ones at the top which are colour coded in different shades of blue. The development of the cortical plate over time is shown in **(b)**. The arrows pointing up from the ventricular zone (VZ) represents the migration of the developing neurons. VZ, ventricular zone; SP, subplate cells (purple); CR, Cajal-Retzius cells (yellow). (Figure adapted from Cooper et al., 2008).

1.1.2 Embryonic, pediatric, and adult brain development

The layers of the cerebral cortex are characterised by the distinct molecular profiles within their different cell types which are established in the embryo and continue to develop into infancy. These distinct cell types are also governed by gene regulatory networks which play significant roles in maintaining cell function and directing cell differentiation (John L.R. Rubenstein, 2011; Kang *et al.*, 2011). Furthermore, we know that brain development is extremely complex requiring tight regulation of gene expression over time (John L R Rubenstein, 2011). Neuronal circuit development relies on the diverse and precise expression of gene products at the right time and place, which is referred to as spatiotemporal regulation (Johnson *et al.*, 2009). Despite the progress that has been made in characterising cortical neurodevelopment

many molecular mechanisms remain unknown, particularly for pediatric neurodevelopment. Previous studies have largely been limited to fetal and adult development (Fan *et al.*, 2018; Fullard *et al.*, 2018; Lake *et al.*, 2018; McKenzie *et al.*, 2018; Zhong *et al.*, 2018; Hodge *et al.*, 2019; Polioudakis *et al.*, 2019; Bakken *et al.*, 2021) and as a result there is limited data, specifically transcriptomic and gene regulation data, for pediatric brain development. This gap in the knowledge forms the basis of this research project.

As a result of its complexity, the brain takes approximately two decades to build; starting its development during the prenatal periods and continuing into the twenties or even thirties depending on the brain region (Gogtay *et al.*, 2004; Cooper, 2008; Stiles and Jernigan, 2010). Some of the first important processes that occur are during the embryonic and early fetal periods such as neurogenesis and neuronal migration. By the end of these periods, most of the cortical neurons have been generated and have migrated to their specific areas of the cortex (Bystron, Blakemore and Rakic, 2008; Stiles and Jernigan, 2010). A timeline of human development and periods designed by Kang *et al*. divides human neurodevelopment into 15 distinct periods spanning from embryonic development to late adulthood (Kang *et al.*, 2011) (**Table 1.1**).

Gene expression over the course of brain development has extremely complex and dynamically regulated patterns. Previous studies have shown that these gene expression profiles differ more temporally and spatially than they do between males and females or individuals (Johnson *et al.*, 2009; Kang *et al.*, 2011; Tebbenkamp *et al.*, 2014; Werling *et al.*, 2020). These studies used human brain tissue samples from a broad range of ages and developmental stages (5.7 post-conceptional weeks (pcw) to 82 years-old). In addition, these studies also focused on different brain regions including the amygdala, hippocampus, striatum, dorsomedial nucleus of the thalamus, cerebellum, cerebellar cortex, and several regions of the neocortex (Johnson *et al.*, 2009; Kang *et al.*, 2011; Werling *et al.*, 2020). Using the transcriptomic data, they were able to analyse the expression dynamics of genes that are associated with certain neurodevelopmental processes (**Figure 1.2**). These analyses show that the most significant changes in gene expression occur during embryonic and fetal development (**Table 1.1**), also known as the late-fetal transition, until it reaches a plateau during late infancy and suggests that prenatal development is the most robust and complex phase of neurodevelopment (Kang *et al.*, 2011; Werling *et al.*, 2020) (**Figure 1.2**). Two neurodevelopmental genes that were enriched during the late-fetal transition include *OPALIN* and *IGF2BP1* (Werling *et al.*, 2020). *OPALIN*, a transmembrane protein also known as *TMEM10*, has been shown to promote the expression of myelin genes during oligodendrocyte differentiation (de Faria et al., 2019) and *IGF2BP1* is found in axons or dendrites of developing neurons and promote mRNA localisation (Farina *et al.*, 2003).

However, the above studies used bulk transcriptomic techniques such as DNA microarrays and RNA-seq, which have certain limitations. These techniques do not provide enough detail about gene expression

profiles at the cellular level (Darmanis *et al.*, 2015) because the final output will be an average measurement of a group of cells, masking important differences between individual cells (Ofengeim *et al.*, 2017; Regev *et al.*, 2017; Hwang, Lee and Bang, 2018). It is possible that there are changes in cell type-specific gene expression during later stages of development which are being masked in these bulk transcriptomic analyses.

Table 1.1: Periods of human development and adulthood (adapted from Kang et. al, 2011)

Periods	Definition	Age
1, 2, 3	Embryonic and early fetal	4 PCW ≤ Age < 13 PCW
4, 5	Early mid-fetal	13 PCW ≤ Age < 19 PCW
6, 7	Late mid-fetal and late fetal	19 PCW ≤ Age < 38 PCW
8, 9	Neonatal and early infancy and late infancy	0M (birth) ≤ Age < 12M
10	Early childhood	1Y ≤ Age < 6Y
11	Middle and late childhood	6Y ≤ Age < 12Y
12	Adolescence	12Y ≤ Age < 20Y
13, 14, 15	Young, middle, and late adulthood	20Y ≤ Age

M, postnatal months; PCW, post-conceptional weeks; Y, postnatal years

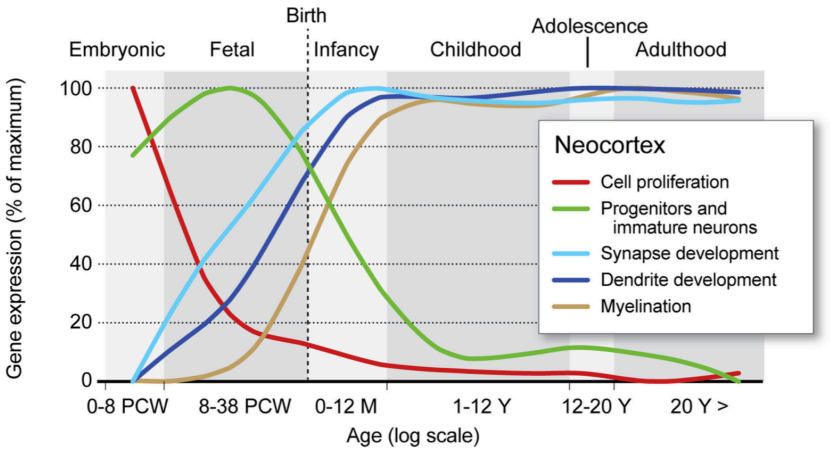


Figure 1.2: ***Timeline of important human neurodevelopmental processes based on gene expression trajectories.*** Gene expression trajectories associated with certain developmental processes such as cell proliferation or myelination reflect when these processes occur and how they progress in the human neocortex (NCX). The expression levels and trajectories have been adopted from Kang et al. 2011. PCW, post-conceptional weeks; M, months; Y, years. (Taken from Tebbenkamp et al. 2014).

1.2 Cell types of the human brain

This research project focuses on the cerebral cortex of the temporal and frontal lobes. As already stated, these different areas of the brain are comprised of hundreds or even thousands of distinct cell types. Each of these cell types express a unique set of genes which allow them to carry out specific functions (Darmanis *et al.*, 2015; Silbereis *et al.*, 2016). To get a better understanding of brain function and the function of neural circuits it is important to understand its cellular components. For over a century, biologists have taken on this challenging task of characterising and grouping cells into distinct types to create a human brain cell atlas. But the description of these cell types specifically in the brain, is constantly evolving. Previous approaches used to classify cells were based on different parameters such as cell morphology, physiological properties or expression of certain marker genes (Heng *et al.*, 2008; Defelipe *et al.*, 2013). With the development of genomic profiling techniques such as DNA microarrays and bulk RNA-sequencing, we have been able to obtain more detailed characterisations of cell types, grouped by brain region or developmental stages (Heller, 2002; Abrahams *et al.*, 2007; Colantuoni *et al.*, 2011). However, as previously mentioned, these techniques have limitations. More recently, there has been advances in these genomic profiling techniques to allow for the interrogation of the entire transcriptome at the level of individual cells, known as single-cell RNA sequencing (scRNA-seq) (Picelli *et al.*, 2013; Macosko *et al.*, 2015). Performing scRNA-seq, allows for the discovery of thousands of differentially expressed genes per cell type. By analysing the dynamics of gene expression, we can

classify different cell types and begin to understand what gene expression changes occur during brain maturation.

These technological advancements have resulted in the drive to complete the Human Cell Atlas – a reference map of all human cells which is based on their features and distributions (Regev *et al.*, 2017, 2018). In addition, there is also the BRAIN Initiative Cell Census Consortium which aims to generate a whole-brain cell atlas in humans (Ecker *et al.*, 2017). It is important to note that these are also developmental atlases that seek to describe how the cellular composition of the brain changes throughout the human lifespan and therefore it is imperative to include data from a wide range of ages (Taylor *et al.*, 2019). This research project will generate transcriptomic data from the pediatric brain, filling in the gap created by the lack of pediatric data and contributing to the human brain cell atlas.

These initiatives, such as the BRAIN Initiative, aim to define all the different neuronal and glial cells found within each layer of the cerebral cortex. The six-layered cerebral cortex of the mature brain consists of hundreds of different neuronal and glial cell types and the classification of these cell types remain largely incomplete. The progenitor cells of the central nervous system (CNS), known as neural progenitor cells (NPCs), gives rise to the neural cells and most glial cells of the CNS (Martínez-Cerdeño and Noctor, 2018). What are some of the major neuronal and glial cell types that we should be able to identify in this study?

1.2.1 Neuronal Cell types

There are two main types of cortical neuronal cells: excitatory pyramidal neurons and inhibitory local interneurons (Cajal, 1906; DeFelipe and Fariñas, 1992). Excitatory neurons, also known as glutamatergic neurons, have axons that extend and can transmit information between areas of the cerebral cortex but also other regions of the brain (Anderson, 2002). Excitatory neurons make up most of the neurons in the cortex and are typically distributed within all cortical layers except layer I (DeFelipe and Fariñas, 1992; Molyneaux *et al.*, 2007). Interneurons make local connections within the cortex that also span the different layers of the cortex, are mostly inhibitory and are GABAergic (Cauli *et al.*, 1997; Molyneaux *et al.*, 2007). This broad classification of neurons is being expanded with the introduction of scRNA-seq and the identification of new molecular markers for neuronal sub-types (Yuste *et al.*, 2020). The function of neuronal cells is to make connections with other neurons across the different layers and other cortical areas, generating highly complex information processing networks (Rakic and Lombroso, 1998; Stiles and Jernigan, 2010).

1.2.2 Glial cell types

Glial cells include four main groups: (1) astrocytes, (2) microglia, (3) oligodendrocytes, and (4) oligodendrocyte precursor cells (OPCs) (Jäkel and Dimou, 2017). Astrocytes are involved in maintaining water and ion homeostasis, uptake of glutamate and GABA and play a role in blood brain barrier (BBB) maintenance (Kimelberg and Nedergaard, 2010). Microglia are the resident immune cells in the brain. They have been shown to survey the extracellular environment of healthy tissue and play a role in synaptic pruning (Jäkel and Dimou, 2017). OPCs generate mature oligodendrocytes and have been proposed to carry out specific functions such as generating new myelinating oligodendrocytes after injury (Gensert and Goldman, 1997) and may play a role in maintaining the BBB (Seo *et al.*, 2014). Mature oligodendrocytes have been shown to produce myelin, allowing for insulation of axons (Simons and Trajkovic, 2006; Kuhn *et al.*, 2019). But what governs the development and maintenance of all these various cell types?

1.3 Gene regulatory networks

Gene regulatory networks (GRNs) are complex systems, showing hierarchical interactions underlying development, cell morphology and function (MacNeil and Walhout, 2011). Ultimately, they control the development and maintenance of all cell types. Gene expression is activated by several active transcription factors (TFs) that interact with a set of *cis*-regulatory elements or modules in the genome (Fiers *et al.*, 2018). Each *cis*-regulatory element (CRE) can interact with multiple TFs produced by multiple different genes. Additionally, each TF can interact with multiple CREs and these interactions between genes and TFs can be represented as a GRN (Davidson, McClay and Hood, 2003; Sanguinetti and Walker, 2019) (**Figure 1.3**). One way that a developmental GRN can be depicted simply is through nodes and edges, with nodes, representing the gene and their regulators and edges, representing the physical and regulatory interactions (MacNeil and Walhout, 2011) (**Figure 1.3**).

1.3.1 Key Players in the GRN

One of the most important components of a GRN are enhancers. Enhancers are non-coding DNA elements or distal CREs that are characterised by low nucleosome occupancy (Erokhin *et al.*, 2015; Carullo and Day, 2019). The overall function of enhancers is to activate and increase transcription by bringing transcription factors to the promoter, thus regulating gene expression (Calo and Wysocka, 2013; Erokhin *et al.*, 2015) (**Figure 1.3**). More specifically, they can regulate gene expression through enhancer-promoter loops where the enhancer will physically interact with the promoter of a target gene (**Figure 1.3**). The Encyclopedia of DNA Elements (ENCODE) project, a ground-breaking study that started in 2003, (ENCODE Project, 2011; Moore *et al.*, 2020) aimed to annotate all the functional elements in the human genome, including enhancers. They performed and integrated several sequence-based studies (such as RNA-seq, DNase I hypersensitive sites sequencing (DNase-seq) and Chromatin Immunoprecipitation with massively parallel sequencing (ChIP-seq)) generating 4 834 datasets from different human tissues or cells. For example, there are approximately 245 DNase-seq datasets from human adult or fetal brain tissue and cell lines. Some of the human brain cell lines include Human Astrocytes-cerebellar (HAc), Human Astrocytes-hippocampal (HAh), Human astrocytes spinal cord (HA-sp) and Human Brain Microvascular Endothelial cells (HBMEC). One part of the project involved identifying and annotating enhancers, where they were able to identify around 926,535 putative enhancers from human tissues or cells. In the human brain, enhancers likely determine which genes are active during neural specification and which ones will remain accessible in mature neurons (Carullo and Day, 2019). How do we identify the enhancers and other key GRN components to build an understanding of the GRNs that are at play in the human brain?

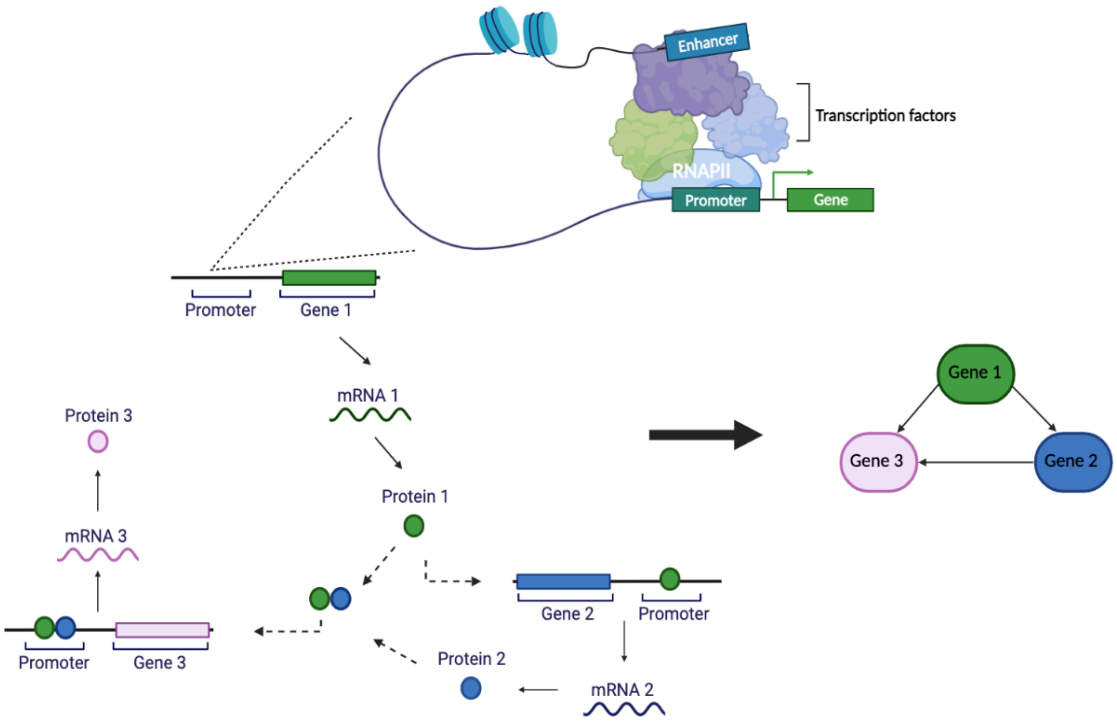


Figure 1.3: ***Schematic representation of a gene regulatory network and its components.*** The interaction between three genes on the left shows direct regulation (gene 1 on gene 2) and combinatorial regulation via complex formation (gene 3 by genes 1 and 2). This complex system is represented as a GRN on the right, where gene 1 is required for the regulation of gene 2 and 3 and gene 2 is required for the regulation of gene 3. Genes are shown as nodes and their interactions represented as the edges. Zooming in on one node (gene 1) and its specific components in the top right section, showing gene regulation in more detail. Chromosomal looping allows for distal enhancer elements to interact with the gene's promoter. It also brings in transcription factors that ultimately increases and activates expression of the gene. (Made in Biorender).

1.3.2 Molecular tools for interrogating GRNs in the brain

Mapping the complex interactions that make up GRNs is a great challenge and more specifically, our knowledge about the developmental GRNs in the human brain remains incomplete. DNA microarrays, RNA-seq, Assays for Transposase-Accessible Chromatin by sequencing (ATAC-seq), DNase-seq and other methylation analysis techniques have been used to identify the components such as genes or CREs of the

GRN (Fullard *et al.*, 2018). The next two sections of this literature review go into detail about what tools have been used and how they have been improved over the years. Studies that have used these techniques to identify genes or enhancers that are important for neurodevelopment are also highlighted.

1.4 Studying gene expression in the developing human brain

Initial approaches to understand transcriptomics in the brain during development, which used techniques such as microarrays and bulk RNA-seq, have led to important discoveries – showing distinct gene expression patterns over the course of neurodevelopment as mentioned previously (Johnson *et al.*, 2009; Colantuoni *et al.*, 2011; Kang *et al.*, 2011). Moreover, the development of scRNA-seq has allowed us to interrogate the entire transcriptome at the level of single cells and thereby fill in the gaps that previous techniques were unable to fill. These techniques have also been applied to identify networks of co-expressed genes, which will allow us to begin to understand developmental GRNs in the human brain (McKenzie *et al.*, 2018). What do these techniques entail and how have they been improved over the years?

1.4.1 Microarray and bulk RNA-seq techniques

The microarray technique involves the covalent binding of known nuclei acid probes to glass slides. Fluorescently labelled cDNA generated from samples is then hybridised to the probes on the slides and imaged to detect differential hybridisation between samples (Heller, 2002; Mantione *et al.*, 2014). RNA-seq technology on the other hand, involves the analysis of gene expression on a genome-wide scale using technologies such as Illumina Next Generation sequencing platforms, without relying on prior knowledge of the gene sequence (Mantione *et al.*, 2014; Marioni *et al.*, 2019). Research groups such as the Allen Brain Institute, used DNA microarrays in their pioneering studies on the human and mouse brain. They were able to characterise broad cell classes of the brain and began generating gene expression maps of different regions of the human and mouse brain (Hawrylycz *et al.*, 2012; Henry and Hohmann, 2012). These datasets formed part of a publicly accessible data resource that researchers from all over the world can use (https://www.brainspan.org/).

In 2012, Hawrylycz *et al.* performed microarray analysis on frozen adult human tissue from two donors (ages 24 and 39 years-old) and generated an extensive map of gene expression across the whole adult brain (Hawrylycz *et al.*, 2012). To identify gene expression patterns related to specific cell types, they applied weighted gene co-expression network analysis (WGCNA) on the microarray datasets, where genes were grouped into 13 distinct modules (Hawrylycz *et al.*, 2012). WGCNA is a data collection method that identifies groups of genes that are related to specific cell types (Zhang *et al*., 2005). This allowed them to identify and characterise four broad classes of cells: neurons, astrocytes, microglia, and oligodendrocytes according to the different brain regions (Hawrylycz *et al.*, 2012). Interestingly, the neocortex showed relatively uniform transcriptional patterns compared to other regions of the brain. However, as mentioned in the paper, they might have seen greater variation in the areas of the neocortex if the analyses were performed at the level of specific cell types (Hawrylycz *et al.*, 2012). When analysing whole tissue, these techniques do not provide as much detail on cell-type specific gene expression because the final output will be an average measurement of the cells, masking the differences between them (Ofengeim *et al.*, 2017; Regev *et al.*, 2017).

Colantuoni *et al.* used oligonucleotide microarrays to analyse the RNA from post-mortem human brain tissue covering a vast range of ages, starting at 14-20 gestational weeks through to 80 years in the prefrontal cortex (Colantuoni *et al.*, 2011). They generated data for 30,176 probes and discovered unique gene expression patterns that begin during fetal development. They showed that during fetal development, gene expression changes are faster than any other life stage (**Figure 1.4**). There is then a significant decline in the rate of gene expression change, which is maintained through to the twenties. Interestingly, after the thirties and forties the rate of gene expression changes starts to increase again (**Figure 1.4**). As mentioned above, several other studies have seen similar gene expression changes, where they also show that the greatest differences in gene expression occurs before birth (Johnson *et al.*, 2009; Kang *et al.*, 2011; Werling *et al.*, 2020).

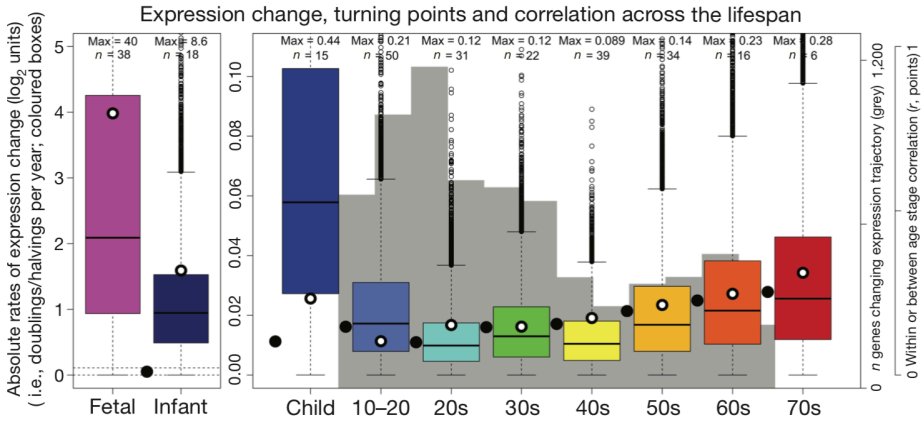


Figure 1.4: ***Rate of gene expression change in the human brain across the lifespan.*** The absolute rate of expression change for each developmental stage was quantified and plotted. The y-axis scale is different for fetal and infant stages than for all the other stages because the gene expression changes are faster than all other developmental stages. The horizontal dotted line on the left represents the entire extent of the y-axis on the right. (From: Colantuoni *et al.,* 2011).

1.4.2 Single cell RNA sequencing (scRNA-seq)

To overcome the challenges presented by bulk RNA-seq and microarray technologies, scRNA-seq, has been used to analyse gene expression profiles on a single cell level. scRNA-seq involves the combination of isolating individual cells, the creation of single cell cDNA libraries and then performing massively parallel RNA-seq on the resulting libraries (Lake *et al.*, 2016; Hu *et al.*, 2017). We can look at the expression of thousands of genes per cell without selecting specific genes. In addition, scRNA-seq allows us to identify rare populations of cells which would not be possible from a pooled population of cells. This technique provides a greater understanding and characterisation of transcriptomic profiles and molecular mechanisms of normal cells (Zeisel *et al.*, 2015; Ofengeim *et al.*, 2017). A lot of work has already been done applying scRNA-seq to the human brain, providing major insights into the complexity of neurodevelopment by identifying different cell types and novel cell subtypes (Darmanis *et al.*, 2015; Lake *et al.*, 2016; Polioudakis *et al.*, 2018; Zhong *et al.*, 2018; Hodge *et al.*, 2019; Bakken *et al.*, 2021).

1.4.3 What are the different scRNA-seq library preparation methods?

Isolating single cells or nuclei is the first and most important step for obtaining cell type-specific gene expression information (Hwang, Lee and Bang, 2018). There are many different techniques, each having their own set of advantages and disadvantages. Since the first scRNA-seq publication (Tang *et al.*, 2009), several different isolation techniques have emerged but in this review, only the most commonly used methods will be covered. These techniques can be classified into two groups: (1) high-sensitivity plate-based (which normally requires sorting cells into wells) and (2) high-throughput droplet-based microfluidic methods.

Plate-based methods such as Smart-seq2 or CEL-Seq2 works by sorting or capturing individual cells into multiwell plates (Picelli *et al.*, 2013; Ofengeim *et al.*, 2017). Smart-seq2 relies on template switching to generate full-length cDNA. In addition to generating full-length cDNA, it provides good read coverage across transcripts and detects very high numbers of genes per cell (Picelli *et al.*, 2013). Importantly, this allows users to detect gene isoforms, rare transcripts or allele-specific expression using single nucleotide polymorphisms (SNPs) (Deng *et al.*, 2014). Even though this method is very sensitive, it is also low throughput as the total number of cells are limited to 96 cells per plate (Picelli *et al.*, 2013). Smart-seq2 also relies on an additional sorting step and fluorescence-activated cell sorting (FACs) has become one of the most commonly used sorting techniques. This is because it can detect cells that express low levels of markers (Julius, Masuda and Herzenberg, 1972). Briefly, the cells or nuclei are labelled with a fluorescent monoclonal antibody which targets cell-surface markers and allows you to sort and isolate specific populations of cells or nuclei according to fluorescence (Julius, Masuda and Herzenberg, 1972). Nuclei can also be sorted and stained using fluorescent dyes such as Hoechst (Latt *et al.*, 1975; Kubbies, 1990). They are then isolated into individual wells of the multiwell plate (Julius, Masuda and Herzenberg, 1972). However, the drawbacks of FACs include the need for monoclonal antibodies to target cell-surface proteins of interest and FACS requires starting with a high number of cells.

The Fluidigm C1 platform is a plate-based microfluidic technique and was the first commercially available cell capture platform (Ofengeim *et al.*, 2017). It uses microfluidics to capture and process 96 to 800 individual cells per plate for transcriptome analysis (Pollen *et al.*, 2014). The most important advantage of the Fluidigm system is that users can visualise and check each captured cell under the microscope. Because of this, users can then exclude doublets, empty wells or wells that contain cell debris (Pollen *et al.*, 2014; See *et al.*, 2018). However, the disadvantage of this system is that only cells of a certain size can be captured in a single run because of the chip that is used.

High-throughput droplet-based methods, such as the 10X Genomics Chromium solution, involves the isolation of individual cells in nano-litre oil droplets along with individual barcoded beads using a chip-based microfluidics system (Zheng *et al.*, 2017; 10x, 2019). Each bead is coated in oligonucleotides that contain a unique barcode, a UMI and a poly(dT) primer (**Figure 1.5**). This is an alternative to the Fluidigm C1 platform which also uses microfluidics but captures cells in wells and not nano droplets. Once the droplets containing individual cells and 10x Gel beads are formed, the beads are dissolved, and the poly(dT) primer captures the poly-adenylated mRNA. Lastly, barcoded cDNA is generated (Zheng *et al.*, 2017). The 10x Genomics Chromium platform has become one of the most popular library preparation protocols for scRNA-seq and has recently been set up in the Hockman lab. This technique will be discussed in more detail throughout this project but here I will summarise why it has become the gold standard technique for single cell or nucleus isolation. Previous benchmark studies have determined that 10x Genomics Chromium was the best performer compared to other high-throughput methods (Ding *et al.*, 2019).

Firstly, the main advantage of this platform is that it allows high-throughput analysis of thousands of cells (up to 10,000) in comparison to well-based approaches which are usually applied to a few hundred cells at a time. As a result, the 10x Genomics Chromium method allows the detection of rarer cell types which may otherwise be missed (Hwang, Lee and Bang, 2018). This platform allows you to detect thousands of cells and is also less labour intensive because the reactions for all the cells are conducted in a single tube and not across multiple well plates. A possible disadvantage of using the UMI tag-based approach is that the sequencing will only be carried out from the 3' end of the RNA and not the full transcript. This means that alternative splice forms cannot be detected (Ofengeim *et al.*, 2017). Finally, 10x Genomics has streamlined or automated most of the downstream processing steps, making this method less time consuming and much simpler in comparison to other methods that might require additional sorting steps.

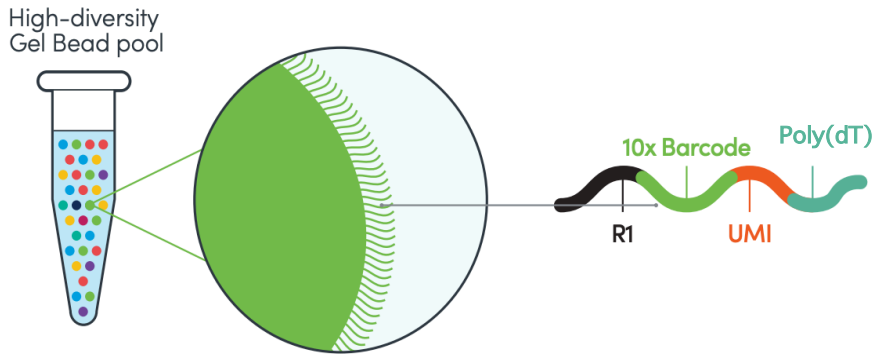


Figure 1.5: ***Structure of the 10x™ Gel bead.*** Each bead in the pool is coated with thousands of oligodendrocytes. These oligodendrocytes consist of a Read1 sequencing adapter (R1), a 10x cell Barcode, Unique Molecular Identifier (UMI) and a Poly(dT) primer sequence. (Adapted from: 10x, 2019).

Darmanis *et al.* conducted scRNA-seq on a group of 466 cells using the Fluidigm C1 platform. These cells were obtained from healthy adult temporal lobe tissue and embryonic cortical brain tissue to establish a cellular map of the human brain. They were able to group each cell into one of the main cell types of the brain, either neuronal, glial or vascular cell types using two different approaches: an unbiased approach and a focused approach. (Darmanis *et al.*, 2015). The unbiased approach classified the cells into 10 clusters according to their molecular signatures. From the set of enriched genes obtained, they were able to identify classic cell-type specific markers that were then used to assign cell types to each cluster resulting in a total of 8 distinct clusters. The focused approach used existing data from the mouse brain, where they were able to identify a set of human cell-type specific marker genes and using these marker genes the cells were grouped into 7 distinct clusters (Darmanis *et al.*, 2015). The results of each clustering approach were compared and there was a high similarity between the identity of the unbiased clusters compared to the biased clusters. This showed that scRNA-seq data can be used to identify the different cell types in the brain without selecting a set of genes beforehand. This is advantageous especially when this information might not always be available (Darmanis *et al.*, 2015). The initial clustering of the 466 cells using the unbiased approach revealed that the fetal brain cells were classified into two distinct groups: replicating and quiescent fetal neurons and in particular, these neuronal cell clusters were separate from the adult brain cell clusters (Darmanis *et al.*, 2015). From this they were able to perform a principal component analysis (PCA) on pre- and postnatal neurons which identified gene expression gradients between the replicating and quiescent fetal neurons and neuronal progenitors (Darmanis *et al.*, 2015). These genes that showed an expression gradient reflect the change between replicating and quiescent neurons. In addition, this analysis allowed them to determine the distinction between neuronal progenitors of the prenatal brain and mature postnatal neurons which has helped us obtain a better understanding of what happens as the brain matures.

A study by Polioudakis *et al.* used scRNA-seq to describe the transcriptomes for all the main cell types during human brain development, focusing on the cortical tissue from gestation week 17 to 18 (Polioudakis *et al.*, 2018). They applied Drop-seq (Macosko *et al.*, 2015) on 40 000 cells and Fluidigm C1 on a smaller subset of cells to compare the two techniques. Using an unbiased clustering approach, they identified 16 distinct cell clusters including: oligodendrocytes progenitor cells (OPCs), radial glia (RG), intermediate progenitors (IP), microglia, interneurons, migrating excitatory neurons, pericytes and endothelial cells (Polioudakis *et al.*, 2018) (**Figure 1.6**). Additionally, using the bioinformatics tool, *Monocle 2*, they were able to cluster the cells by pseudo-time in an unbiased manner and perform lineage trajectory reconstruction (Polioudakis *et al.*, 2018). This analysis was able to show how radial glia cells transition to intermediate progenitors and subsequently to migrating neurons (Polioudakis *et al.*, 2018). Using this method, they were

able to find connections between these different cell types during fetal brain development and provide insight into the process of neurogenesis (Polioudakis *et al.*, 2018).

1.4.4 Single nucleus RNA sequencing (snRNA-seq)

scRNA-seq has become a widely used technique to study and generate transcriptomic profiles at a cellular level (Darmanis *et al.*, 2015; Lake *et al.*, 2016; Zhong *et al.*, 2018). Even though this has proven to be a very powerful tool, it only works well with cells that are easily isolated. It has been shown that non-neuronal cells survive harsh dissociation better than neurons (Habib *et al.*, 2017). Complex cells such as neurons, specifically mature neurons, have shown to be a challenge to isolate because they are highly interconnected, and they end up being underrepresented in the final dataset compared to non-neuronal cells (Hu *et al.*, 2017). scRNA-seq also requires fresh tissue to prepare single cell suspensions which is extremely limiting (Darmanis *et al.*, 2015; Habib *et al.*, 2017; Hu *et al.*, 2017). As a result, snRNA-seq has been shown to be more advantageous, especially when using human brain tissue samples (Grindberg *et al.*, 2013; Hu *et al.*, 2017). Habib *et al*. and others have developed methods that use a single nuclei suspension in place of a single cell suspension (Grindberg *et al.*, 2013; Krishnaswami *et al.*, 2016; Lake *et al.*, 2016; Habib *et al.*, 2017; Hu *et al.*, 2017). The main difference between using single cells and single nuclei is how they are obtained from the tissue. A cell suspension is obtained through tissue digestion using enzymes. It has been shown that enzymatic treatment will favour specific cell types, especially cells that are less complex and introduces aberrant transcription (Lacar *et al.*, 2016). A nuclei suspension is obtained using dounce homogenization on the tissue, followed by cell lysis to release the nuclei and no enzymatic digestion is required (Krishnaswami *et al.*, 2016). This has been shown to be an advantageous method when using human brain tissue because there are no biases towards less complex cells and the neuronal population will not be underrepresented (Lacar *et al.*, 2016). Once a nuclei suspension has been obtained, the same isolation techniques that were mentioned above can be applied to generate snRNA-seq libraries. These methods allow the use of frozen, fresh, or fixed tissue. In this study, we use snRNA-seq instead of scRNA-seq.

Lake *et al.* developed a snRNA-seq pipeline, using postmortem tissue from a 51-year-old sample to perform transcriptome analysis (Lake *et al.*, 2016). This pipeline starts with using neuronal nuclear antigen (NeuN) to isolate neuronal nuclei (instead of single cells) of the cerebral cortex using FACS followed by uses of the Fluidigm C1 platform. They were able to resolve the nuclei, first into inhibitory (mainly interneurons)

and excitatory neurons (mainly projection neurons) and then into 16 distinct subclusters based on the expression of specific marker genes. Ultimately, they were able to identify 8 specific types of excitatory and inhibitory neurons respectively (Lake *et al.*, 2016).

A subsequent study conducted by Lake *et al.* used single-nucleus droplet-based sequencing (snDrop-seq) on human adult brain tissue from the visual cortex, frontal cortex, and cerebellum to characterise the cell types and subtypes in the adult brain (Lake *et al.*, 2018). This study builds on their previous study, where they developed a snRNA-seq pipeline to identify different neuronal subtypes across different cortical areas in the adult brain (Lake *et al.*, 2016). However, they discovered this previous method has several limitations: high cost, the microfluidic chip was limited to 96 cells, and sampling bias because smaller nuclei from glial cells are not captured on microfluidic chips. They realised there was a need for a higher-throughput method and this is why they developed snDrop-seq, adapted from a previously published droplet-based method (Macosko *et al.*, 2015). This improved method allowed them to identify 35 distinct clusters which were region and cell type specific, including: inhibitory and excitatory subtypes, endothelial cells, astrocytes, oligodendrocytes and OPCs (Lake *et al.*, 2018) (**Figure 1.6**). More specifically, they identified 14 types of excitatory neurons, 13 types of inhibitory neurons, 2 distinct astrocyte clusters, 1 microglia cluster, 2 OPC clusters and an oligodendrocyte cluster (Lake *et al.*, 2018) (**Figure 1.6**). This method was able to resolve more excitatory and inhibitory subtypes of neurons compared to their previous study (Lake *et al.*, 2016). Furthermore, both studies also showed that using nuclei instead of cells, they were able to identify all the major cell types of the brain in addition to several neuronal subtypes.

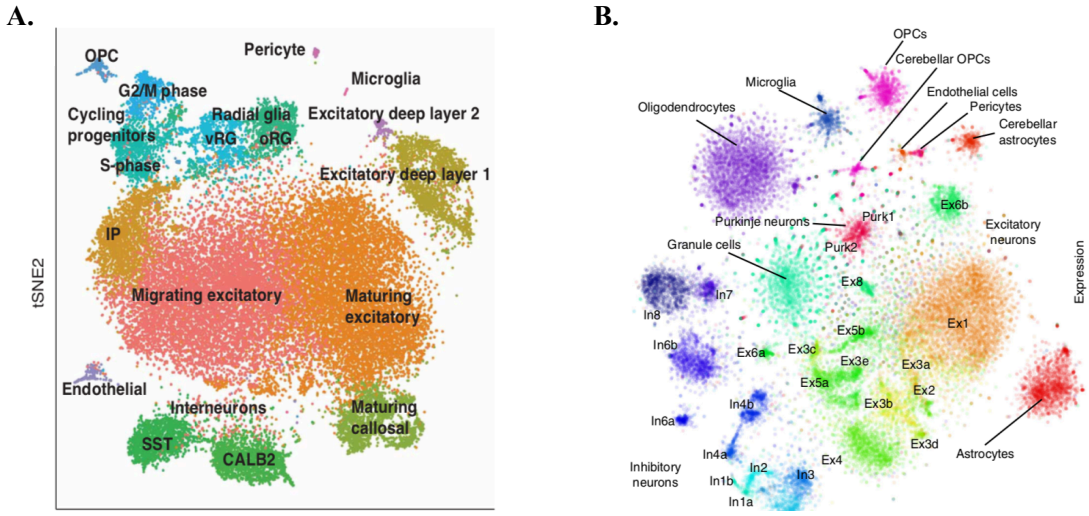


Figure 1.6: ***Classification of major cell types using t- Distributed stochastic neighbour embedding (t-SNE) from fetal and adult human brain tissue.*** **(A)** The cell-type specific clusters identified from fetal brain tissue, after conducting scRNA-seq (taken from Polioudakis *et al.* 2018). **(B)** The region and cell-type specific clusters identified from adult brain tissue, after conducting snDrop-seq (taken from Lake *et al.* 2018). The colours and annotations represent the major cell types and clusters. t-SNE is a non-linear algorithm used to distinguish the distinct cell types (Ofengeim *et al.,* 2017).

Studies conducted by the Allen Brain institute utilised snRNA-seq on post-mortem adult brain tissue to develop a consensus classification of cell types in the human brain (Hodge *et al.*, 2019; Bakken *et al.*, 2021). For example, Hodge *et al*., were able to define 69 neuronal and 6 non-neuronal cell types using the nomenclature they developed for each cluster. More specifically, they defined 45 inhibitory neuron types, 24 excitatory neuron cell types, 1 OPC cell type, 2 astrocyte types, 1 oligodendrocyte type, 1 endothelial type and 1 microglia cell type (Hodge *et al.*, 2019). This ontology was made available through the BioPortal resource (https://bioportal.bioontology.org/ontologies/PCL). Overall, these studies show the power of snRNA-seq to explore the cell atlas of the human brain. However, the majority of snRNA-seq datasets, currently come from adult brain samples. This project will begin to supplement these datasets with pediatric data.

The techniques used to study changes in gene expression are continuously evolving, especially in the last few years. scRNA-seq technologies are now allowing us to build comprehensive cell-type specific GRNs and have allowed us to reveal cell-type specific transcriptomes. In addition to identifying important genes that make up these developmental GRNs, it is also important to identify the CREs such as enhancers that regulate their expression. In turn, this will allow us to build even more accurate GRNs. The most popular methods to identify enhancers are described below including studies that have applied these techniques to identify important enhancers that are involved in neurodevelopment.

1.5 Studying gene regulation in the developing human brain

As already stated above, GRNs are comprised of genes interacting with multiple TFs and CREs which control the development and maintenance of all cell types (MacNeil and Walhout, 2011). Gene expression profiles are produced by a combination of active TFs that interact with a set of *cis*-regulatory modules in the genome (Fiers *et al.*, 2018). These diverse spatiotemporal gene expression patterns determine cell fates and their functions (Carullo and Day, 2019). CREs, such as enhancers, play a key role in facilitating this cell-type specific regulation of gene expression throughout development (John L R Rubenstein, 2011; Buenrostro *et al.*, 2013; Carullo and Day, 2019; Fujiwara *et al.*, 2019). Additionally, our DNA is highly condensed into chromatin. Chromatin can switch or exist as two states: inactive heterochromatin or active euchromatin. How these different states of condensation co-operate with each other plays an important role in gene regulation (Richmond and Davey, 2003). Chromatin regulation in the brain has been shown to be important for transcriptional programmes underlying development, neuronal maturation and plasticity (Su *et al.*, 2017).

1.5.1 DNAse I hypersensitive sites sequencing (DNAse-seq) and Chromatin Immunoprecipitation with massively parallel sequencing (ChIP-seq)

A greater understanding of gene regulation in the human brain has been made possible through the use of genome-wide methods that assay the regulatory regions of chromatin (Thurman *et al.*, 2012; Baek, Goldstein and Hager, 2017). Some of the tools we can use to study gene regulation include DNAse-seq (Boyle *et al.*, 2008; Kundaje *et al.*, 2015). DNaseI hypersensitive sites (DHSs) have been used to identify regulatory regions for many years and it has also been used along with other methods to map regulatory elements of the human genome, such as enhancers, promoters or silencers (Gross and Garrard, 1988; Boyle *et al.*, 2008; ENCODE Project, 2012). DNase-seq works by sequencing regions that are sensitive to cleavage by DNase I, which will theoretically be the nucleosome-depleted DNA (Boyle *et al.*, 2008). Briefly, nuclei are isolated and digested with the DNAse I enzyme, which will cut chromatin at sites where there are nearby specific non-histone proteins. These cut sites are sequenced and then the location of these sites that are 'hypersensitive' to DNase I can be determined, which corresponds to open chromatin (Boyle *et al.*, 2008). Several large studies have been conducted using DNase-seq (along with other methods) to identify CREs such as in enhancers in the human genome (Bernstein *et al.*, 2010; ENCODE Project, 2012; Thurman *et al.*, 2012; Kundaje *et al.*, 2015). This begun with the ENCODE project, a follow up to the Human Genome Project, where they set out to identify and annotate all the functional elements in the human genome (ENCODE Project, 2012; Moore *et al.*, 2020). They profiled DNase sensitivity genome-wide and were able to map 2.89 million unique, non-overlapping DHSs in 125 human cell and tissue types. The generated data was then uploaded onto a public database (https://www.encodeproject.org/) and is continuously being updated and expanded (ENCODE Project, 2012; Moore *et al.*, 2020). To date, there are approximately 245 DNase-seq datasets from human adult or fetal brain tissue and cell lines. In addition, ENCODE has now incorporated and processed their datasets with the data from the Roadmap Epigenomics project (Bernstein *et al.*, 2010). The DNase-seq brain datasets in this project come from nine areas including angular gyrus, anterior caudate, inferior parietal lobule, inferior temporal lobe, medulla, mid frontal gyrus, midbrain, occipital pole, and pons (Kundaje *et al.*, 2015). ENOCODE also includes DNA accessibility and chromatin modification data to generate a categorised list of candidate CREs (cCREs) (Moore *et al.*, 2020).

Another popular tool to identify CREs, ChIP-seq, takes advantage of transcription factor binding or chromatin marks associated with open chromatin to identify the location of putative enhancers and other regulatory elements (Thurman *et al.*, 2012; Carullo and Day, 2019). Firstly, using formaldehyde, the chromatin is crosslinked, followed by sonication to obtain fragments of DNA that are 200-600 base pairs in length. Next, the DNA-protein complex is immunoprecipitated with antibodies. The DNA is then uncross-linked and subjected to several library preparation steps before being sequenced (Ma and

Zhang, 2020). A study conducted in 2012 selected ENCODE ChIP-seq datasets (transcription-related factors (TRF) binding site data) from five cell lines: GM12878, H1-hESC, HeLa-S3, Hep-G2 and K562. They developed machine learning methods to identify genomic features associated with different levels of ChIP-seq signal (Yip *et al.*, 2012). For example, they identified genomic regions proximal to gene promoters or distal to genes that displayed high levels of ChIP-seq signal. Six of these (out of the top fifty predictions) were tested for enhancer activity using a mouse enhancer-reporter assay. From this first round of validation assays, they showed that most of these enhancers were active in tissue associated to neurodevelopment such as the forebrain, midbrain, hindbrain, neural tube and cranial nerve (Yip *et al.*, 2012).

One of the major limitations of DNAse-seq and ChIP-seq is that they both require large number of cells as input (ENCODE Project, 2012). Furthermore, these techniques involve complicated and time-consuming sample or library preparation steps. So, how have researchers been able to improve these techniques?

1.5.2 Assays for Transposase-Accessible Chromatin by sequencing (ATAC-seq)

ATAC-seq has become a popular method for assaying regions of accessible chromatin where CREs are likely to be located due to it being a simple and time-efficient technique, only requiring around 50 000 cells (Buenrostro *et al.*, 2013). This protocol involves nuclei isolation, followed by transposition of DNA using a hyperactive Tn5 transposase (Buenrostro *et al.*, 2016) (**Figure 1.7**). The transposase will cut and insert sequencing adapters into regions of open chromatin which is then amplified and sequenced (Buenrostro *et al.*, 2013) (**Figure 1.7**). By performing ATAC-seq it is possible to identify open regions of chromatin that are accessible to transcription factors and thus identify the location of CREs genome-wide.

A study conducted by Fullard *et al.* performed ATAC-seq on nuclei isolated from adult post-mortem brain tissue. They isolated nuclei from neuronal (NeuN+) and non-neuronal (NeuN-) cells using fluorescence-activated nuclear sorting (FANS) (Fullard *et al.,* 2017). The age of these samples ranges from 64 to 90 years-old and were obtained from one brain region, the frontopolar prefrontal cortex (Fullard *et al.*, 2017). The aim of this study was to identify and characterise open chromatin regions (OCRs) from the human prefrontal cortex. They were able to identify 60,653 differentially accessible OCRs between neurons and non-neurons and of the total OCRs that were identified, 33,054 were neuronal. In addition, Fullard *et al.* carried out functional enrichment analysis by annotating peaks to nearby genes. They found that the neuronal peaks were enriched for the terms such as cellular morphogenesis, cell-cell signalling and synaptic transmission (Fullard *et al.*, 2017).

This is relevant as these terms are all related to neuronal development. For example, neurons undergo unique and complex cellular morphogenesis, which involves the coordinated assembly of functionally distinct axons and dendrites (Poulain and Sobel, 2010).

A subsequent study conducted by Fullard *et al.* in 2018 used ATAC-seq on FANS-sorted neuronal and non-neuronal nuclei from adult post-mortem brain tissue across 14 brain regions to identify both cell type-specific and region specific OCRs (Fullard *et al.*, 2018). To analyse regional differences and cell-type specific differences, a consensus set of OCRs were generated by combining all peaks called in the different brain regions and individual cells. They then determined how many reads overlapped each other, performed t-SNE clustering using the read counts and this result showed a clear division between the neuronal and non-neuronal samples (Fullard *et al.*, 2018). More specifically, open chromatin regions in neurons showed increased regional variability and these OCRs were further away from TSSs compared to OCRs in non-neuronal populations. Finally, using the open chromatin patterns, they performed TF footprinting analysis to infer downstream gene regulation and expression. Ultimately they found that OCRs mainly overlapped with genes expressed in the same brain region, highlighting the cell type- and region-specificity of these OCRs (Fullard *et al.*, 2018). This study supplemented their previous study conducted in 2017 as that study only obtained samples from one brain region. Both studies provide evidence that generating and using an open chromatin atlas can provide insights into regulation of gene expression in the brain.

While Fullard *et al.* focused on adult brain tissue, de la Torre-Ubieta *et al.* performed ATAC-seq and RNA-seq on embryonic human cortical brain tissue to gain a better understanding of how gene expression is regulated during human cortical neurogenesis (Torre-Ubieta *et al.*, 2018). In addition, they also performed Hi-C, a method used to measure chromatin interaction between two regions. By integrating the ATAC-seq data, RNA-seq data and the Hi-C data, they were able to identify CREs, such as the enhancers for the genes *FGFR2* and *EOMES*, that are known to be involved in cortical neurogenesis (Torre-Ubieta *et al.*, 2018). To confirm the activity of these enhancers, they generated partial and complete deletions of the *EOMES* and *FGFR2* enhancers in primary human neural progenitors (phNPCs). These results showed that for both sets of deletions, there was a reduction in the expression of the two genes (Torre-Ubieta *et al.*, 2018). Even though this study is extremely important they have only included one developmental stage in their analysis, and it would be very useful to expand this and include additional stages such as juvenile and adult developmental stages.

Even though ATAC-seq has mainly been used on fresh tissue, it has also been used on motor neurons that have been slow-frozen, derived from human embryonic stem cells (Milani *et al.*, 2016) and hematopoietic B cells (Scharer *et al.*, 2016). In addition, an adapted ATAC-seq protocol (Omni-ATAC) was developed for human brain tissue and thyroid cancer tissue that has been flash-frozen (Corces et al., 2017).

Lastly, another study compared ATAC-seq results obtained when using fresh, flash-frozen or cryopreserved breast cancer cells and mouse mammary tissues (Fujiwara *et al.*, 2019). They also used and modified three ATAC-seq protocols: the original protocol (Buenrostro *et al.*, 2013), Omni-ATAC protocol (Corces *et al.*, 2017) and the Takaku-ATAC protocol (Takaku *et al.*, 2016). Interestingly these results showed that ATAC-seq can generate data of high quality from cells that were stored by cryopreservation, and these results are equivalent to results obtained from fresh cells or tissue (Fujiwara *et al.*, 2019). In this project, I sought to optimise the ATAC-seq technique for use with human brain tissue from the sample preparation to the different bioinformatic pipelines. In addition, I also determined if the type of sample tissue used (either fresh, frozen, or cryopreserved) affected the quality of the data.

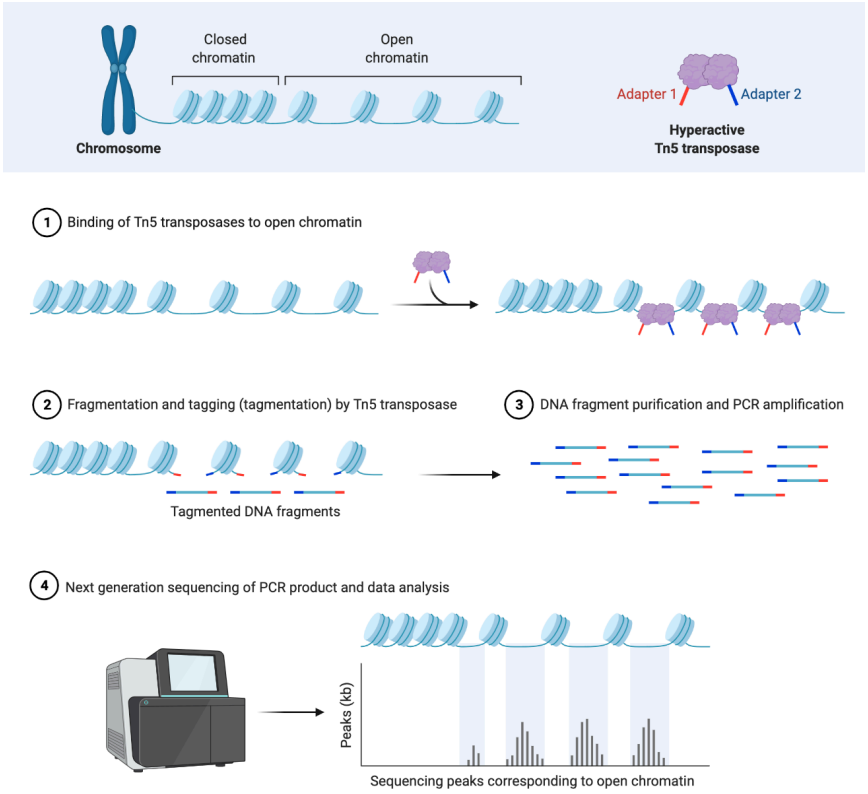


Figure 1.7: ***Schematic overview showing the steps of the ATAC-seq reaction.*** **(1)** Tn5 transposases bind to regions of open chromatin. **(2)** Tn5 transposase cuts and tags the DNA fragments with sequencing adapters. **(3)** The DNA fragments are then purified and amplified. **(4)** Finally, the amplified product is sequenced, and the sequencing data is analysed through mapping to target genome. (Made in Biorender).

1.6 Diseases and disorders that affect the human brain

1.6.1 Disorders and diseases that manifest during different developmental stages

What happens when gene expression is dysregulated, and gene regulatory networks are disrupted? Transcriptional dysregulation is thought to play a major role in numerous disorders of the human brain (Luthi-Carter and Cha, 2003; Li *et al.*, 2018; Meng and Mei, 2019). In addition, ATAC-seq and sn/scRNA-seq are now being applied to determine these regulatory changes (Polioudakis *et al.*, 2019; Pfisterer *et al.*, 2020; Boulting *et al.*, 2021). There are many different infectious diseases and neurological disorders that affect the brain. Epilepsy and Tuberculosis meningitis (TBM) are such diseases or neurological disorders that present a significant burden in Southern Africa.

TBM is a type of *Mycobacterium tuberculosis* (TB) infection of the central nervous system (CNS) and there are several types of CNS TB infections that have different presentations in the brain (Chatterjee, 2011). In general, CNS TB begins with the formation of small tuberculosis foci in the brain, meninges or spinal cord (Rich and McCordock, 1944). Where these foci are located and the ability to control them, determines which form of CNS tuberculosis occurs (Rich and McCordock, 1944). Symptoms in children with TBM include fever, seizures, and neck stiffness. Depending on the stage of the presentation they can also present with neurological symptoms such as lethargy and agitation to coma (Cherian and Thomas, 2011). TBM is considered to be the most lethal extrapulmonary form of TB and the second most common form in children (Starke, 1999; Chiang *et al.*, 2014). Another form of CNS TB in children are tuberculomas, which are known as space occupying lesions, but this presentation is very rare compared to TBM (Chatterjee, 2011). TBM is associated with high mortality rates in children and individuals with an HIV-1 co-infection. Unfortunately, the global prevalence of TBM remains poorly understood. This is because it is rarely reported and the diagnosis of TBM is not always microbiologically confirmed but several reviews have estimated that the global burden of TBM is approximately 100 000 cases per year (Wilkinson *et al.*, 2017). The Western Cape province has the highest incidence of TB in South Africa. More specifically, the incidence rate of TBM in the Western Cape is age-specific and ranges from 31.5 per 100 000 (<1 year) to 0.7 per 100 000 (10 – 14 years) (Van Well *et al.*, 2009). This shows that studying gene expression and gene regulation is very important, making this project very relevant. Pediatric transcriptomic data from the brain is needed to help us obtain a better understanding of TBM especially because of the high incidence of the disease in this population. By using advanced technologies such as ATAC-seq and sc/snRNA-seq, genes which are differentially expressed in brain tissues of TBM patients could be identified. This would provide

a better understanding of the molecular component of the infection and identify potential biomarkers (Cai *et al.*, 2020; Guo *et al.*, 2022).

Epilepsy is a neurological disorder that is described by spontaneous and periodic seizures. It has been shown that these seizures are generated in the cerebral cortex, with the temporal and frontal cortex having been found to be common sites of epilepsy (Avanzini and Franceschetti, 2003). Approximately 70 million people around the world are diagnosed with epilepsy and more importantly it is one of the most common disorders in childhood (Shinnar and Pellock, 2002). A study conducted in 2019 evaluated the burden of epilepsy in 195 countries from 1990 to 2016 (Beghi *et al.*, 2019). Interestingly in 2016, regions in southern sub-Saharan Africa showed the highest prevalence. In addition, they found that prevalence increased with age, with specific peaks at ages 5 to 9 and 80 years and above (Beghi *et al.*, 2019) (**Figure 1.8**). A subsequent study conducted by Owolabi *et al.* used available data to determine the prevalence of epilepsy in sub-Saharan Africa (Owolabi *et al.*, 2020). They determined the overall prevalence of active epilepsy (i.e. requiring treatment or ongoing seizures) to be 9 per 1 000 persons and 16 per 1 000 persons for lifetime epilepsy (Owolabi *et al.*, 2020). Despite the increase in the incidence of epilepsy, not much is known about how it affects the brain at the level of gene expression. This is important because if we can identify epilepsy-associated genes and cell types, this would help us further understand the mechanisms underlying epileptogenesis. With the development of cutting-edge technologies such as scRNA-seq, there has been an increase in our understanding of the pathophysiology of epilepsy nevertheless it is still an ongoing task (Wang *et al.*, 2017).

Using snRNA-seq on human brain temporal cortex tissue from epileptic and non-epileptic subjects, Pfisterer *et al.* identified specific neuronal subtypes from principal neurons, (layer L5-6_Fezf2 and L2-3_CUX2) and GABAergic interneurons, (Somatostatin (*Sst*) and Parvalbumin (*Pvalb*)), that showed significant transcriptomic changes compared to other cell subtypes in the same families (Pfisterer *et al.*, 2020). Their results showed that epilepsy is defined by the dysregulation of thousands of genes and these genes were up- or downregulated in specific cell types. 6, 900 differentially expressed (DE) genes were identified for GABAergic neurons and 13, 700 DE genes were identified for principal neurons (Pfisterer *et al.*, 2020). As mentioned above, several layer L5-6 principal neuron subtypes showed significant transcriptomic changes, specifically, dysregulation in gene expression for several glutamate receptor- and action potential-associated proteins. For example, their results showed an increase in the expression of Glutamate Ionotropic Receptor AMPA Type Subunit 1*(GRIA1* or *GLUR1*) and a decrease in *GRIA2* (or GLUR2) which are both part of the AMPA receptor complex (Pfisterer *et al.*, 2020). This study was particularly interesting because it showed that these affected neuronal subtypes could be grouped according to their shared epilepsy-related transcriptional changes which shows that they could all belong

to a common network that underlies epileptogenesis (Pfisterer *et al.*, 2020).

In addition to mutations in the coding regions of genes, non-coding regions of the genome are also implicated and shown to be involved in brain disorders. For example, enhancer dysregulation has been involved in several neurodegenerative and psychiatric diseases such as Schizophrenia (SCZ) and Alzheimer's disease (AD) (Fullard *et al.*, 2017; Meng and Mei, 2019). Genome wide association studies (GWAS) and other studies have identified SNPs that are strongly associated to SCZ and these SNPs fall into non-coding regions with enhancer activity (Ripke *et al.*, 2013; Li *et al.*, 2018). As mentioned above, Fullard *et al.* used ATAC-seq to identify thousands of OCRs that were differentially accessible between neurons and non-neurons from adult frozen post-mortem brain tissue (Fullard *et al.*, 2017). TF footprinting analysis was also performed to identify OCRs and TF-binding sites that overlap with known SCZ risk loci. One of the risk variants identified, rs10750450, was a SNP proximal to the gene encoding sorting nexin 19 (*SNX19*). Functional validation analysis confirmed that this SNP does lead to an increase in transcriptional activity. The gene expression levels of *SNX19* has previously been linked to SCZ (Zhu *et al.*, 2016), so Fullard *et al.* has supplemented this by identifying the putative functional regulatory region for this gene, and showing that an increase in gene expression of *SNX19* increases the risk of SCZ (Fullard *et al.*, 2017).

A study conducted by Li *et al.,* performed several experiments to allow for the analysis of multiple genomic data modalities to gain a better understanding of the causes of neuropsychiatric disorders (Li *et al.*, 2018). Firstly, they generated bulk, scRNA-seq and snRNA-seq data from human post-mortem tissue. The age of the samples ranged from 5 PCW to 64 postnatal years (PY). Epigenomic data was generated using ChIP-seq for three histone marks: H3K4me3, H3K27me3 and H3K27ac. This data was generated from samples from mid-fetal, infant and adult brain tissue (Li *et al.*, 2018). Additionally, they collected GWAS data related to neuropsychiatric disorders or personality traits such as SCZ, AD, autism spectrum disorder (ASD) and several others. The GWAS data can be used to identify genomic loci associated with risk for these particular disorders. Lastly, they obtained Hi-C data from adult and fetal brain tissue samples which is used to measure chromatin interaction. They then integrated the GWAS, Hi-C and ChIP-seq data and converged them on 10 disease-associated gene modules where they were able to identify genes associated with disease risk (Li *et al.*, 2018). Of particular interest was module ME37, consisting of 145 genes that were enriched for fetal enhancers. In addition, these enhancers were enriched in neurons and not neural progenitors or glia. They also identified several genes in this module such as *SATB2*, *MEF2C*, *ZNF184* and *TCF4* that were associated with different traits and disorders such as SCZ, ADHD, ASD and major depressive disorder (MDD). More specifically, myocyte enhancer factor 2C (*MEF2C*) has been shown to control activity-dependent expression of neuronal genes and in particular those related to synapse function and ASD

(Parikshak *et al.*, 2013). The progression in genetic technologies has now made it possible for us to identify genes associated with hundreds of diseases and neurological conditions. This will not only help us to further understand the mechanism of the disease but also improve diagnosis and treatment.

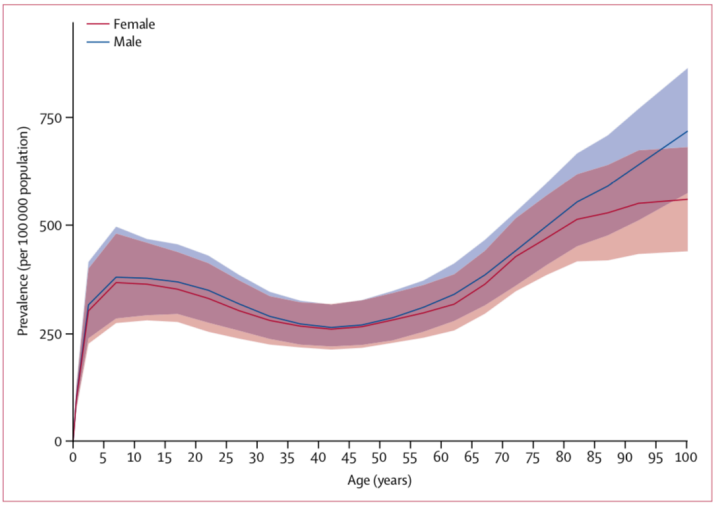


Figure 1.8: ***The global prevalence of epilepsy according to age and sex.*** Burden of epilepsy was assessed in 195 countries from 1990 to 2016. The burden was measured as prevalence (y-axis), by age (years) and grouped according to sex (x-axis). (Taken from Beghi *et al.* 2019).

1.7 Toward understanding the dynamics of gene regulatory networks during brain development

In this review, I have highlighted numerous techniques that can be applied to study the human brain. Sc/snRNA-seq and ATAC-seq have now become popular methods to use to understand how the brain develops over time. However, there are still disadvantages to these approaches and very few benchmark studies, comparing different approaches, have been conducted to help researchers decide which technique would be best suited to their study. Additionally, it is extremely important that these technologies are established and made accessible to researchers in South Africa. Therefore, in this project, I have sought to optimise the sample preparation and isolation techniques which will allow us to train and assist researchers wanting to use snRNA-seq and ATAC-seq to study the human brain. Ultimately, this will allow us to boost cutting edge research such as neuroscience in South Africa, which, for a long time, has been dominated by the Global North.

This study was facilitated by our collaboration with neurosurgeons at Red Cross War Memorial Children's Hospital and Mediclinic Constantiaberg. This has allowed us to generate a biobank of brain tissue obtained from samples that are removed during neurosurgical procedures to treat conditions such as epilepsy. Through generating snRNA-seq and ATAC-seq data from these pediatric and adult tissue samples, I have been able to identify all the major cell types of the brain, genes that are being expressed by these specific cell types and build a dataset of putative enhancers or promoters that may be regulating their expression. With the use of our unique resources and by performing optimisation experiments, I have addressed many of the challenges mentioned previously and provide new datasets to supplement what we currently know about normal gene expression and regulation during brain maturation. Furthermore, these datasets can be compared to adult datasets in the future to show the dynamics of the gene regulatory networks that guide brain function over the course of the human lifespan. Once we determine and characterise normal gene expression and regulation profiles during brain maturation, we can begin to understand how these gene expression profiles change when the brain is affected by an infectious disease or neurological disorder.

1.8 Aim and objectives

Aim:

This project aims to provide a better understanding of how the brain cell atlas changes over time by contributing to the current brain cell atlas with pediatric single cell data. This will be achieved by utilising cutting-edge single cell technology (snRNA-seq) to identify the major cell types in pediatric brain tissue samples and identify genes being expressed by these cell types. A dataset of putative enhancers and promoters that may be regulating their expression, will also be generated using the latest tools for assessing chromatin dynamics (ATAC-seq).

Objectives of the project:

1. Optimisation of the nuclei isolation protocol and generation of snRNA-seq libraries for pediatric brain tissue samples from the frontal and temporal lobes using the 10x Genomics platform

2. Establishment of a bioinformatics pipeline to analyse the snRNA-seq libraries generated in Objective 1 in order to identify the major brain cell types and conduct a pilot differential expression analysis study between snRNA-seq libraries from the frontal and temporal lobe

3. Generation of bulk ATAC-seq libraries from pediatric and adult brain tissue samples to assess chromatin dynamics over the course of brain maturation

4. Testing of bioinformatic pipelines to analyse the bulk ATAC-seq data generated in Objective 3 in order to generate a consensus list of cis regulatory elements, including promoters and enhancers, and to identify those that are being dynamically used over the course of brain maturation

Chapter 2: Materials and methods

2.1 Sample collection and storage

Pediatric brain tissue samples were obtained from consenting patients during neurosurgical procedures at Red Cross Children's hospital. Adult brain tissue samples were obtained from neurosurgeons at Constaniaberg Mediclinic. This was done through a collaboration between the Neuroscience Institute and the division of Neurosurgery of the University of Cape Town. Tissue fragments were obtained during surgery to remove epileptic regions. The surgeons removed the access tissue to obtain the epileptic tissue and this access tissue likely represents normal brain tissue which was collected for this study. The majority of these tissue fragments were then transferred to the lab in ice-cold artificial CSF for sample preparation or slow frozen in −80 ℃ immediately after surgery. The information about each sample (adult and pediatric) is detailed in **Table 2.1**. Ethics was granted for the use of pediatric and adult human brain tissue by the Human Research Ethics Committee (HREC REF numbers: 016/2018 and 145/2022).

Table 2.1: Patient data including sex, age, and diagnosis. The brain region, number of technical replicates, type of replicate, type of experiment performed, and lysis buffer used for each experiment are also given.

sex	*Age (year. months)*	*brain region*	*diagnosis*	*number of replicates*	*fresh replicate*	*frozen replicate*	*cryopreserved replicate*	*type of experiment performed*	*lysis buffer used*	*size of cell strainer used*
male	14*	frontal cortex	Right frontal refractory epilepsy	2	0	2	0	snRNA-seq	EZ lysis buffer	35
female	15**	temporal cortex	Left temporal lesion and epilepsy	4	0	4	0	snRNA-seq	PURE lysis buffer	40
male	1.11	temporal cortex	Right hemimegaloencephaly and intractable seizures	3	2	0	1	ATAC-seq	PURE lysis buffer	35
female	1.7	temporal cortex	Left frontal cortical dysplasia with Epilepsy	3	2	0	1	ATAC-seq	EZ lysis buffer	35
female	5	temporal cortex	Sturge Weber syndrome	2	2	0	0	ATAC-seq	PURE lysis buffer	35
male	6	frontal cortex	Left frontal dysplasia with epilepsy	2	2	0	0	ATAC-seq	PURE & EZ lysis buffer	35
female	9	temporal cortex	Right sided Rasmussen's with refractory EPC	2	2	0	0	ATAC-seq	EZ lysis buffer	35
male	14*	frontal cortex	Right frontal refractory epilepsy	1	0	0	0	ATAC-seq	PURE lysis buffer	40
female	15**	temporal cortex	Left temporal lesion and epilepsy	2	1	0	1	ATAC-seq	PURE lysis buffer	40
male	23	temporal cortex	Temporal lobe epilepsy; previous cavernous malformation	1	1	0	0	ATAC-seq	PURE lysis buffer	40
male	31	temporal cortex	Temporal lobe epilepsy	2	2	0	0	ATAC-seq	PURE lysis buffer	40
male	46	frontal cortex	Frontal lobe epilepsy	2	2	0	0	ATAC-seq	PURE lysis buffer	40

*, ** These tissue samples come from the same patient, respectively

2.2 Single nucleus RNA-sequencing (snRNA-seq)

SnRNA-seq is used to perform transcriptome-wide analysis at the single nucleus level. Briefly, the brain tissue from each sample is collected and flash frozen. Nuclei suspensions are prepared from the frozen brain tissue using the nuclei isolation protocol (**Figure 2.1**). The nuclei suspensions are then used as input for the chromium controller to generate single nuclei droplets. snRNA-seq libraries are then generated, sequenced, and analysed using Cell Ranger and other bioinformatics tools (**Figure 2.1**).

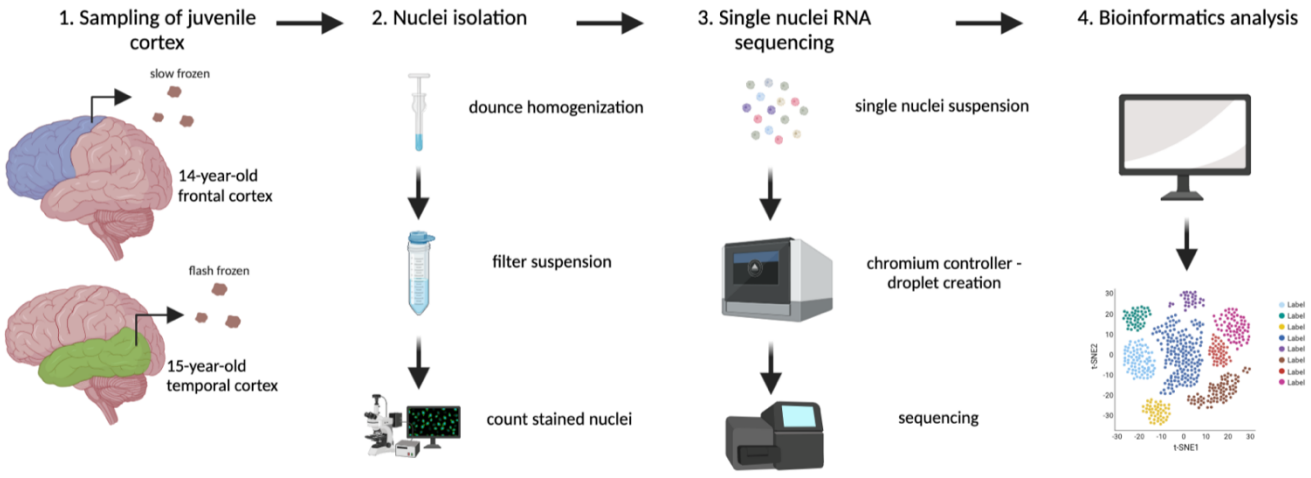


Figure 2.1: ***Schematic overview of experimental workflow for snRNA-seq of human temporal and frontal cortex.*** **(1)** Tissue fragments were collected and prepared from each sample. **(2)** These frozen tissue fragments were used as starting material for the nuclei isolation experiment. **(3)** A single nuclei suspension was generated and used as the input material for the 10x Genomics chromium controller. **(4)** The raw sequencing data was processed and analysed using 10x Genomics Cell Ranger and other bioinformatics tools. (Made in biorender).

2.2.1 Nuclei isolation from human brain tissue

snRNA-seq was performed on 14-year-old frontal cortex tissue and 15-year-old temporal cortex tissue in two separate sessions (see Table 2.1). Two fragments of 14-year-old frontal cortex tissue and one fragment of 15-year-old temporal cortex tissue were used. The 14-year-old tissue fragments had been slow frozen and then transferred to a -80 ℃ freezer (**Figure 2.2**). The 15-year-old tissue had been flash frozen in liquid nitrogen after transferal to the lab in ice-cold artificial CSF and stored in the -80 ℃ freezer. Frozen tissue fragments were disrupted to form a single cell suspension and the nuclei isolated from the cells using an adapted method from protocols described previously (10x, 2017; Habib *et al.*, 2017) (**Figure 2.2**). Tissue samples were placed in a glass dounce homogeniser (KIMBLE Dounce tissue grinder set, Sigma-Aldrich, D8938) (on ice) containing 2 ml of ice-cold Nuclei EZ Lysis Buffer (Sigma-Aldrich, NUC101) or Nuclei PURE Lysis Buffer (Sigma-Aldrich, NUC201) (**Table 2.1**). The tissue was ground 20-25 times with pestle A followed by 20-25 times with pestle B until there were no visible tissue pieces. The ground up tissue was transferred to a 15 ml conical tube. 2 ml of EZ lysis buffer or PURE lysis buffer was added to the tissue and incubated on ice for 5 minutes. The samples were then centrifuged at 500 $x\ g$ for 5 minutes at 4 ℃ (**Figure 2.2**). The supernatant was discarded, and the pellet resuspended in another 2 ml of EZ lysis buffer or PURE lysis buffer. The nuclei suspension was incubated for 5 minutes and centrifuged. The supernatant was discarded, and the pellet resuspended in 2 ml of ice-cold Nuclei Wash & Resuspension buffer (NSB) (1 X PBS with 0.01 % BSA and 0.1 % RNase inhibitor). A cell strainer was placed onto a 50 ml tube and the resuspended nuclei was passed through the strainer and centrifuged at 500 $x\ g$ for 5 minutes at 4 ºC. This was done to remove any cell debris or large clumps. 10 μl of the filtered nuclei suspension was stained with Hoechst, loaded onto a hemocytometer, and then counted under an upright microscope (Zeiss) (**Figure 2.2**).

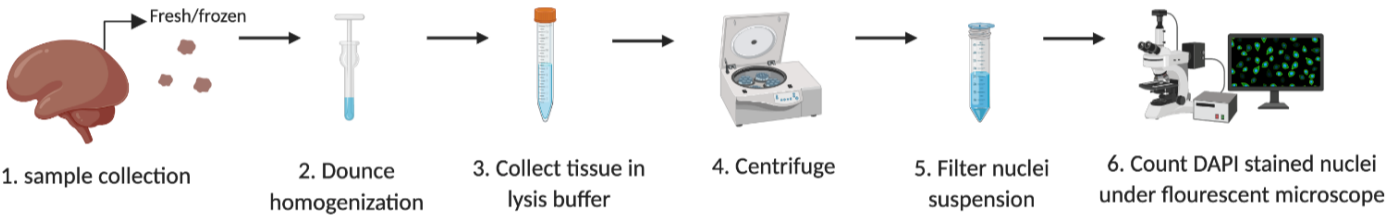


Figure 2.2: ***Detailed overview of the nuclei isolation procedure for snRNA-seq and ATAC-seq.*** **(1)** Brain tissue samples were collected, flash frozen, or used fresh. **(2,3)** Tissue fragments were homogenised using a dounce homogeniser and cells were lysed to release the nuclei. **(4,5)** The nuclei were centrifuged and then passed through a filter to remove cell debris. **(6)** Finally, an aliquot of nuclei was stained and counted under the microscope. (Made in biorender).

2.2.2 Isolating nuclei into Gel beads-in-emulsion (GEMs) and mRNA barcoding to generate snRNA-seq libraries

The 14-year-old and 15-year-old samples were run in two separate sessions on the 10x Genomics Chromium Controller. 10x Genomics recommends using 1 000 nuclei/μl as the input concentration and each sample was diluted accordingly. An aliquot of each nuclei suspension generated for the two 14-year-old tissue fragments at $1\,000\ \text{nuclei}/\mu l$ was used to generate the master mix to load onto the chromium chip. For the 15-year-old sample, only one tissue fragment was used and four aliquots of the resulting nuclei suspension at $1\,000\ \text{nuclei}/\mu l$ were used to generate the master mix to load onto the chromium chip. The 10x Genomics single cell protocol was used to isolate individual nuclei and generate single nuclei RNA-seq libraries using the Chromium Single Cell 3' GEM kit v3 kit for the 14-year-old sample (10x Genomics, PN-1000075) and Chromium Single Cell 3' GEM v3.1 kit for the 15-year-old sample (10x Genomics, PN-1000128). The aim was to target and sequence at least $10\,000$ nuclei per replicate, so the samples were diluted according to the target cell recovery table in the respective version of the 10x Genomics protocol (**Table 2.2**). For each sample, the master mix (reverse transcription [RT] reagent, template switch oligo, reducing agent B and RT enzyme C) and isolated nuclei were combined with 10x barcoded Single Cell 3' v3 Gel beads and partitioning oil to form Gel beads in emulsion (GEMs), through the use of a chromium chip which was loaded into the Chromium Controller (**Figure 2.3**). The gel beads are coated with thousands of oligonucleotides which consist of a poly(dT) primer sequence, a 10x Barcode and a unique molecular identifier (UMI). The poly(dT) primer captures the poly-adenylated mRNA (**Figure 2.3**) and during the GEM incubation barcoded cDNA is produced from the poly-adenylated mRNA released from the nuclei (**Figure 2.3**). Each nucleus is given a unique 10x Barcode, so all generated cDNA molecules share a common 10x Barcode and each transcript molecule is given a unique UMI.

Table 2.2: Master mix and nuclei suspension preparation. The volume of nuclei (1000 nuclei/μl), nuclease free water and master mix used to prepare each technical replicate solution that was loaded onto the chromium chip according to the specifications of the indicated version of the Chromium Single Cell 3' GEM kit.

Sample	**Number of technical replicates**	**Nuclei stock volume (μl)**	**Nuclease free water (μl)**	**Volume of master mix (μl)**	**10x Genomics kit used (version)**
14-year-old	2	16.0	30.6	33.4	v3
15-year-old	4	16.5	26.7	31.8	v3.1

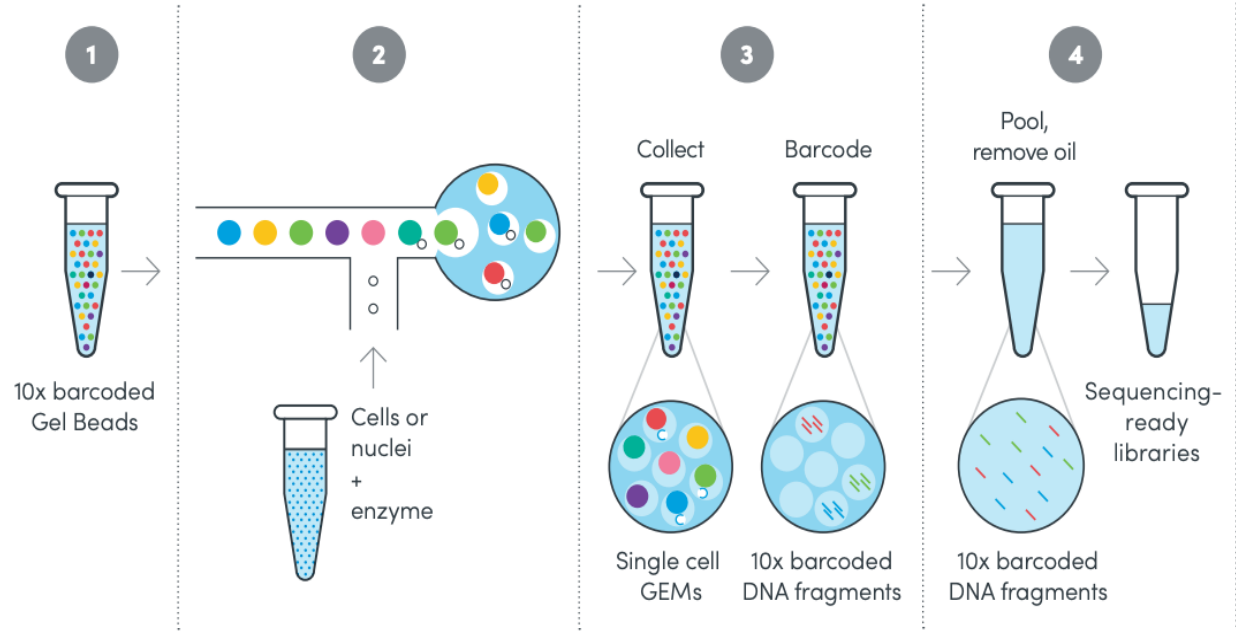


Figure 2.3: ***Overview of Gel beads in emulsion (GEMs) generation and barcoding***. Schematic diagram showing the steps of generating GEMs in the Chromium Chip using the 10x microfluidic device. **(1)** The pool of Gel beads were coated with thousands of oligonucleotides which each consist of a 10X Barcode, UMI and poly(dT) primer sequence. **(2)** The cell or nuclei suspensions were loaded onto the chromium chip, combined with 10x barcoded gel beads and partitioning oil to form thousands of GEMs. **(3)** During GEM incubation, the Gel bead was dissolved, the poly(dT) primer captured the poly-adenylated mRNA and barcoded cDNA was produced. **(4)** Finally, the barcoded fragments were pooled, and sequencing libraries prepared which were then sent for sequencing. (Adapted from: https://www.10xgenomics.com/technology and https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_BR025_Chromium-Brochure_Letter_Digital.pdf).

2.2.3 GEM-RT incubation and cDNA amplification

After GEM-RT incubation, the barcoded cDNA was amplified via polymerase chain reaction (PCR) through the use of a thermal cycler (BioRad, 1851197). The number of cycles used is dependent on the targeted nuclei recovery. In order to target 10 000 nuclei, 11 cycles were used during the cDNA amplification step (10x, 2019). To check the quality (i.e., cDNA fragment size distribution) and quantity of the cDNA library, 1 μl of each sample was run on an Agilent Bioanalyzer High Sensitivity chip on the Agilent 2000 Bioanalyzer instrument (Agilent, G293BA) by the Central Analytical Facility (CAF), (Stellenbosch University).

2.2.4 snRNA library preparation and sequencing

The snRNA-seq libraries were prepared by adding unique sample index primers, which allow sample multiplexing and add the necessary sequences for Illumina bridge amplification during sequencing. The concentration of each cDNA library (determined using the bioanalyzer readings from the previous step) was used to identify the number of cycles required during the sample index PCR (10x, 2019). The concentration of the DNA libraries was between 25-150 ng, and therefore 13 cycles were used. To check the quality (average fragment size) of the snRNA libraries, 1 μl of each sample was run on the Agilent Bioanalyzer High Sensitivity chip on the Agilent 2000 Bioanalyzer instrument (Agilent, G293BA) by CAF (Stellenbosch University) or the Centre for Proteomic and Genomic Research (CPGR) (Cape Town). Library concentration was determined using Qubit (Thermo Fisher Scientific) by CAF (Stellenbosch University) or CPGR (Cape Town). The 14-year-old snRNA-seq libraries were sequenced by the CPGR (Cape Town) on the Illumina NextSeq 500 platform (Illumina, SY-415-1002) using the Illumina NextSeq High Output v2.5 kit (150 cycles) and using the sequencing conditions from the 10x Genomics protocol (**Table 2.3)**. The 15-year-old snRNA libraries were sequenced by Novogene (Singapore) on the NovaSeq High Output v2.5 kit (150 cycles) using standard paired-end sequencing conditions. Sequencing the libraries generated a standard BCL file output folder.

Table 2.3: 3' Gene Expression library run parameters used for the 14-year-old sample. The recommended number of cycles to run for each sequencing read.

Sequencing Read	Recommended number of cycles
Read 1	28
i7 index	8
i5 index	0
Read 2	91

2.2.5 snRNA-seq analysis

The bioinformatics pipeline that was used to analyse the snRNA-seq libraries is summarised in **Figure 2.4** and elaborated on in subsequent sections. Briefly, after library generation and sequencing, the data files

were processed to generate fastq files. Filtered feature-barcode count matrices were generated using Cell Ranger, which was used as input for the *Seurat* pipeline. Quality control, doublet removal and other pre-processing steps were performed on the samples before integration. Clustering and cell-type annotation was performed, followed by additional downstream analyses such as differential expression and functional enrichment analysis (**Figure 2.4**).

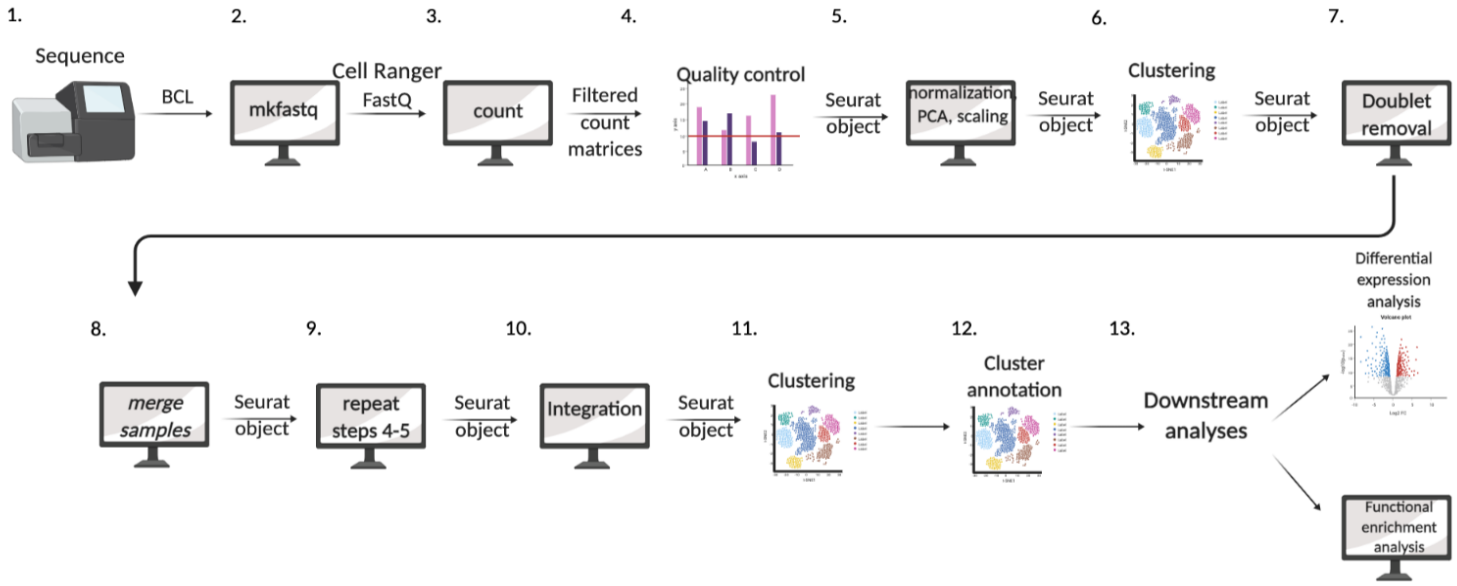


Figure 2.4: ***Schematic of bioinformatics workflow for snRNA-seq of temporal and frontal cortex.*** **(1-2)** Raw sequencing data was processed to produce fastq files which represents the start of the workflow. **(3)** The fastq files were used as input to produce filtered count matrices for each sample. **(4, 5, 6)** In *Seurat*, quality control and other pre-processing steps such as clustering were performed on each technical replicate for each sample separately. **(7)** A doublet removal step was performed on each technical replicate. **(8, 9)** All of the samples were then merged before undergoing another round of quality control and pre-processing. **(10, 11)** The merged *Seurat* object was integrated, and clustering analysis performed. **(12)** Marker genes were identified, and the clusters annotated. **(13)** Lastly, downstream analysis steps were performed such as differential gene expression analysis. (Made in biorender).

Building a custom reference package and raw data processing

Raw base call (BCL) files were received from the sequencing facilities and Cell Ranger (v3.1, 10x Genomics) software was used, specifically the *cellranger mkfastq* and *cellranger count* commands, to process the snRNA-seq data by aligning the sequences to the human genome assembly GRCh38 and quantify gene expression (10x, 2018) (**Figure 2.4: steps 1-2**). These commands were carried out using 10x Genomics online tutorials as guides, detailed in the section below. cDNA libraries were generated from a single nuclei suspension rather than a single cell suspension which, resulted in the capturing of unspliced precursor mRNA

(pre-mRNA) and mature mRNA. However, the Cell Ranger software only used the reads aligned to the exons. To ensure that all the intronic reads from the pre-mRNA are also captured, a custom "pre-mRNA" reference was generated using the cellranger *mkref* command (cellranger mkref --genome=path/to/ GRCh38-3.0.0_premrna --fasta=path/to/refdata-cellranger-GRCh38-3.0.0/fasta/genome.fa --genes=/path/to/GRCh38-3.0.0.premrna.gtf) (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/references#premrna).

Cell Ranger's (v3.1) *mkfastq* command was used to demultiplex and convert BCL files to FASTQ files (cellranger mkfastq --id= sampleID --run=/path/to/BCL/files --csv=path/to/sample_sheet.csv) (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/mkfastq). Using the FASTQ files as input, the *count* command performs alignment, filtering, barcode counting and UMI counting. The alignment was carried out using the custom reference data (pre-mRNA) generated from the hg38 reference genome. The barcodes are used to produce the feature-barcode matrices: an unfiltered feature-barcode matrix and a filtered feature-barcode matrix (cellranger count --id=sampleID --transcriptome=/path/to/refdata --fastqs=/path/to/fastq/outs/folder --sample=sampleID --=expect-cells=1000). Each element of the matrix is the number of UMIs associated with a feature (row) and a barcode (column) (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/count) (**Figure 2.4: step 3)**. The full script used for this step can be found in **Supplementary file 1.**

Pre-processing of snRNA-seq data

The *Seurat* (v4) (Butler *et al.*, 2018; Stuart *et al.*, 2018) pipeline in R studio (v4.1) was used to analyse the snRNA-seq gene expression matrices generated by Cell Ranger using the HBC training scRNA-seq workshop (https://github.com/hbctraining/scRNA-seq) and the *Seurat* tutorials (https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html) as guides, as detailed in the following sections. This pipeline allows for the identification of cell subtypes and clusters (Stuart *et al.*, 2018).

Quality control

Quality control metrics and plots were generated to visually explore the data. The quality control steps were performed on each technical replicate separately and then on the merged *Seurat* object as described in **Figure 2.4 (see steps 4 and 9**). The *count* data was read in using the *Read10X* function and a *Seurat* object was created using the *CreateSeuratObject* function (Technical_replicate <- CreateSeuratObject(counts =

T1.data, project = "T1", min.cells = 5, min.features = 100)). Nuclei that expressed a minimum of 100 genes were kept. The quality control metrics were analysed to determine which nuclei were of low quality and they were removed from the dataset. The quality control metrics assessed include cell counts, UMI counts per nuclei, UMI vs. genes detected and the proportion of reads that map to the mitochondrial genome indicative of low-quality nuclei. After visualising the QC metrics, nuclei with a UMI count lower than 500 were filtered out and nuclei that expressed a minimum of 500 genes detected were kept. Nuclei that had an overall complexity less than 0.85 were removed from the final dataset and low-quality nuclei that showed a mitochondrial count ratio greater than 0.15 were also removed (filtered_seurat <- subset(x = merged_sample, subset= (nUMI >= 500) & (nGene >= 500) & (log10GenesPerUMI > 0.85) & (mitoRatio < 0.15))). Next, genes that had zero expression in all nuclei were removed and the dataset was filtered to keep genes which were expressed in 10 or more cells (nonzero <- counts > 0, keep_genes <- Matrix::rowSums(nonzero) >= 10, filtered_counts <- counts[keep_genes,]) (**Figure 2.4: step 4**) (**Supplementary file 2**).

Normalisation

The cell cycle score was checked before performing the *sctransform* normalisation method to determine if cell cycle is a major source of variation in the dataset. The nuclei were assigned a cell cycle score based on their expression of G2/M and S phase markers (Seurat_object <- CellCycleScoring(Seurat_object, g2m.features = g2m_genes, s.features = s.genes). To determine if cell cycle is a source of variation, principal component analysis (PCA) was used, and the most variable genes were identified. In addition, the *ScaleData* function was applied to the data, which shifts and scales the expression of each gene to ensure that the highly expressed genes do not dominate in the downstream analyses. PCA was performed on the scaled data by using the *RunPCA* function and the top principal components (PCs) were plotted and coloured by cell cycle phase. This plot determines whether or not it is required to regress out the variation due to cell cycle. The filtered datasets were normalised based on the greatest sources of variation using the *sctransform* function, which normalises the expression measurements for each nucleus by the total expression. In addition, variation due to mitochondrial expression was regressed out (**Figure 2.4: step 5**) (**Supplementary file 2**).

Cell clustering

Seurat (v4) uses a graph-based clustering approach, and the nuclei are clustered based on the similarities between their gene expression profiles (Butler *et al.*, 2018; Stuart *et al.*, 2018). The optimal number of PCs were determined which are required for the clustering analysis. The *ElbowPlot* function was used,

which plots the standard deviations of principal components and the optimal number of PCs to use would be where there is a clear elbow in the graph (ElbowPlot(object = seurat_integrated, ndims = 35)). To cluster the nuclei, the *FindNeighbors* and *FindClusters* functions with 20 PCs and a resolution of 0.4 was used for all datasets. *Seurat* applies the Louvain algorithm to group the nuclei (Butler *et al.*, 2018; Stuart *et al.*, 2018), resulting in a specific number of clusters for the integrated dataset. The *RunUMAP* and *DimPlot* functions were used to generate the UMAP plots (DimPlot(seurat_integrated, reduction = "umap", label = TRUE, label.size = 6)) (**Figure 2.4: step 6**) (**Supplementary file 2**).

Doublet Removal

One of the disadvantages of scRNA-seq technology is that it can easily produce technical artifacts such as doublets where droplets are filled with two cells or nuclei. It has been shown that this type of technical artifact can confound scRNA-seq data, so it is important to remove them (DePasquale *et al.*, 2019). Two different tools were used to perform doublet removal: *DoubletFinder* (v2.0.3) (McGinnis, Murrow and Gartner, 2019) and *DoubletDecon* (v1.15) (DePasquale *et al.*, 2019) in R studio (v4.1) (**Figure 2.4: step 7**). This was done to compare the accuracy of the results from these tools. *DoubletFinder* first simulates artificial doublets and then incorporates them into the snRNA-seq data. *DoubletFinder* then identifies real doublets based on each "real" cell's distance to artificial doublets (McGinnis, Murrow and Gartner, 2019). *DoubletDecon* applies a deconvolution method to the snRNA-seq data which involves the generation of deconvolution references, based on previously defined marker genes and cell clusters, from clustering approaches such as *Seurat* (DePasquale *et al.*, 2019). Next, *DoubletDecon* identifies a set of putative doublets based on this deconvolution analysis and cells that show any similarity to the artificial doublets are removed from the original cell clusters. Finally, the tool also "rescues" cells that were incorrectly predicted as doublets based on their unique gene expression (DePasquale *et al.*, 2019). The scripts for both tools can be found in the supplementary material (**Supplementary file 2**). The input files required were the individual filtered count matrices corresponding to each technical replicate for each sample (fourteen_T1.data <- Read10X(data.dir = "/path/to/outs/filtered_feature_bc_matrix/")). The *Seurat* objects were then generated from these matrices (fourteen_T1 <- CreateSeuratObject(counts = fourteen_T1.data, project = "14_T1", min.cells = 5, min.features = 100)). Only the barcodes that were identified as doublets by both tools were removed from each *Seurat* object. After the doublets were identified, they were removed from the *Seurat* object, and the filtered seurat objects for each sample were merged and then used as input for the integration, clustering, and annotation steps (merged_sample = merge(x = T1, y = T2, add.cell.ids = c("T1", "T2"))). The QC, normalisation and clustering steps were performed again on the merged Seurat object (**Figure 2.4: step 9**).

Integration

The next step in the pipeline, is to integrate the datasets using shared variable genes (Stuart *et al.*, 2019). *Seurat* uses the shared highly variable genes from each dataset which serves as a reference and integrates those datasets to overlay the cells that are similar between the two datasets. Integration ensures that cell types of the one dataset align with the same cell types of the other dataset (Stuart *et al.*, 2019). The *SCTransform* object was used as input to perform the integration across the datasets and the 3 000 most variable features were specified for integration (integ_features <- SelectIntegrationFeatures(object.list = split_seurat, nfeatures = 3000). The *SCTransform* object was then prepared for integration (split_seurat <- PrepSCTIntegration(object.list = split_seurat, anchor.features = integ_features)), canonical correlation analysis (CCA) was performed to identify the shared sources of variation in the data and the anchors were identified and filtered to remove any incorrect anchors from the datasets (integ_anchors <- FindIntegrationAnchors(object.list = split_seurat, normalization.method = "SCT", anchor.features = integ_features)). Anchors are pairs of nuclei (one from each dataset) that have a shared biological state and they are used for batch effect correction and integration. Finally, the integration was performed across the datasets using the *IntegrateData* function (seurat_integrated <- IntegrateData(anchorset = integ_anchors, normalization.method = "SCT")) (**Figure 2.4: step 10**). After integration, the data was visualised using dimensionality techniques, PCA and uniform manifold approximation and projection (UMAP). The *RunPCA* function was used to determine the principal components (Seurat_integrated <- RunPCA(object = Seurat_integrated)). UMAP will take the information from the top PCs to arrange the cells in the multidimensional space and take the distances in the multidimensional space and attempt to plot it in two dimensions using the *RunUMAP* function (seurat_integrated <- RunUMAP(seurat_integrated, dims = 1:30, reduction = "pca")). As a result, the nuclei from the merged samples are clustered based on the similarities between their gene expression profiles (Butler *et al.*, 2018; Stuart *et al.*, 2018) (**Supplementary file 3**).

Cluster annotation

To assign cell type identities to the clusters, marker identification analysis was performed to identify the genes that significantly differ in their expression between clusters using the *FindMarkers* function (markers <- FindAllMarkers(object = seurat_integrated, only.pos = TRUE, logfc.threshold = 0.25)). The automated cell type assignment tool, *SCSA* was used to assign cell types to the clusters. This tool takes the table of gene expression markers from the *FindMarkers* analysis as input (python3 SCSA.py -d whole.db -s seurat -I all_markers.csv -k brain -E -g Human -p 0.01 -f 1.5) (Cao, Wang and Peng, 2020). *SCSA* then compares this table of markers to known markers from the *CellMarker* database and the *CancerSEA* database (Yuan *et al.*, 2019; Zhang *et al.*, 2019).

Briefly, the *CellMarker* database is comprised of thousands of markers of approximately 467 different cell types and *CancerSEA* is a single cell database comprised of 14 distinct functional states of cancer cells from 25 different cancer types. To verify the cell types identified from the automated assignment, manual annotation was also performed, where the cell clusters were annotated by comparing data-derived marker genes with marker genes from the literature (Lake *et al.*, 2018; Song *et al.*, 2021). The expression of different cell-type specific markers across the entire dataset was visualised using the *FeaturePlot* function (e.g FeaturePlot(seurat_integrated, reduction = "umap", features = c("PLP1", "ST18"), sort.cell = TRUE, min.cutoff = 'q10', label = TRUE)) and *DoHeatmap* function (DoHeatmap(subset(seurat_integrated), features = markers.to.plot, cells = 1:500, size = 4, angle = 90)). The total number of nuclei in each cluster was determined using the *FetchData* function (number_of_cells <- FetchData(seurat_integrated, vars = c("ident", "orig.ident")) %>% dplyr::count(ident,orig.ident) %>% tidyr::spread(ident, n)). With all this information, the identity of the cell clusters was reassigned to the different cell types using the *RenameIdents* function and the *DimPlot* function was used to plot the labelled UMAP (**Figure 2.4: step 12**) (**Supplementary file 3**).

Differential gene expression analysis

Differential gene expression analysis was performed to identify genes that showed any significant changes in expression between two different brain regions, temporal, or frontal regions of the cerebral cortex within the major cell types of the brain. The two different brain regions were compared to each other, and this comparison was performed as a proof of principle for the DE analysis pipeline. After assigning cell types to the distinct clusters, differential expression analysis was performed using *DESeq2* (Love, Huber and Anders, 2014) and using a pseudo bulk approach. The HBC training scRNA-seq workshops were used as a guideline for this analysis step (https://github.com/hbctraining/scRNA-seq). Briefly, the processed integrated *Seurat* object was used as input and the raw counts and metadata were extracted (counts <- seurat_integrated_labelled@assays$RNA@counts, metadata <- seurat_integrated_labelled@meta.data). This was used to create the SingleCellExperiment object using the *SingleCellExperiment* tool in R studio (v4.1) (sce <- SingleCellExperiment(assays = list(counts = counts), colData = metadata)) (Amezquita *et al.*, 2020). A *DeSeq2* object was then generated for each cluster or cell type (dds_object <- DESeqDataSetFromMatrix(counts, colData = cluster_metadata, design = ~ group_id)). Next, *DESeq2* was used to normalise the count data, and this was used to generate PCA plots to ensure that the replicates for each sample are clustering together and to determine how similar the samples are to each other (rld <- rlog(dds_object, blind = TRUE), (DESeq2::plotPCA(rld, intgroup = "group_id"))). The differential expression analysis step was run using the *DESeq* command,

where the raw counts were fitted to the negative binomial model and hypothesis testing was performed using the Wald test (dds_object <- DESeq(dds_object)). These results were explored for the comparison between the two different brain regions and visualised using heatmaps, scatterplots and volcano plots (**Figure 2.4: step 13**) (**Supplementary file 4**).

Functional enrichment analysis

A list of significant differentially expressed genes (DEGs) for each cell type was obtained in the previous step. Functional enrichment analysis was performed on these gene lists to determine if there was any enrichment of known biological functions, interactions, or pathways. The Yu lab *ClusterProfiler* tutorial workbook was used as a guideline for these steps (https://yulab-smu.top/biomedical-knowledge-mining-book/). *ClusterProfiler* (Yu *et al.*, 2012; Wu *et al.*, 2021) was used to perform over-representation on Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Disease Ontology (DO) terms associated with the list of significant DEGs for each cell type. GO terms are arranged into three controlled ontologies: Biological process (BP), Molecular function (MF) and Cellular component (CC) (http://geneontology.org/). KEGG is a collection of pathway maps depicting molecular interaction and reaction networks. These pathways are divided into seven broad groups: genetic and environmental information processing, metabolism, organismal systems, cellular processes, human diseases, and drug development (https://www.genome.jp/kegg/). The DO database provides gene annotations associated to disease (https://disease-ontology.org/). This analysis was performed on the DE genes from all the cell types for the frontal vs. temporal comparison.

Firstly, the data was prepared, and a gene list created, so that it was in the correct format to be used as input. The significant DEGs for each cell type from the frontal lobe were combined and this was repeated for the temporal lobe. To calculate enriched functional categories for individual gene clusters, the *CompareCluster* function was used (Wu *et al.*, 2021). The full script for this analysis is outlined in **Supplementary file 5**. Briefly, the significant genes from each brain region were used as input (frontal_significant_genes = read.csv("significant_genes_frontal.csv"). A data frame with the cell type, fold change and gene ID was generated (data_frame_frontal <- frontal_significant_genes[,c(5,3,1)],). Next, the *CompareCluster* function was used for the three different gene ontologies ("enrichGO", "enrichKEGG" and "enrichDO" ontologies) (results_frontal <- compareCluster(Entrez~group, data=data_frame_frontal, fun="enrichGO", OrgDb='org.Hs.eg.db')). These results were then visualised using a dotplot. (**Figure 2.4: step 13**) (Wu *et al.*, 2021).

2.3 Filtering and lysis condition optimisations for future analyses

The initial bioinformatic analysis from the two snRNA-seq experiments showed that the nuclei preparations may have contained a large amount of cell debris or dead cells. To prepare for future experiments, further optimisation experiments were performed to assess the quality of the nuclei and reduce the amount of debris in the final nuclei preparation.

Nuclei isolation trial 1 and 2: Assessing nuclei quality

Nuclei suspensions were prepared as described in the nuclei isolation protocol above (**Figure 2.2**) and the age and tissue type are detailed in **Table 2.4**. The Lysis solution was prepared fresh for these experiments (Nuclei PURE lysis buffer (Sigma-Aldrich, NUC201), 1 M DTT, 10 % Triton-X 100). After each lysis and centrifugation step, an aliquot of nuclei was taken and stained with trypan blue to check the quality of the nuclei and the amount of cell debris.

Nuclei isolation trial 3: Guablomme et al. 2019 protocol

The Guablomme *et al.* nuclei isolation protocol was tested on 46-year-old frozen brain tissue and the results were compared to the results from the nuclei isolation protocol described above. This was done to see if this protocol would reduce the amount of cell debris as it only has one lysis step and makes use of a different lysis buffer and different cell filters. Frozen brain tissue was used as input material (**Table 2.4)**. Tissue fragments were disrupted to form a single cell suspension and the nuclei were isolated from the cells using methods described previously (Gaublomme *et al.*, 2019) (**Figure 2.5**). The buffers described below were all prepared fresh. Tissue samples were placed in a glass dounce homogeniser (KIMBLE Dounce tissue grinder set, Sigma-Aldrich, D8938) (on ice) containing 1 ml of ice-cold NP40 Lysis Buffer (NST) (0.1 % NP40, 10 mM Tris, 146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$). The tissue was ground 20 times with pestle A followed by 20 times with pestle B until there were no visible tissue pieces. 500 μl of ST Wash Buffer (10 mM Tris, 146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 0.01 % BSA) was added to the dounce homogeniser and the nuclei suspension was passed through a 30 μm cell strainer (Miltenyi Biotec, 130-041-407) (**Figure 2.5**). The suspension was transferred to a 2 ml Eppendorf tube. The dounce homogeniser was rinsed with $2X$ 200 μl of ST buffer, filtered through a 30 μm cell strainer and added to the filtered homogenate to add up to a final volume of 2 ml. The sample was centrifuged at 500 x g for 5 minutes at 4 ℃. The supernatant was discarded, and the pellet resuspended in 200 μl of ST Staining Buffer (ST-SB) (2 % BSA, 0.02 % Tween-20, 10 mM Tris,

146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$) and the nuclei suspension was passed through a 20 μm cell strainer (Miltenyi Biotec, 130-041-407). 10 μl of the filtered nuclei suspension was stained with trypan blue, loaded onto a hemocytometer, and counted under a fluorescent microscope (Zeiss) (**Figure 2.5**).

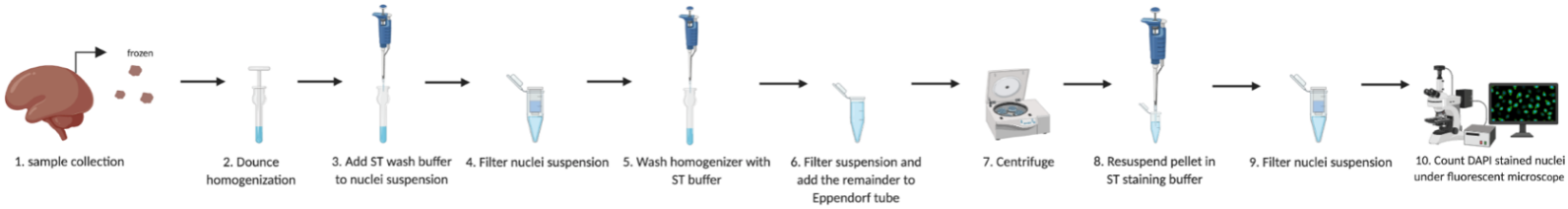


Figure 2.5: ***Schematic of sample preparation and nuclei isolation for isolation trial 3 (Guablomme et al. 2019 protocol).*** Flow diagram showing the steps of nuclei isolation using human brain tissue fragments as the input material. **(1)** Brain tissue samples were collected and flash frozen. **(2, 3)** Tissue fragments were homogenised, cells lysed to release the nuclei and ST wash buffer was added to the nuclei suspension. **(4)** The nuclei w ere passed through a **30 μm** filter and **(5)** then the homogeniser was washed with ST buffer. **(6)** The remaining nuclei in the homogeniser were passed through a **30 μm** filter and added to the **2 ml** Eppendorf tube. **(7)** The nuclei suspension was centrifuged. **(8)** The pellet was re suspended in ST staining buffer. **(9)** The nuclei suspension was then filtered through a **20 μm** filter. **(10)** Finally, an aliquot of nuclei was st ained and counted under the upright microscope. (Made in biorender).

Nuclei isolation trial 4 and 5: Standard nuclei isolation protocol with additional myelin removal step (Allen Brain Institute 2019 protocol)

Nuclei suspensions were prepared as described in the nuclei isolation protocol above using freshly prepared lysis solution (**Figure 2.2**) with an additional myelin removal step using a method described previously (Allen brain institute, https://dx.doi.org/10.17504/protocols.io.y6rfzd6). After passing the suspension through the 40 μm cell strainer, the nuclei were centrifuged at 900 x g for 10 minutes at 4 ^{o}C. The supernatant was discarded leaving 50 μl of solution sitting above the pellet. 3 ml of freshly prepared blocking buffer (10 X PBS, 1 % BSA and nuclease free water) was added to the pelleted nuclei and gently resuspended (**Figure 2.6**). 30 μl of myelin removal beads (Miltenyi Biotec, 130-096-733) were added to the solution, mixed 5 times, and incubated in the 4 ^{o}C refrigerator for 15 minutes. Another 3 ml of blocking buffer was added and mixed 5 times. The nuclei suspension was centrifuged at 300 x g for 5 minutes at 4 ^{o}C. The supernatant was removed and 2 ml of blocking buffer was added and the pellet resuspended and transferred to a 2 ml Eppendorf tube (**Figure 2.6**). The 2 ml tube was placed on a DynaMag™-2 magnet (ThermoFisher, 12321D) and incubated in the 4 ^{o}C refrigerator for 15 minutes. After 15 minutes, the supernatant was transferred to a new tube. 10 μl of

the nuclei suspension was stained with trypan blue, loaded onto a hemocytometer, and then counted under an upright microscope to determine the nuclei concentration (**Figure 2.6**).

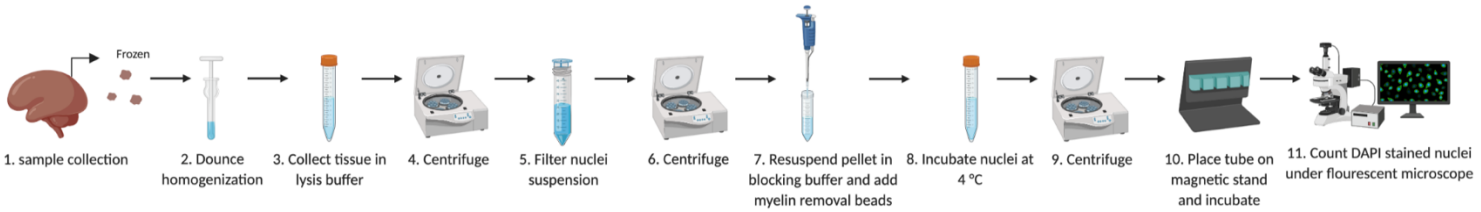


Figure 2.6: ***Schematic of sample preparation and nuclei isolation for isolation trial 4 and 5 (Allen Brain institute 2019 protocol).*** **(1)** Brain tissue samples were collected and flash frozen. **(2, 3)** Tissue fragments were homogenised, and cells lysed to release the nuclei. **(4,5)** The nuclei were centrifuged and then passed through a **40 µm** filter to remove cell debris. **(6)** The nuclei were centrifuged, and the supernatant discarded, leaving **50 *µl*** of the solution above the pellet. **(7)** The pellet was resuspended in blocking buffer and then myelin removal beads were added to the solution. **(8)** The nuclei were then incubated for 15 minutes. **(9)** More blocking buffer was added, and the solution was centrifuged. **(10)** The supernatant was discarded, **2 ml** of blocking buffer was added to the pellet and then placed on a magnetic stand. **(11)** After incubating for 15 minutes on the magnetic stand, the supernatant was transferred, and an aliquot of nuclei were stained and counted under the microscope. (Made in biorender).

Table 2.4: Summary of conditions used for the nuclei isolation trial experiments. Experiment trial number, tissue type, age of sample, size of cell strainer used, and number of tissue pieces used for each trial.

Nuclei isolation trial	Tissue type	Age	Cell strainer used (size)	Number of tissue pieces used
1	frozen frontal cortex	6-year-old	40	2
2	frozen frontal cortex	46-year-old	40	1
3	frozen frontal cortex	46-year-old	30 and 20	1
4	frozen temporal cortex	15-year-old	40	1
5	frozen temporal cortex	31-year-old	40	1

2.4 Assays for Transposase-Accessible Chromatin (ATAC-seq)

ATAC-seq is a method that has been developed for assaying accessible regions of chromatin, as this is where putative active enhancers are likely to be located. This method involves exposing nuclei to a Tn5 transposase that cuts and insert Illumina sequencing adapters into the open regions of chromatin **(Figure 2.7)**. The DNA is then amplified and sequenced (Buenrostro *et al.*, 2013) **(Figure 2.7)**. After sequencing, the libraries are processed, filtered, and analysed using several bioinformatic tools.

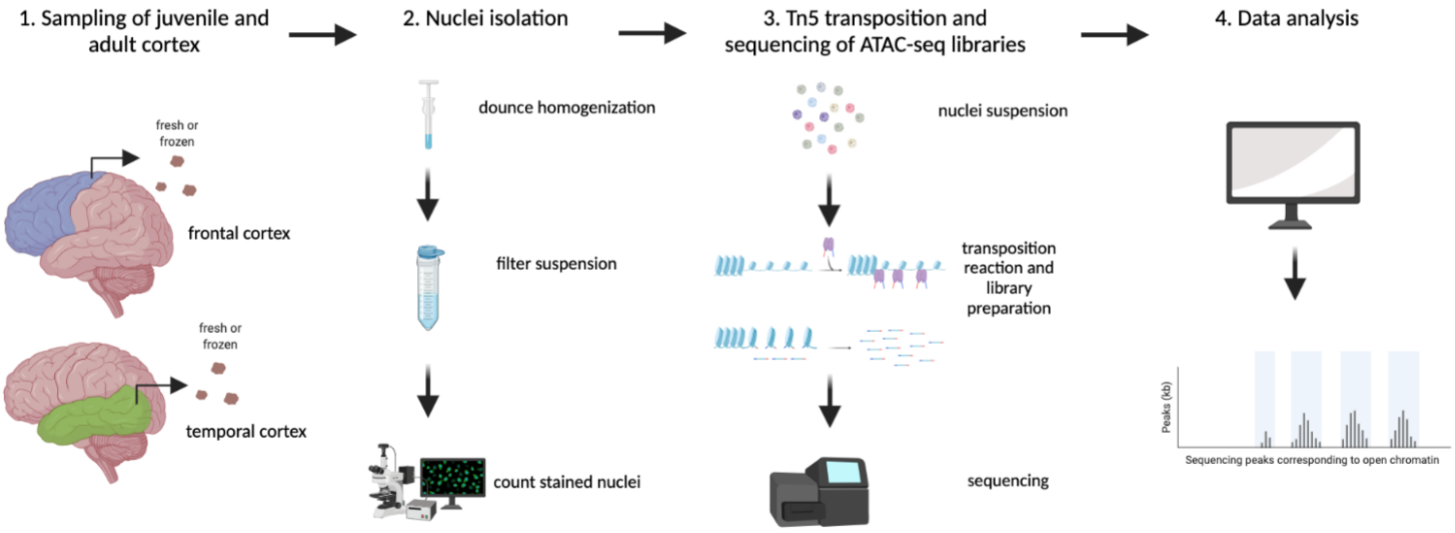


Figure 2.7: ***Overview of experimental workflow for ATAC-seq of temporal and frontal cortex.*** **(1)** Tissue fragments were collected and prepared. **(2)** Fresh, frozen, or cryopreserved tissue fragments were used as starting material for the nuclei isolation experiment. This generated a suspension of nuclei. **(3)** The nuclei suspension was used as the input material for the Tn5 transposition reaction and ATAC-seq library preparation. These ATAC-seq libraries were then sent for sequencing**. (4)** The raw sequencing data was processed and analysed using Bowtie2 and other bioinformatic tools. (Made in biorender).

2.4.1 Sample preparation and Transposition Reaction

ATAC-seq was performed on seven pediatric tissue samples and three adult tissue samples (see **Table 2.1**). The samples were either used in a fresh, frozen, or cryopreserved state (see **Table 2.1**). The fresh tissue fragments were kept on ice and used immediately after collection. Frozen fragments had been slow frozen and then transferred to a -80 ℃ freezer immediately after surgery. Cryopreserved fragments were placed in cryopreservation solution (10 % DMSO, 5 % FBS, 40 % DMEM). These fragments were then slow-frozen in a -20 ℃ freezer for two hours before being transferred to the -80 ℃ freezer for long term storage. The nuclei isolation protocol was adapted from the Habib *et al.* protocol and the ATAC-seq protocol was adapted from the Buenrostro *et al.* 2016 protocol (Buenrostro *et al.*, 2016; Trinh *et al.*, 2017). Nuclei suspensions were prepared as described above (**section 2.2.1**) in the nuclei isolation protocol for snRNA-seq, without the use of 0.1 % RNase inhibitor in the final NSB solution (**Figure 2.2**). 10 μl of the filtered nuclei suspension was stained with Hoechst or trypan blue, loaded onto a hemocytometer, and counted under an upright microscope (Zeiss) (**Figure 2.2**). The concentration of the nuclei was determined, and the volume of the nuclei suspension required for 50 000 nuclei per sample was used in the following steps. Buenrostro *et al.* has shown that 50 000 cells are required as input because performing ATAC-seq with a smaller number of cells causes extensive digestion of chromatin but using too many cells can result in some of the chromatin not being digested, creating high molecular weight fragments which are difficult to sequence (Buenrostro *et al.*, 2016).

50 000 nuclei were spun down at 500 $x\ g$ for 10 min at 4 ℃. The supernatant was discarded, and the pellet kept on ice. The transposition reaction mix (2x TD buffer, Tn5 transposase (Tagment DNA enzyme 1) and Nuclease free H_2O [Illumina]) was added to the pellet and gently resuspended. The transposition reaction was incubated at 37 ℃ for 30 min. 500 mM of EDTA was added to the reaction a final concentration of 50 nm and the reaction incubated at 50 ℃ for 30 min to stop tagmentation. The reaction mix was purified using a PCR purification MinElute kit (Qiagen, 28004). The transposed DNA was eluted in 20 μl of Elution buffer (Qiagen, 28004) (Buenrostro *et al.*, 2016).

2.4.2 PCR Amplification and sequencing

PCR amplification was performed to amplify the DNA, add sample specific barcodes, and generate libraries for sequencing. The PCR amplification mix (**Table 2.5**) was added to the transposed DNA on ice. The DNA libraries were amplified in a thermocycler (**Table 2.6**). The amplified DNA was purified using the PCR purification MinElute kit according to the manufacturer's instructions (Qiagen, 28004) and eluted in 10 μl

of elution buffer. Lastly, an additional clean up using KAPA pure beads (KAPA, KR1245-v3.16) was performed to remove excess primers. The beads were warmed to room temperature and vortexed thoroughly. 30 μl of the beads were added to the library preparation, pipette mixed 10 times, and incubated at room temperature for 10-15 min. The solution was spun down briefly and placed in a magnetic stand. Once the solution was clear, the supernatant was removed and discarded, without removing the magnetic beads as they contain the DNA. 200 μl of 80 % ethanol was added to the solution while in the magnetic stand and removed after 30 seconds. This wash step was repeated and then the sample was spun down briefly to remove the excess ethanol. The tubes were placed back into the magnetic stand with the lids open for the beads to air dry. To elute the DNA from the beads, 20 μl of TE was added and mixed thoroughly. The solution was spun down briefly and placed in the magnetic stand for 5 mins. Approximately 18 μl of the supernatant which contained the eluted DNA library was removed and placed into a new tube. To check the quality (average fragment size) of the ATAC-seq libraries, 1 μl of each sample was run on the Agilent Bioanalyzer High Sensitivity chip on the Agilent 2000 Bioanalyzer instrument (Agilent, G293BA) by CAF (Stellenbosch University) or CPGR (Cape Town). Library concentration was determined using Qubit (Thermo Fisher Scientific) by CAF (Stellenbosch University) or CPGR (Cape Town). The ATAC-seq libraries were sequenced by Novogene (Singapore) on the NovaSeq High Output v2.5 kit (150 cycles) using standard paired-end sequencing conditions.

Table 2.5: PCR amplification mix. List of reagents and volumes used for the PCR amplification step.

Reagent	**Volume (μl)**
Transposed DNA	10
Nuclease Free H_2O	10
customized Nextera PCR primer 1 (universal)	2.5 (10 μM)
customized Nextera PCR primer 2 (barcode)	2.5 (10 μM)
KAPA HiFi Hotstart ReadyMix	25 (2x)

Table 2.6: PCR cycling conditions. PCR cycling conditions used for the amplification step of the ATAC-seq protocol, showing the temperature, duration, and cycle number for each step of the PCR protocol.

Steps	Temperature (°C)	Duration	Cycle
1	72	5 min	1
2	98	3 min	1
3	98	20 sec	12
4	63	30 sec	
5	72	1 min	
6	4	Hold	1

2.4.3 ATAC-seq Analysis

Sequencing was performed and the data files were processed to generate fastq files. After quality control, the reads were aligned to the human reference genome using *Bowtie2* (**Figure 2.8**). QC steps and peak calling was performed using two different pipelines: Reske *et al.* pipeline and the Harvard Informatics pipeline. This was followed by additional downstream analyses such as peak annotation (**Figure 2.8**).

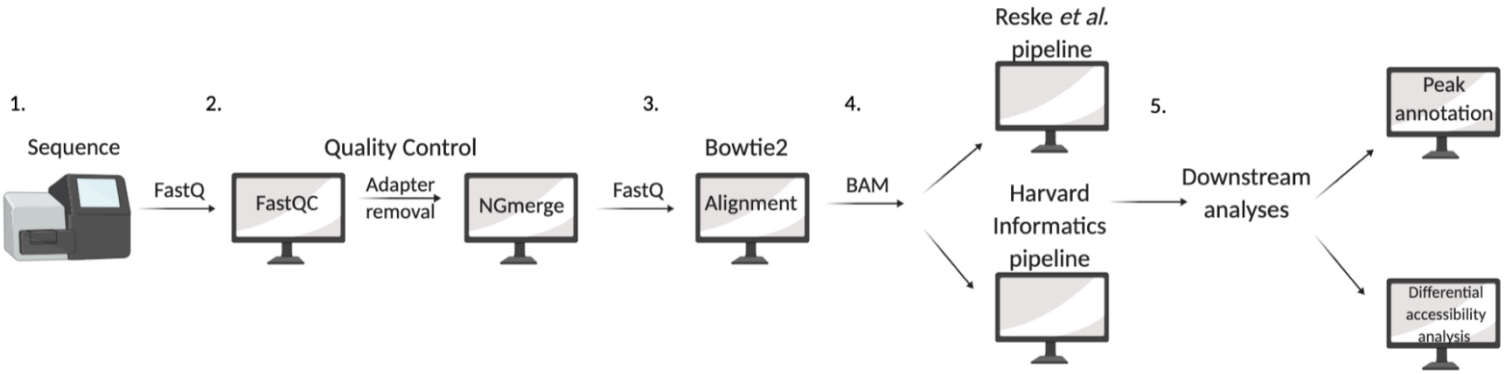


Figure 2.8: ***Overview of bioinformatics workflow for ATAC-seq of temporal and frontal cortex.*** **(1-2)** ATAC-seq libraries were sequenced, generating FastQ files and quality control was performed using *FastQC*. After running *FastQC*, the adapters were removed using *NGmerge*. **(3)** Reads were aligned to the reference genome using *Bowtie2*. **(4)** BAM files were used as input and the remaining QC steps were performed using two different pipelines: Reske *et al.* pipeline and the Harvard Informatics pipeline which used different bioinformatic tools and two different peak-calling methods. **(5)** Finally, downstream analyses steps were performed such as peak annotation and differential accessibility analysis, using several bioinformatic tools. (Made in biorender).

Pre-alignment quality control

FastQC (v0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to check the quality of the raw sequence data (fastqc <sample>.R1.fastq.gz). The FastQC report generated was also used to check for Illumina adapter contamination. The adapters were removed using the *NGmerge* tool (Gaspar, 2018b) (NGmerge -a -1 <sample>.R1.fastq.gz -2 <sample>.R2.fastq.gz -o -v) (**Figure 2.8: step 2**).

Alignment and post-alignment quality control

The reference genome (human GrCh38, https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/latest/) was indexed using the bowtie2-build command (bowtie2-build <genome.fa> <genomeIndexName>). The reads were aligned to the reference genome using *Bowtie2* (Langmead and Salzberg, 2012) and sorted using *samtools* (Li *et al.*, 2009) (bowtie2 --phred33 -p 4 -X 2000 --very-sensitive -x $GENOME -1 <name>_1.fastq -2 <name>_2.fastq | samtools view -b - > $A3\.bam) (**Figure 2.8: step 3**) (**Supplementary file 6**). After alignment, the distribution of the fragment sizes was assessed using the Picard Tools (http://broadinstitute.github.io/picard/) (v2.22.8) *CollectInsertSize* command. This gives an indication of the quality of the aligned ATAC-seq data.

The quality of the ATAC-seq libraries were also checked on the UCSC genome browser (https://genome.ucsc.edu/index.html). First bigwig (bw) files were generated from the BAM files using *bedtools* (Quinlan and Hall, 2010) (v2.29) (genomeCoverageBed -bg -split -sample.bam -g $CHROM > sample.bg, bedGraphToBigWig sample.bg $CHROM sample.bw). These files were then uploaded onto a private human genome assembly hub on the UCSC genome browser (https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#UseOneFile). In addition to checking the quality of these libraries, they were also compared to the publicly available datasets: the GeneHancer track set (Fishilevich *et al.*, 2017) and the ENCODE regulation track set (Thurman *et al.*, 2012).

Peak-calling

One of the most important steps in the ATAC-seq pipeline is peak calling, which involves the use of peak-calling tools to identify regions of genomic enrichment or peaks. These represent regions of significantly accessible chromatin. After the reads were aligned to the reference genome generating BAM files, two different peak-calling methods were tested and compared. Firstly, the BAM files for each sample were

processed and filtered using the Reske *et al.* ATAC-seq pipeline (https://github.com/reskejak/ATAC-seq) which uses the *MACS2* peak-caller (Gaspar, 2018a) (Reske, Wilson and Chandler, 2020) (**Figure 2.8: step 4**, **Figure 2.9**). Secondly, the BAM files were processed using the Harvard FAS Informatics pipeline (https://informatics.fas.harvard.edu/atac-seq-guidelines.html#qc), which uses the *Genrich* peak-caller (https://github.com/jsh58/Genrich) (**Figure 2.8: step 4, Figure 2.9**).

Additionally, these peak files were used to generate bigbed (bb) files using *bedtools* (Quinlan and Hall, 2010) (v2.29) (bedToBigBed sample.bed $CHROM sample.bb). These files were then uploaded onto a private human genome assembly hub on the UCSC genome browser (https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#UseOneFile).

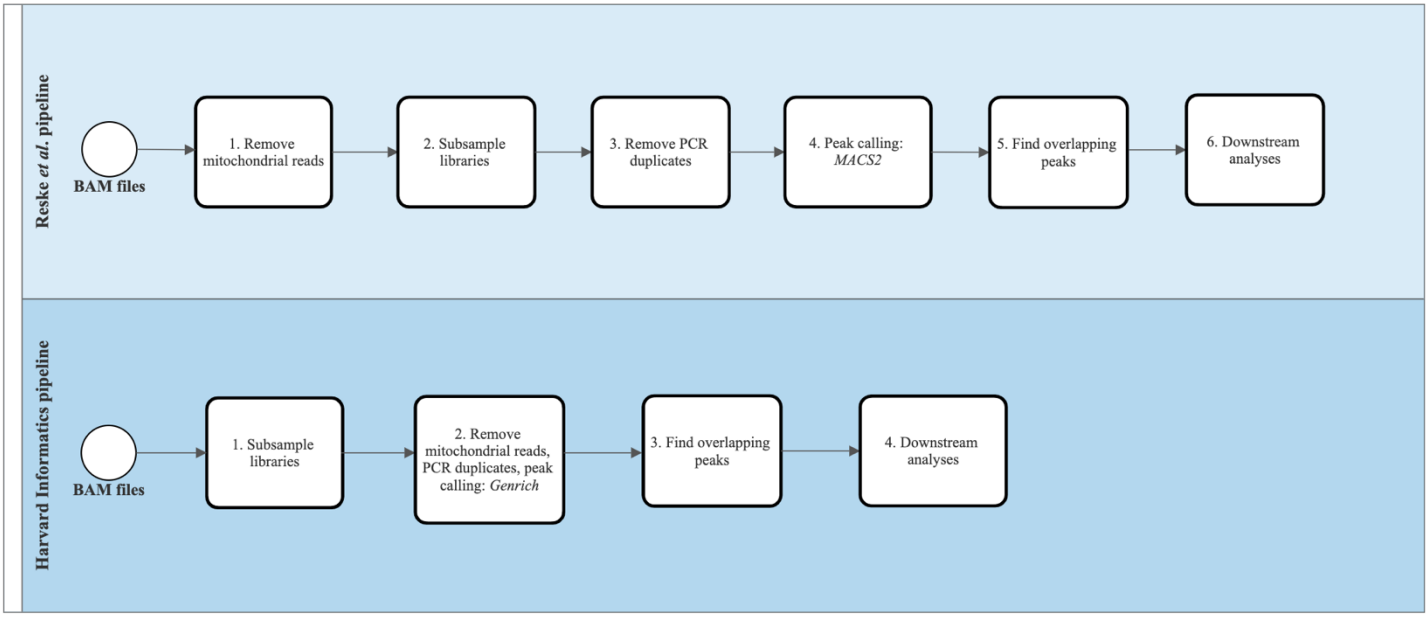


Figure 2.9: ***Processing workflow using two different pipelines.*** The overall data flow and processing steps for the Reske *et al.* pipeline and Harvard Informatics pipeline.

Reske *et al.* pipeline using *MACS2*

Removal of mitochondrial reads

ATAC-seq datasets usually contain a large percentage of reads derived from mitochondrial DNA which needed to be removed as this study focuses on nuclear DNA. The python script *removeChrom* (https://github.com/harvardinformatics/ATAC-seq) was used to remove the mitochondrial reads (samtools view -h sample_sorted.bam | python removeChrom -- chrM | samtools view -b - > Sample_sorted_noMT.bam). An additional filtering step was performed to only retain properly paired reads for downstream analysis using *samtools* view (Li *et al.*, 2009) (samtools view -bh -f 3 sample.bam > sample.filtered.bam) (**Figure 2.9**).

The complexities of all samples were determined in R studio (v4.1) using the *preseqR* (Deng, Daley and Smith, 2015) and *ATACQC* (Ou *et al.*, 2018) packages. First the duplication frequency matrix was determined using the *readsDupFreq* function (sample.dups <- readsDupFreq(sample, index = sample.bai)). Then the library complexity was estimated using the *estimateLibComplexity* function (sample.complexity <- estimateLibComplexity(sample.dups, times=100, interpolate.sample.sizes=seq(0.1, 1, by=0.01))). *Samtools view* was used (Li *et al.*, 2009) to subsample or normalize all sample libraries to equivalent molecular complexity i.e. to the complexity value of the library with the lowest complexity (e.g. samtools view -h -b -s 1.35 sample_filt.sorted.noMT_fresh.bam > sample_sub.filt.noMT_fresh.bam). The library with the lowest complexity was not subsampled (**Supplementary file 7**).

PCR duplicates

After normalizing library complexity, PCR duplicates were removed using the *Picard MarkDuplicates* command (http://broadinstitute.github.io/picard/) (v2.22.8) (java -jar picard.jar MarkDuplicates I=Sample_sorted_filt_noMT.bam O=Sample_noDups_filt_noMT.bam M=Sample_dups.txt REMOVE_DUPLICATES=true) (**Figure 2.9**).

Paired-end BED format (BEDPE) conversion

The BAM files were sorted by read name and associated read mate information was fixed using *samtools* (Li *et al.*, 2009). The BAM files were then converted to BEDPE format using the *bedtools bamtobed* command (Quinlan and Hall, 2010) (samtools view -bf 0x2 Sample_fixed.bam | bedtools bamtobed -i stdin

-bedpe > Sample_fixed.bedpe). The next step was to shift BEDPE coordinates +4 and -5 bp respectively to compensate for the Tn5 insertion in the ATAC-seq data which was performed using the bash script *bedpeTn5shift.sh* (https://github.com/reskejak/ATAC-seq). Finally, the bash script *bedpeMinimalConvert.sh* was used to convert the standard 10-column *bedtools* format to the minimal 4-column BEDPE format that is required by *MACS2* (https://github.com/reskejak/ATAC-seq).

Peak-calling

Broad peaks were called from the minimal BEDPE files using *MACS2* (macs2 callpeak -t Sample2_tn5.bedpe -f BEDPE -n Sample2 -g hs --broad --broad-cutoff 0.05 --keep-dup all). ENCODE-defined blacklisted regions were removed from the peaks using *bedtools intersect* (bedtools intersect -v -a sample2.broadPeak -b hg38.blacklist.bed | grep -P 'chr[\dXY] + [\t]' > sample2.filtered.broadPeak) (Quinlan and Hall, 2010; Amemiya, Kundaje and Boyle, 2019). The *bedtools multiinter* command was used to identify peaks that were common between the technical replicates for each sample (i.e. peaks that showed any amount of overlap) (bedtools multiinter -I sample1.bed sample2.bed | awk '$4 == 2' > common.bed) (Quinlan and Hall, 2010). Peaks that displayed at least 50% overlap between technical replicate samples were identified using the bash script *naiveOverlapBroad.sh* (https://github.com/reskejak/ATAC-seq). Peaks that were unique to each sample were determined using *bedtools intersect* command (bedtools intersect -a sample.bed -b common.bed -v > sample_unique.bed) (**Figure 2.8: step 4, Figure 2.9**).

Harvard FAS Informatics pipeline using *Genrich*

Before peak calling, the libraries were subsampled using the *preseqR* (Deng, Daley and Smith, 2015) and *ATACQC* (Ou *et al.*, 2018) packages as described above and the library with the lowest complexity was not subsampled. *Genrich* (v0.6) (https://github.com/jsh58/Genrich) was used to perform all filtering steps, including the removal of PCR duplicates, mitochondrial reads and ENCODE blacklisted regions, as well as peak calling (Genrich -t <BAM> -E hg38.blacklist.bed -o <OUT> -j -y -r -e chrM -v). Next, the peaks that displayed at least 50 % overlap between technical replicate samples were identified using methods described above. The *Genrich* peak calling command generates narrowPeak files so these files had to be converted to broadPeak files so that it was compatible with the *naiveOverlapBroad.sh* script (**Figure 2.8: step 4, Figure 2.9**).

Peak Annotation

HOMER (Heinz *et al.*, 2010) (v4.10) was used to annotate identified peaks using the *annotatePeaks.pl* command (annotatePeaks.pl sample.bed hg38 -go annotate_peaks -annStats common__S9_peaks.stats > common_all_S9_peaks.annotated.txt) (http://homer.ucsd.edu/homer/ngs/annotation.html). *HOMER* associates the ATAC-seq peaks with nearby genes and assigns them to different genomic categories such as: Promoter-TSS, exonic region, intronic region, intergenic region, and transcription termination site (TTS) (**Figure 2.8: step 5**).

Differential accessibility analysis

Differential accessibility analysis of ATAC-seq datasets can also be used to determine regions that are differentially accessible (DA) between two states. Here, the ATAC signal at enriched regions was quantified and compared between the pediatric and adult samples (Reske, Wilson and Chandler, 2020) using *csaw* (v1.22.1) (Lun and Smyth, 2015) in R studio (v4.1) following the Reske *et al.* csaw workflow (https://static-content.springer.com/esm/art%3A10.1186%2Fs13072-020-00342-y/MediaObjects/13072_2020_342_MOESM6_ESM.txt). The adapted version of the Reske *et al*. csaw script was used for the DA analysis (**Supplementary file 8**) on the *MACS2* peaks and *Genrich* peaks separately. The broadPeak, nave overlap broadPeak and sorted paired-end BAM files (generated in the previous step) for each sample were used as input ((sample1.peaks <- read.table("path/to/file/sample1.broadPeak", sep="\t")[,1:3])(paired-end.bams <- c("samples.bam"))). Next, a consensus peak set was generated using the *union* command which involves the union of the nave overlap peaks or a normal broadPeak file. The normal broad peak files were used for each cryopreserved replicate (they were treated as separate samples) and they were also used if there was not a replicate for that sample (peaks1 <- union (sample1.overlap.peaks, sample2.overlap.peaks), peaks2 <- union (peaks1, sample3.overlap.peaks) ,etc.). The final consensus peak sets used for both the *MACS2* and *Genrich* analyses consisted of the peaks that showed at least 50 % overlap between the technical replicates, identified using the *naiveOverlapBroad.sh* bash script or the normal broadPeak files when necessary.

Two different normalisation methods were applied to both the *MACS2* and *Genrich* peaks: trimmed mean of M values (TMM) (Robinson and Oshlack, 2010) and non-linear locally estimated scatterplot smoothing (loess) normalisation. The TMM method identifies a reference region and estimates the fold change and signal abundance based on that reference. The regions are trimmed by these values and the trimmed mean of the fold changes (M-values) is determined for each region. This method assumes that the majority of

regions are not differentially accessible (Robinson and Oshlack, 2010). The loess normalisation method is highly conservative and normalises the signal distribution locally based on the extent of ATAC signal abundance. Therefore, loess normalisation assumes a symmetric global distribution where there are no true biological differences. It assumes that these differences are technical and need to be removed (Reske, Wilson and Chandler, 2020). After normalisation, *EdgeR* (Robinson, McCarthy and Smyth, 2009) (v3.30.3) was used for DA analysis to determine significant DA regions. MA plots were generated using the *ggplot* function to identify global accessibility patterns between samples. The samples were grouped into four developmental periods or categories and pairwise DA analysis was performed comparing the different categories to each other (**Table 2.7**) (**Figure 2.9**). The sample groupings were based on developmental periods defined by Kang *et al.* and Werling *et al.* (Kang *et al.*, 2011; Werling *et al.*, 2020). Again, this was performed separately for the *MACS2* and *Genrich* peaks as well as the two different normalisation methods. This resulted in six different comparisons being performed using four different analysis approaches. Approach *I* utilised regions defined by the *MACS2* consensus peak set and a TMM normalisation method, while Approach *II* used the Loess normalisation method on the same peaks. Approach *III* utilised regions defined by the *Genrich* consensus peak set and a TMM normalisation method, while Approach *IV* used the Loess normalisation method on the same peaks.

Table 2.7: The four developmental categories used for ATAC-seq DA analysis and corresponding samples.

Category	Samples
Early childhood	19-month-old, 23-month-old, 5-year-old
Middle childhood	6-year-old, 9-year-old
Late childhood	14-year-old, 15-year-old
Adulthood	23-year-old, 31-year-old, 46-year-old

Functional enrichment analysis

Functional enrichment analysis can be performed to identify biological functions, processes or interactions that are enriched in the genes of interest. After performing DA analysis, BED files containing the genomic coordinates of the DA peaks for each comparison were generated and used for functional enrichment

analysis. These DA peaks were annotated using *HOMER* and assigned to the different genomic category annotations: Promoter-TSS, exonic region, intronic region, intergenic region, and TTS. This list of genes that are linked to the DA peaks from Approach *I*, *III* for the early childhood vs adult comparison and Approach *I* for the late childhood vs adult comparison were used as input to identify over-represented GO terms, utilising the *ClusterProfiler* tool (Yu *et al.*, 2012; Wu *et al.*, 2021). The analysis performed was similar to that performed on the snRNA-seq data described above, using the Yu lab online book as a guide (https://yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.html). The adapted version of this script was used for enrichment analysis (**Supplementary file 9**). Briefly, the gene lists for each comparison were prepared and used as input (early_vs_adult_genes = read.csv("early_vs_adult_all_genes.csv")). The significant genes were then specified using the false discovery rate (early_vs_adult_significant_genes <- names(early_vs_adult_genes)[abs(early_vs_adult_genes) < 0.05]). Next, the *enrichGO* function was used to identify enriched genes (early_adult_GO <- enrichGO (gene = early_vs_adult_significant_genes, universe = names(early_vs_adult_genes), OrgDb = org.Hs.eg.db, ont = "BP", pAdjustMethod = "BH", pvalueCutoff = 1, qvalueCutoff = 1, readable = TRUE)). The *simplify* function was performed on the enrichGO output file to remove any redundant GO terms (early_adult_simplified <- simplify(early_adult_GO, cutoff=0.7, by="p.adjust", select_fun=min)) (Wu *et al.*, 2021). The top 10 GO terms were then visualized using a dot plot (dotplot(early_adult_simplified, showCategory=10) + ggtitle("dotplot for GO enrichment: early vs adult")).

Chapter 3 - Results

3.1 Sample collection and preparation

The brain tissue samples in this study come from the temporal and frontal cortex, which were used as input material for the snRNA-seq and ATAC-seq experiments. Tissue fragments were obtained during surgery to remove epileptic regions and these fragments were the access tissue that likely represent normal brain tissue. Four female samples and six male samples were collected, ranging from 20-months-old to 46-years-old (see **Table 2.1**).

3.2 SnRNA-seq using the 10x Genomics platform

3.2.1 Brain tissue nuclei isolation

14-year-old frontal cortex and 15-year-old temporal cortex tissue were prepared for 10x Genomics snRNA-seq using the standard nuclei isolation protocol (see **Figure 2.2**). After the final centrifugation step, the nuclei were stained with Hoechst or trypan blue and counted under the fluorescent microscope (**Figure 3.1**). A higher nuclei yield was obtained for 14-year-old T2 compared to T1, while a higher nuclei yield was obtained for the 15-year-old compared to the 14-year-old sample (**Table 3.1**). A large amount of cell debris was present for both samples and this was something to consider for the remaining processing steps (**Figure 3.1**).

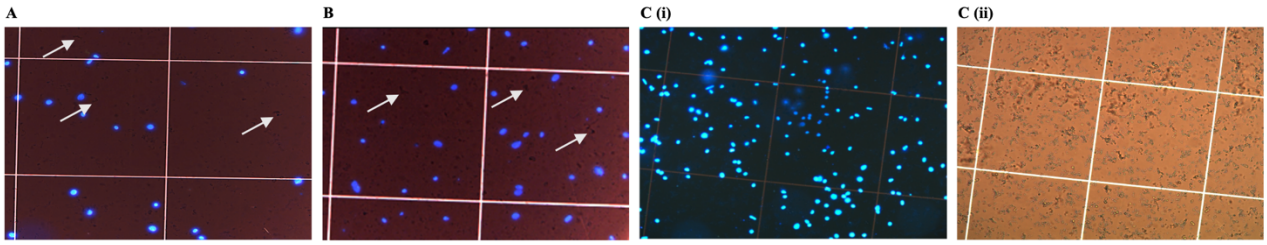


Figure 3.1: ***Hoechst-stained nuclei.*** Nuclei were isolated and counted under the fluorescent microscope. (**A-B)** Nuclei suspensions for two technical replicates from the 14-year-old sample stained with Hoechst, with arrows indicating cell debris. (**C**) Nuclei suspension from the 15-year-old sample stained with Hoechst **(i)** alongside the corresponding brightfield image showing cell debris **(ii)**.

Table 3.1: 14-year-old and 15-year-old nuclei isolation conditions and results. The nuclei yield for each sample is shown. T, Technical replicate

Experiment	Sample	Yield (nuclei/ul)
14-year-old 10x run	T1	1 800
14-year-old 10x run	T2	3 030
15-year-old 10x run	-	4 785

3.2.2 10x Genomics snRNA-seq library preparation and alignment

The 14-year-old sample was processed using the 10x Genomics Single Cell 3' Reagent kit V3 and the 15-year-old sample was processed with the 3' Reagent kit V3.1. After generating snRNA-seq libraries, the quality of the libraries was checked by assessing the average fragment size, using the Agilent Bioanalyzer (**Supplementary figure 1**). The cDNA fragment size distribution for the 14-year-old and 15-year-old technical replicates were $300 - 600\ bp$, which is expected and an indication of a good quality library. The libraries were sent for sequencing and the raw data was processed using Cell Ranger (v3.1, 10x Genomics).

The Cell Ranger *count* pipeline was used to align the sequencing reads to the human reference genome, generating alignment statistics (**Supplementary table 1**) and several files that were used for further analysis. The targeted nuclei recovery for both samples was 10 000 nuclei, however only 2 354 nuclei and 4 517 nuclei were obtained for 14-year-old T1 and T2 respectively. 8 985, 9 082, 8 972 and 8 791 nuclei were recovered for 15-year-old T1, T2, T3 and T4 respectively (**Supplementary table 1**). 10x Genomics advises that approximately $40\ \% - 60\ \%$ of nuclei should be recovered, however more than 50 % of nuclei were lost for both technical replicates from the 14-year-old sample. Additionally, the ideal fraction of reads in the nuclei is 70 % and above but for T1 and T2, the fraction of reads in the nuclei was 45.2 % and 66.3 % respectively. For the 15-year-old sample, the fraction of reads in the nuclei ranged from 59 % to 61.3 % (**Supplementary table 1**). Values below 70 % could be a possible indication of high levels of ambient RNA (10X Genomics, 2020).

3.2.3 Pre-processing of snRNA-seq data and quality control (QC)

Seurat (v4) and other tools such as *DoubletDecon* (v1.16) and *DoubletFinder* (v2.0.3) were used in R studio (v4.1) to perform several pre-processing steps, which generated quality control metrics and plots. These

QC metrics were then used to filter and select the best quality nuclei. For example, snRNA-seq datasets can contain a large number of reads from the mitochondria which is an indication of dead or dying cells (Zhao, 2002). Therefore, visualizing these QC metrics, allows the identification of dead or dying cells to be filtered from the data.

QC checks such as identification of reads from mitochondrial genes and pre-processing steps such as clustering were performed on each technical replicate for each sample separately. The quality control plots before filtering (not shown) were used to determined what filtering parameters to use and then the filtered data metrics were replotted (**Supplementary figure 2)**. These filtering parameters removed the low-quality nuclei without removing any high-quality nuclei, so they were used for subsequent filtering steps. The number of nuclei removed after these pre-processing steps were 153 (14 T1), 340 (14 T2), 215 (15 T1), 160 (15 T2), 155 (15 T3) and 177 (15 T4) respectively. The doublet removal tools mentioned above were then used to remove any doublets from the individual technical replicates for each sample (Methods, **Figure 2.4: step 7**). First, *DoubletDecon* was used to identify potential doublets on the data and then *DoubletFinder* was run on the same data for each technical replicate (**Table 3.2**). Based on previously published data, it was recommended to use a combination of doublet removal tools and to filter out nuclei that fell in the intersection of these predictions (DePasquale *et al.*, 2020). The results in **Table 3.2**, showed that there was a clear difference in the number of doublets identified for each tool and *DoubletDecon* identified a greater amount of doublets compared to *DoubletFinder*. To avoid the possibility of removing false positive doublets identified using *DoubletDecon,* only the barcodes that were identified as doublets by both *DoubletFinder* and *DoubletDecon* were removed from the datasets (**Table 3.2**). Overall, after all filtering steps, a small percentage ($7.6\ \% - 11.71\ \%$) of the total number of nuclei was removed from each replicate (**Table 3.2**). A greater percentage of nuclei was removed for the 14-year-old replicates compared to the 15-year-old replicates (**Table 3.2**).

Table 3.2: The number of nuclei identified for each doublet removal tool, the intersection between both tools and the total number of nuclei before and after doublet removal and after the final filtering steps are shown. The percentage of nuclei removed after all filtering steps is also shown.

	Total nuclei before initial filtering	Total nuclei before doublet removal (i.e., after initial filtering)	DoubletDecon	DoubletFinder	Intersection	Total nuclei after doublet removal	% of total doublets removed	Total nuclei after final filtering steps	% of nuclei removed after final filtering
14-year-old T1	2 354	2 201	393	65	41	2 160	1.86	2 089	11.26
14-year-old T2	4 517	4 177	274	167	30	4 147	0.72	3 988	11.71
15-year-old T1	8 985	8 902	931	534	290	8 612	3.26	8 362	8.28
15-year-old T2	9 082	9 038	1 323	542	338	8 700	3.74	8 492	7.68
15-year-old T3	8 972	8 817	858	543	229	8 588	2.6	8 505	7.6
15-year-old T4	8 791	8 614	702	535	228	8 386	2.65	8 379	7.85

3.2.4 Pre-processing of merged snRNA-seq data

The 14-year-old and 15-year-old replicates were combined into a single merged Seurat object and the same QC checks were performed on the merged data (**Figure 2.4: step 8-9**). The largest difference between the two samples was the total number of nuclei recovered out of the 10 000 targeted nuclei. Only 20 % and about 45 % of nuclei were recovered for the 14-year-old replicates compared to the 15-year-old replicates, where 86.5 % – 87 % of nuclei were recovered (**Figure 3.2A**). There was also a greater range of UMI counts and gene counts detected for the 14-year-old sample compared to the 15-year-old sample (**Figure 3.2A, B**). Lastly, the 14-year-old T2 replicate had a higher density of nuclei with high mitochondrial gene read count compared to T1 and the 15-year-old samples (**Figure 3.2E**). It is important to note that the 14-year-old sample was processed first, and the overall quality of the resulting snRNA-seq libraries for both replicates was not ideal. The 15-year-old sample was processed after performing more test runs with the nuclei isolation protocol, and there is a clear improvement in the quality of the resulting snRNA-seq libraries, especially with regards to the similarity between the replicate samples the number of nuclei recovered, when compared to the 14-year-old snRNA-seq libraries (**Table 3.2**; **Figure 3.2**).

Normalisation was performed using Seurat's *sctransform* which removes variation due to sequencing depth and mitochondrial gene expression. In addition, the cell cycle score was checked before performing normalisation to determine if cell cycle is a major source of variation in the dataset. The nuclei were assigned a score based on its expression of G2/M and S phase markers after which, principal component analysis (PCA) was used to identify the most variable genes. The PCA plot showed that it was not necessary to regress out any variation due to cell cycle (**Supplementary figure 3**), however based on the QC checks described above, any variation due to mitochondrial contamination was regressed out. The filtered dataset contains the higher-quality nuclei which were then used for downstream analysis, specifically the clustering analysis and cell type identification.

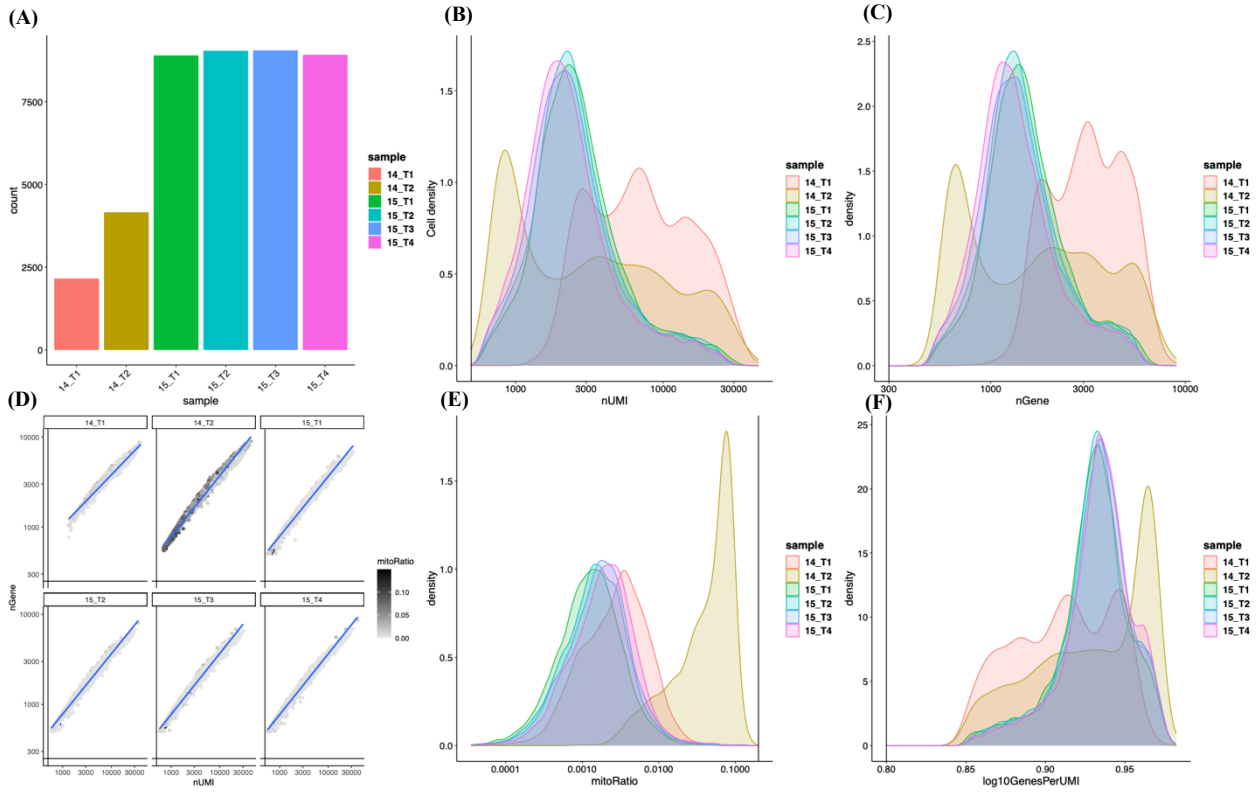


Figure 3.2: ***Visualization of QC metrics after filtering of snRNA-seq datasets.*** **(A)** QC plots showing total number of nuclei per technical replicate, **(B)** UMI counts per nuclei, **(C)** distribution of genes detected per nuclei, **(D)** correlation between genes detected and number of UMIs, **(E)** distribution of mitochondrial gene expression detected per nuclei and **(F)** genes detected per UMI .

3.2.5 Integration analysis of merged datasets

After the filtering and normalisation steps, the merged datasets were integrated. While normalisation regresses out biological effects due to mitochondrial expression, the normalised data can still contain unwanted variability. Therefore, it is also important to perform data correction or integration which can correct for additional variability between samples, such as batch effects. Seurat was used to perform integration to remove the unwanted variation, allowing the resulting nuclei clusters to be annotated more easily. Integration ensured that the cell types of the 14-year-old replicates align with the same cell types of the 15-year-old replicates. To confirm that the data needed to be integrated Uniform Manifold Approximation and Projection (UMAP) was performed to visualise the unintegrated (**Figure 3.3A**) and integrated data (**Figure 3.3B**). This analysis reveals an improvement in sample alignment after integration was performed.

3.2.6 Clustering of nuclei

After integration, the next step is to perform clustering. This involves the separation of the various cell types into distinct clusters. The number of PCs to include is required for the clustering analysis because *Seurat* assigns nuclei to the different clusters based on their PCA score. An elbow plot was used to determine the optimal number of PCs to use (**Supplementary figure 4**). For the clustering analysis, 20 PCs were used along with a resolution of 0.4, resulting in a total of 23 clusters **(Figure 3.3C**). The resulting clusters were assessed for each replicate individually to determine if there are any sample-specific clusters (**Figure 3.3D**; **Supplementary table 2**). The majority of nuclei in cluster 6 were from 14-year-old T2 (30.86 %) and 15-year-old T3 (27.75 %) combined compared to the other replicates which made up $1.07\,\% - 13.68\,\%$ of this cluster (arrows in **Figure 3.3D**). A similar result is seen for cluster 8 (**Supplementary table 2**), where the majority of nuclei were from 14-year-old T2 (arrowheads in **Figure 3.3D)**. Clusters 0, 1, 2, 4, 5, 12, 17 and 19 were largely made up of nuclei from the 15-year-old samples compared to the 14-year-old samples (**Supplementary table 2**, **Figure 3.3D**). This difference in cluster composition is likely due to the fact that a greater number of nuclei were captured for the 15-year-old replicates compared to the 14-year-old replicates (**Table 3.2**). This could also be because of the lower quality of the 14-year-old sample. In addition, cluster 18 was not found in 14 T1 but was present in all the other replicates (**Figure 3.3D, Supplementary table 2**).

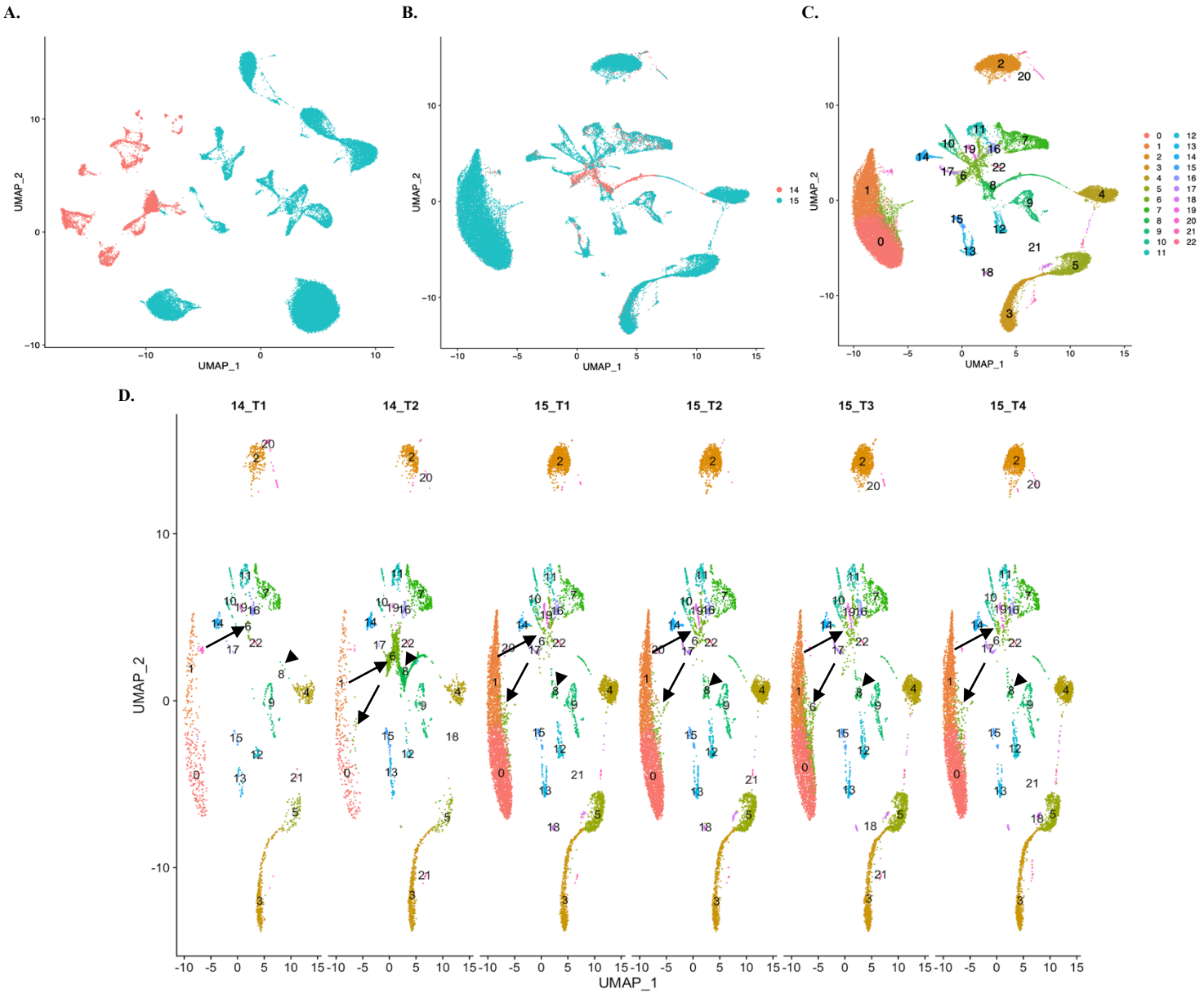


Figure 3.3: ***Clustering analysis of the nuclei before and after integration*****. (A, B)** Sample-specific clustering of nuclei after Uniform Manifold Approximation and Projection (UMAP) before (**A**), and after integration **(B)**, coloured by sample. **(C, D)** Clustering analysis of the integrated samples **(C)** and the clustering split by replicate **(D).** Each cluster is represented by a different colour and number. Each dot represents an individual nucleus. Black arrows indicate cluster 6 and black arrowheads indicate cluster 8.

3.2.7 Annotation of nuclei clusters

After grouping the nuclei into distinct clusters, the cellular identity of these clusters was determined using a combination of manual and automated annotation methods. The automated cell type assignment tool, *SCSA,* was used to assign cell type identities to the distinct clusters (**Supplementary table 3**) using a list of marker genes for each cluster (Cao, Wang and Peng, 2020). Manual annotation was performed to verify the identity of the clusters from the automated annotation and determine the identity of the 'unknown' clusters. For this analysis, the top 10 data-driven markers for each cluster were compared with marker genes from the literature (Lake *et al.*, 2018; Song *et al.*, 2021). Using both annotation approaches and the UMAP visualization tool, the 23 clusters were assigned to one of the major cell types of the brain: inhibitory neurons, excitatory neurons, OPCs, astrocytes, oligodendrocytes, endothelial cells, and microglia cells (**Figures 3.4-3.5**; **Table 3.3**). Clusters 7, 10, 11, 14, 16, 19, 22 express excitatory neuron markers *CUX2* and/or *DLGAP2.* Clusters 9 and 12 express inhibitory neuron markers *GAD1* and *GRIK1* (**Figure 3.4A**). Cluster 4 expressed classic OPC marker *PCDH15* and *PDGFRA*, while astrocyte markers *SLC1A2* and *AQP4* were expressed by clusters 3, 5, 15 and 21 (**Figure 3.4B**). Cluster 4 also expressed *GRIK1* but at a lower level than clusters 9 and 12. This has been shown previously, where OPCs express *GRIK1* (Allen brain Institute, https://celltypes.brain-map.org/rnaseq/human_m1_10x?selectedVisualization=Heatmap&colorByFeature=Cell+Type&colorByFeatureValue=GAD1). Clusters 0 and 1 both express classic oligodendrocyte markers *MBP, MOBP* and *PLP1* (**Figure 3.4C**). In addition, cluster 13 expresses *FLT1*, a classic endothelial marker (**Figure 3.4C**). Lastly, *APBB1IP*, a classic microglia marker is most highly expressed by clusters 2 and 20 (**Figure 3.4C**). After performing both manual and automatic annotations, there were still two unknown clusters, 17 and 18 (**Table 3.3**). Cluster 17 did not highly express any of the classical markers shown in **Figure 3.4** and it was also classified as an "unknown" cluster by *SCSA* (**Supplementary table 3**), therefore, the final annotation of this cluster was "unknown". Cluster 18 expressed inhibitory neuron, OPC and astrocyte markers and was classified as an astrocyte cluster by *SCSA* (**Supplementary table 3**). This made it difficult to annotate this cluster, so the final annotation of this cluster was "unknown". This cluster might be made up of doublets or multiplets from various cell types, which again, made it difficult to annotate.

In addition, clusters expressing high levels of mitochondrial genes were identified (**Supplementary figure 5**). The mitochondrial genes *MT-ND2* and *MT-CO2* were expressed in all the clusters (**Supplementary figure 5**). Cluster 6 and Cluster 8 showed the highest expression for all four mitochondrial genes **Supplementary figure 5**). The expression of mitochondrial genes may indicate a large amount of cell debris or dying cells, generated during the nuclei isolation step, and as a result these clusters were annotated

as cell debris clusters. The proportion of the cell types in each sample was also determined (**Figure 3.6**). The 14-year-old samples contained a larger proportion of excitatory neurons than the 15-year-old samples. As expected, 14-year-old T2 displayed a higher proportion of nuclei classified as cell debris compared to the other samples (**Figure 3.6**). Overall, the 14-year-old technical replicates showed low similarity in terms of the proportion of cell types, while the 15-year-old technical showed similar cell type proportions.

Table 3.3: Cell type assignment for each cluster after performing manual and automatic annotation.

Clusters	Cell type
9, 12	Inhibitory neuron
7, 10, 11, 14, 16, 19, 22	Excitatory neuron
4	OPC
3, 5, 15, 21	Astrocyte
0, 1	Oligodendrocyte
13	Endothelial cell
2, 20	Microglia
17, 18	Unknown
6, 8	Cell debris

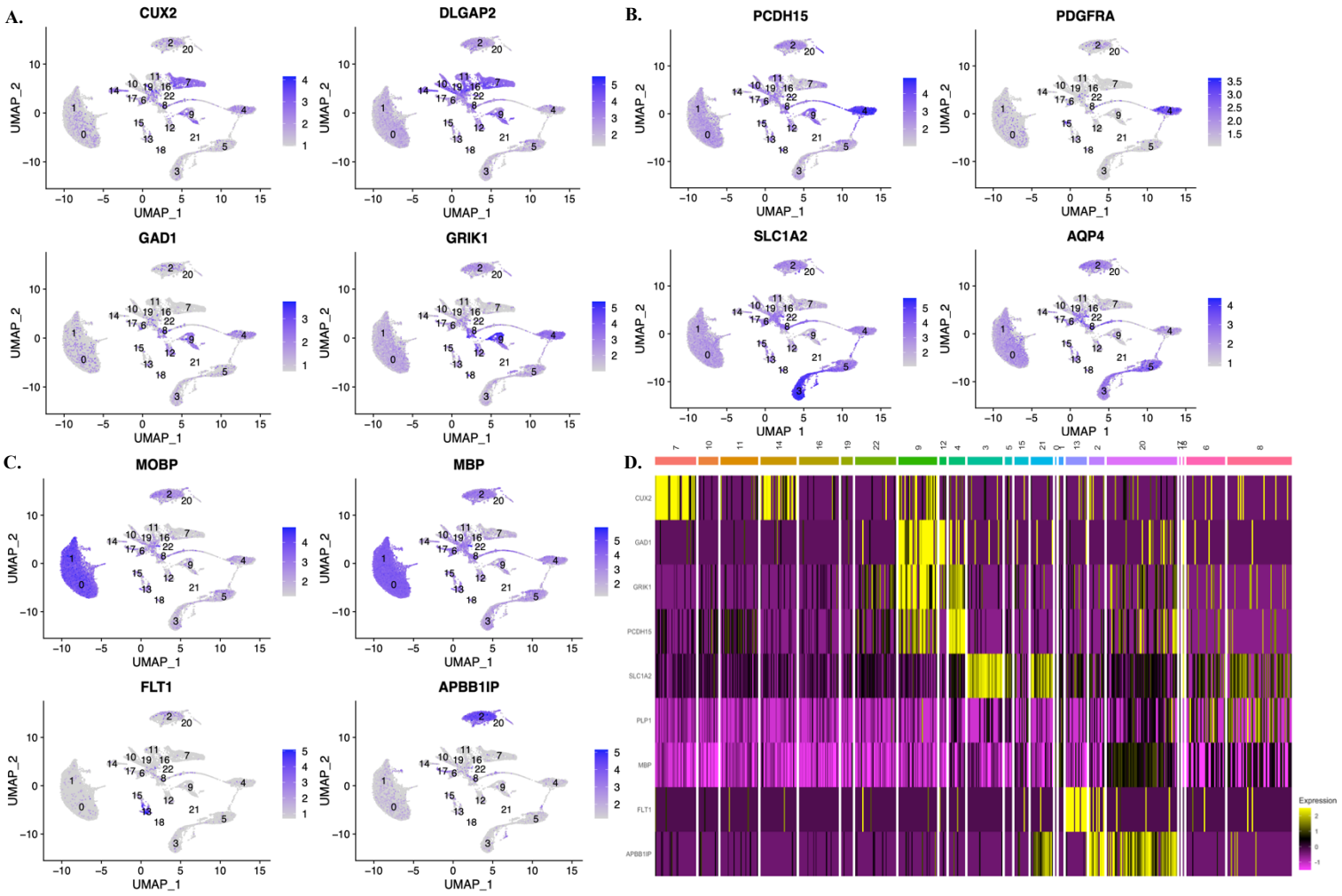


Figure 3.4: ***Classification of cell types of the integrated dataset.*** Feature plots generated in Seurat showing the differentially expressed genes or classic cell-type specific markers for the different cell types. **(A)** shows the expression of: *CUX2* (excitatory neurons), *DLGAP2* (excitatory neurons), *GAD1* (inhibitory neurons) and *GRIK1* (inhibitory neurons), **(B)** shows the expression of: *PCDH15* (OPC), *PDGFRA* (OPC), *SLC1A2* (astrocytes) and *AQP4* (astrocytes). **(C)** shows the expression of: *MOBP* (oligodendrocytes), *MBP* (oligodendrocytes), *FLT1* (endothelial) and *APBB1IP* (microglia). Nuclei are colored purple if the markers are highly expressed and grey if there is no expression. **(D)** Heatmap showing the expression of a selection of the cell-type specific markers shown in the feature plots. Individual clusters are indicated at the top with the cell-type specific marker genes shown on the left. Columns represent individual cells. Pink – purple, no expression; black, low expression; yellow, high expression.

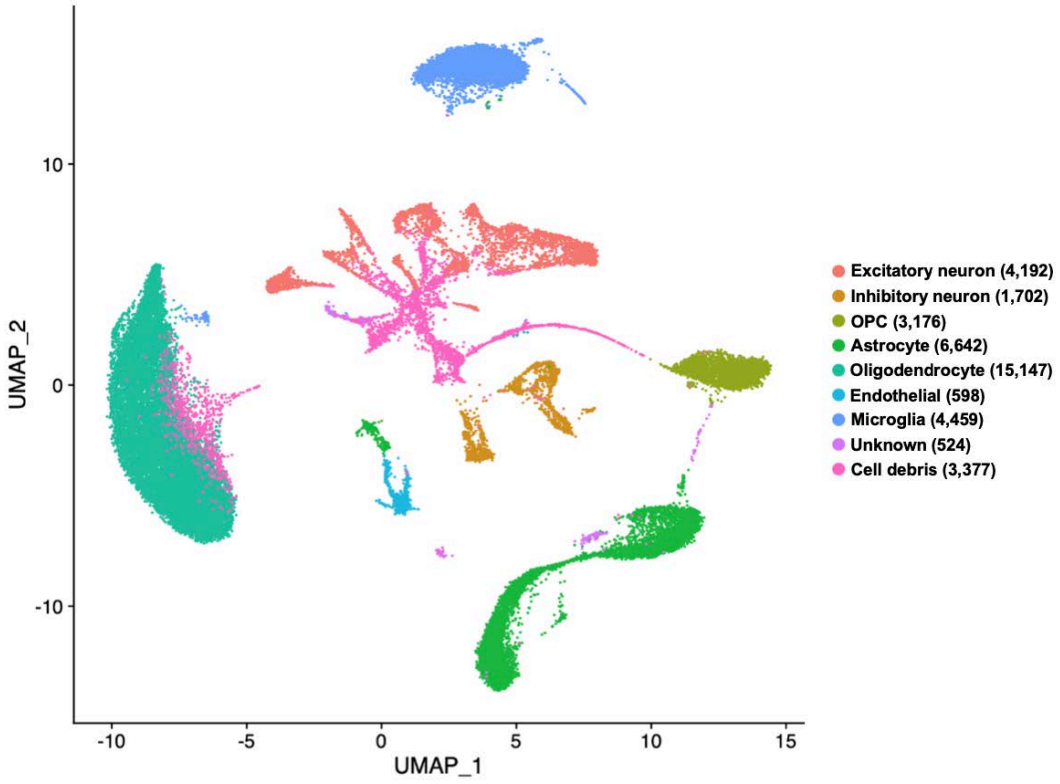


Figure 3.5: ***Classification and visualization of cell types of the integrated dataset.*** UMAP visualization shows the classification of the clusters showing the 7 major cell types of the brain using automated and manual annotation methods. The unknown and cell debris clusters are also shown. Each cell type is represented by a different colour and the total number of nuclei per cluster is shown in brackets.

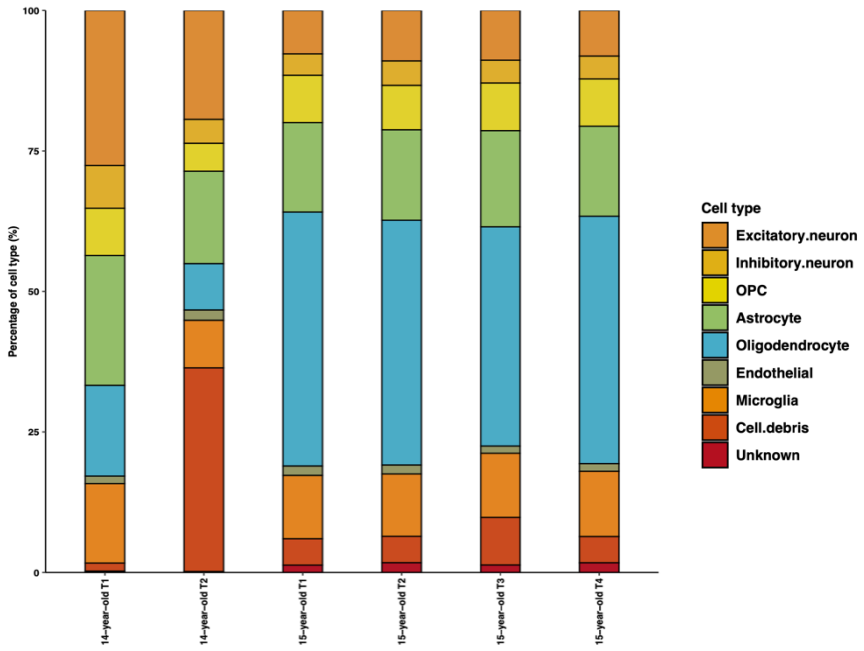


Figure 3.6: ***The proportion of each cell type, cell debris and unknown cluster in each replicate.***

3.2.8 Differential expression (DE) analysis

Previous studies have shown that there are hundreds of distinct cell types in the brain, and it is possible that the gene expression profiles of these different cell types may differ by brain region (Kang *et al.*, 2011; Lake *et al.*, 2018; Li *et al.*, 2018). The current analysis consists of six datasets, two from the frontal cortex and four from the temporal cortex from two patients of similar age. Using *DESeq2* (Love, Huber and Anders, 2014), differential gene expression analysis was performed, to determine if there were any genes that showed significant changes in expression between the two brain regions (temporal and frontal) for each of the annotated cell types. This analysis was performed as a pilot test of the DE pipeline and in the future, when additional snRNA-seq libraries are generated, this pipeline can be applied to samples of different ages.

First, PCA was performed on the gene expression count data for each of the annotated cell-types to confirm that the technical replicate samples for each brain region were correlated with one another (**Supplementary figure 6)**. The technical replicates clustered together, for all the cell types, and there was a clear separation between the samples from different brain regions on PC1 which accounts for the > 92 % of the variance between samples (**Supplementary figure 6**). There was also a small amount of separation between the 14-year-old technical replicates on PC2 and this could be indicative of the quality of these replicates (especially 14 T2).

DESeq2 was then used to assess changes in gene expression between the two brain regions for the seven distinct cell types annotated above (**Supplementary file 10**). The excitatory neurons showed the greatest number of DE genes (8 543), and the endothelial cells had the lowest number of DE genes (1 505) compared to the other cell types (**Figure 3.7**). In addition, the astrocyte (8 019) and oligodendrocyte (7 647) cells also showed a large number of DE genes. The total number of genes enriched for each brain region was 19 886 for the temporal lobe and 17 812 for the frontal lobe. This analysis revealed several interesting genes enriched in the temporal lobe when compared to the frontal lobe (**Figure 3.7**). *MEG3* or maternally expressed 3, a lncRNA was enriched in the temporal lobe for both neuronal groups, OPCs, as well as the astrocyte, endothelial and microglia groups (**Figure 3.7**). This gene has been shown to function as a tumor suppressor gene (Zhou, Zhang and Klibanski, 2012). *MEG3* has also been shown to be involved in the regulation of neuronal synapse plasticity (Tan *et al.*, 2017). *KCNMB2* was enriched in the temporal lobe for the excitatory neurons, inhibitory neurons and OPCs (**Figure 3.7**). This is in line with previous studies that have shown *KCNMB2* to be enriched in OPCs and inhibitory neurons (Sjöstedt *et al.*, 2020) (https://www.proteinatlas.org/ENSG00000197584-KCNMB2).The *KCNMB2* gene encodes for large

conductance Ca^{2+}- and voltage-activated K^+ channels, which have been shown to play important roles in regulating smooth muscle tone and neuronal excitability (Hu *et al.*, 2001; Ko *et al.*, 2008). Furthermore, previous studies have also demonstrated that these channels are involved in the regulation of learning and memory processing (Typlt *et al.*, 2013). Another gene encoding a potassium channel, *KCNAB1* was enriched in the frontal lobe for both neuronal groups and OPCs (**Figure 3.7**). Previous studies have shown *KCNAB1* to be enriched in neuronal cells and OPCs (https://www.proteinatlas.org/ENSG00000169282-KCNAB1). *KCNAB1* was also previously identified as a susceptibility gene for lateral temporal epilepsy (LTE) (Busolin *et al.*, 2011). Interestingly, there was one gene unique to astrocytes that was enriched in the frontal lobe, *RNF219-AS1* (also known as *OBI1-AS1*) (https://www.genecards.org/). Not much is known about the function of *RNF219-AS1*, but several studies have shown this gene to be associated to ADHD (Fu *et al.*, 2021).

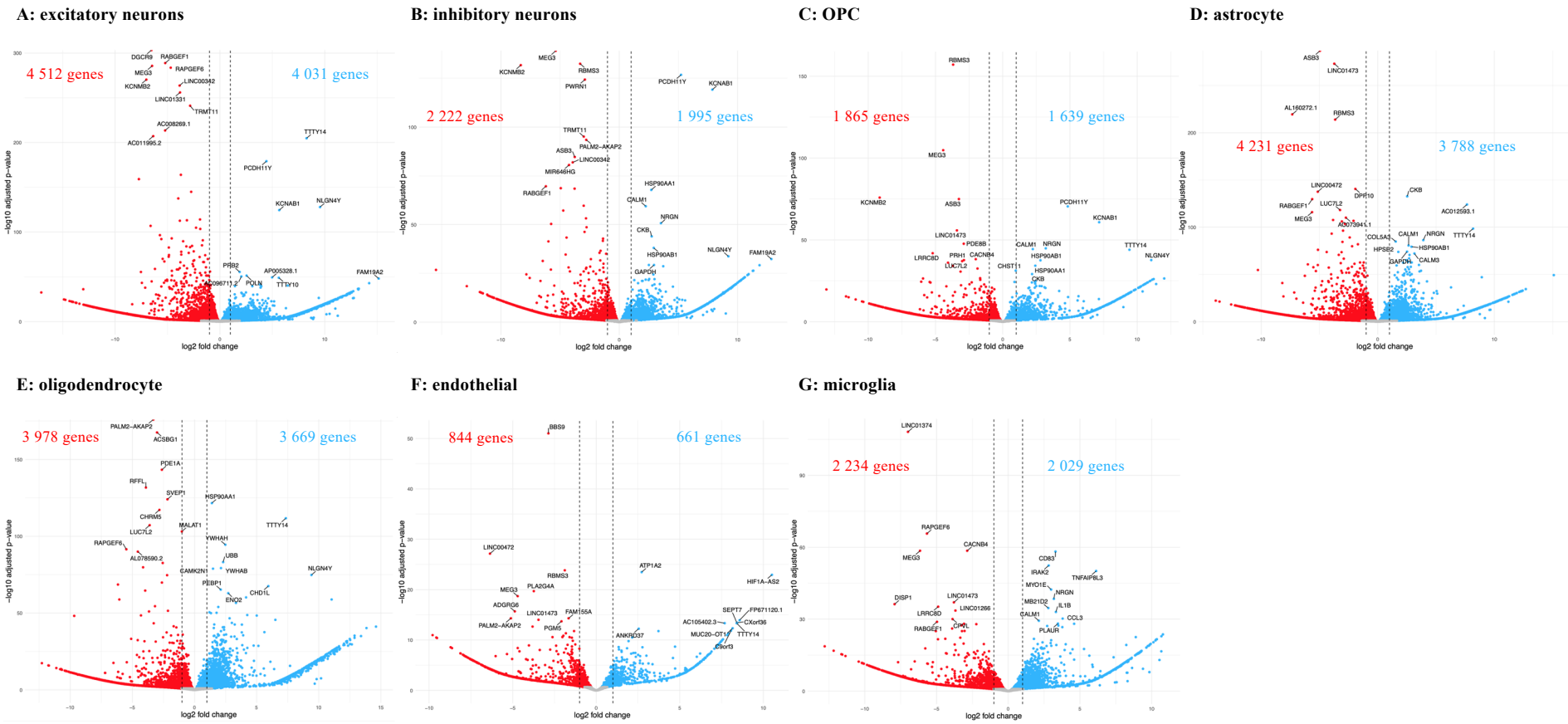


Figure 3.7: ***Volcano plots*** **showing the** ***top 10 most significant genes for each brain region and cell type (A-G).*** **(A)** excitatory neurons, **(B)** inhibitory neurons, **(C)** OPCs, **(D)** astrocytes, **(E)** oligodendrocytes, **(F)** endothelial cells **(G)** microglia. Adjusted p-value < 0.05, blue, enriched in the frontal lobe, red, enriched in the temporal lobe (by adjusted p value). Dashed lines show log2 fold change = 1/-1. The top 10 enriched genes (by adjusted p Value) for each brain region are labeled.

3.2.9 Functional enrichment analysis

After running *DESeq2* and obtaining the list of significant differentially expressed genes (DEGs) for each cell type, functional enrichment analysis was performed to determine if there was any enrichment of known biological functions, interactions, or pathways. GO, KEGG and DO analysis was performed on the DEGs for each cell type and brain region. These analyses highlighted terms associated with neurodevelopment and neurodevelopmental diseases, however term enrichment was not often significant (i.e., p.adjust < 0.05), which is likely due to the small sample size (**Figure 3.8**).

The enriched GO terms were similar across all cell types for the frontal lobe compared to the temporal lobe, where the enriched GO terms were more distinct for each cell type (**Figure 3.8A**). A similar result was seen for the KEGG enrichment analysis (**Figure 3.8B**). There were no enriched GO terms identified for the endothelial cluster for the temporal lobe (**Figure 3.8Ai**). There were two GO terms that were shared across the temporal and frontal brain regions: modulation of chemical synaptic transmission and regulation of trans-synaptic signaling (**Figure 3.8**). Terms associated with nervous system development and function such as neuron migration, glutamatergic synaptic transmission, regulation of postsynaptic membrane potential and synapse organization were enriched in excitatory neurons, inhibitory neurons and OPCs for the temporal lobe (**Figure 3.8Ai**). In addition, terms associated with processes such as phagocytosis and macroautophagy were enriched in microglia (**Figure 3.8Ai**). Interestingly, DO terms associated with neurodevelopmental disorders such as autistic spectrum disorder, were only enriched in excitatory neurons, inhibitory neurons and OPCs in the temporal lobe (**Figure 3.8Ci**). For the frontal lobe, DO terms associated with neurological disorders such as epilepsy syndrome, Alzheimer's disease and focal epilepsy were enriched in all cell types except for the endothelial group (**Figure 3.8Cii**). Terms associated with cancer such as neuroblastoma, connective tissue cancer, prostate cancer and cervix carcinoma were mainly enriched in the endothelial group (**Figure 3.8Cii**). To note, the terms enriched in the temporal and frontal lobe are not significant as none of them have a p value < 0.05.

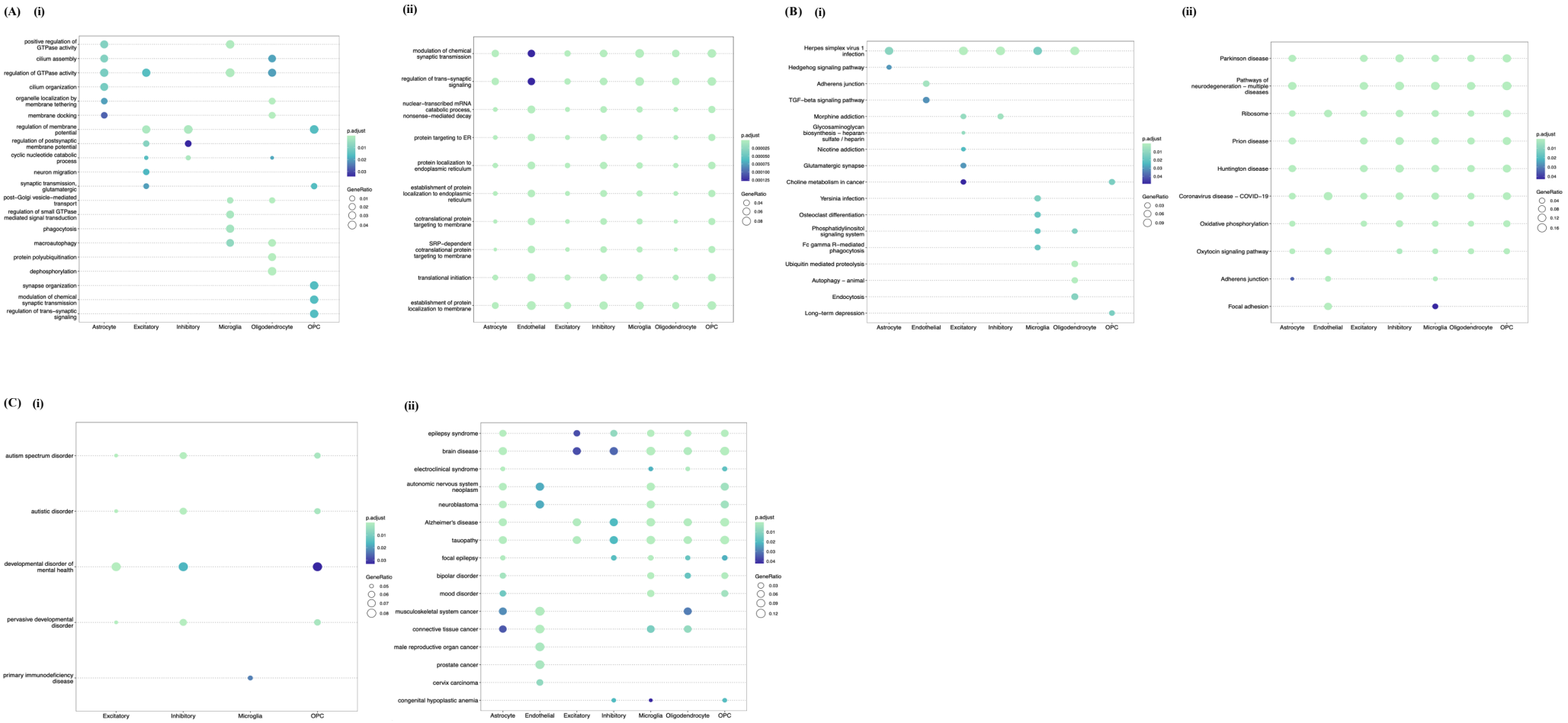


Figure 3.8: ***Enrichment analysis on significant DEGs from each cell type.*** **(A)** GO pathway enrichment analysis results for **(i)** the temporal lobe and **(ii)** the frontal lobe. **(B)** shows the KEGG pathway enrichment analysis results for **(i)** the temporal lobe and **(ii)** the frontal lobe. **(C)** shows the DO pathway enrichment analysis results for **(i)** the temporal lobe and **(ii)** the frontal lobe, represented as dot plots. Each dot represents the indicated enrichment term, with green indicating more significantly enriched terms. The size of each dot corresponds to the GeneRatio (number of genes belonging to a particular cell type/ total number of genes).

3.3 Filtering and lysis condition optimisations for future analyses

One of the sources of technical variation in snRNA-seq analysis comes from sample preparation which affects the sample quality and interpretation of results (Tung *et al.*, 2017). Because of this, it was important to conduct several nuclei isolation experiments, testing previously published protocols (Habib *et al.*, 2017; 10x, 2019) to determine the optimal method which would result in a single nuclei suspension with little to no cell debris and cell clumping. These methods are very similar; however, they can differ in terms of the lysis buffer used, size of cell strainer or the use of additional filtering steps. It was also important to conduct preliminary tests to determine the nuclei concentration as the 10x Genomics protocol recommends $1\,000\ nuclei/\mu l$ as the starting material (10x, 2019). The results from the 14-year-old and 15-year-old snRNA-seq experiments described above, showed that these libraries may have been contaminated by cell debris or dead cells. In order to reduce the amount of cell debris and dead cells in nuclei isolation for future snRNA-seq experiments, further nuclei isolation optimization experiments were conducted.

Nuclei isolation trial 1 and 2: Assessing nuclei quality

Firstly, the nuclei isolation protocol was repeated with no changes. After each lysis and centrifugation step, an aliquot of nuclei was taken and stained with trypan blue to check the quality of the nuclei and amount of cell debris. The nuclei isolation test was performed on 6-year-old frozen tissue from the frontal cortex and 46-year-old frozen tissue from the frontal cortex. This was done to see if there were any differences in nuclei concentration and cell debris between the pediatric and adult tissue samples. The steps of the protocol and summary of the results are outlined in **Table 3.4**. The concentration of the nuclei from the 6-year-old sample was $375\ nuclei/\mu l$ and $1\,200\ nuclei/\mu l$ from the 46-year-old sample. These results showed that there is a difference in the final concentration and amount of cell debris between the pediatric and adult tissue with the adult tissue yielding a higher concentration of nuclei (**Figure 3.9**) (**Table 3.4**). The pediatric tissue had less cell debris, however, there was a very low final nuclei concentration compared to the adult tissue.

Nuclei isolation trial 3: Gaublomme et al. 2019 protocol

The Gaublomme *et al.* nuclei isolation protocol uses a different lysis buffer, different cell filters and shorter lysis time compared to the previously used protocol. This protocol was tested on frozen brain tissue and the results were compared between the different methods to see if it would reduce the amount of cell debris.

The nuclei isolation protocol was performed on the 46-year-old frozen tissue from the frontal cortex and is summarized in **Table 3.4**. The final concentration of nuclei from the 46-year-old sample was 4 640 $nuclei/\mu l$. The microscope images showed that there was a significant amount of cell debris, clumping of the nuclei and after the second filtering step, some of the nuclei started blebbing or had burst (**Figure 3.9**). These results showed a greater amount of cell debris compared to the previous isolation trial results above and because of this, the standard nuclei isolation protocol was used for future experiments. In addition, passing the nuclei through two cell strainers, did not improve or reduce the amount of cell debris and clumping.

Nuclei isolation trial 4: Standard nuclei isolation protocol with additional myelin removal step (Allen Brain institute 2019 protocol)

Finally, an additional myelin removal step was tested with the standard nuclei isolation protocol (Allen brain institute, https://dx.doi.org/10.17504/protocols.io.y6rfzd6). This was done to determine if a myelin removal step would reduce the amount of cell debris in the final nuclei preparation. This protocol was performed on the 15-year-old frozen tissue from the temporal cortex and 31-year-old frozen tissue from the temporal cortex to determine if there were differences in nuclei concentration and cell debris between the pediatric and adult tissue samples. The final concentration of nuclei from the 15-year-old sample was 445 $nuclei/\mu l$ and the final concentration of nuclei from the 31-year-old was 755 $nuclei/\mu l$. The myelin removal step resulted in a large reduction in the amount of cell debris step compared to previously used protocols (**Table 3.4**) (**Figure 3.9**). There was also a decrease in the nuclei concentration for both pediatric and adult samples after the myelin removal, however, the concentrations were still within range of the suggested concentrations in the 10x Genomics protocol.

snRNA-seq datasets from the 14-year-old and 15-year-old tissue samples were successfully generated using the 10x Genomics platform. All the major cell types of the brain were also identified, with oligodendrocytes making up the largest proportion of the dataset. A pilot DE analysis step was performed, comparing the gene expression changes that occurred between the temporal and frontal lobes. This analysis identified several important differentially expressed genes and functionally enriched terms that are relevant to neurodevelopment. Nuclei isolation tests were also performed and a protocol that resulted in a cleaner nuclei preparation was successfully generated.

Table 3.4: Table detailing results from each nuclei isolation trial. Trial 1 and 2: describes the quality of the nuclei from the 6-year-old and 46-year-old samples after each step of the adapted nuclei isolation protocol outlined in the Methods. The nuclei quality and presence of cell debris for each step in the protocol are described below. **Trial 3:** describes the quality of the nuclei from the 46-year- old sample after specific steps of the Guablomme et al. 2019 protocol. **Trial 4 and 5:** describes the quality of the nuclei and presence of cell debris from the 15-year-old sample and 31-year-old sample at specific steps of the adapted nuclei isolation protocol with the myelin removal step as outlined in the Methods. +++, large amount of cell debris and clumping, ++, presence of cell debris with very little clumping, + little to no cell debris and no clumping.

	Adult sample		**Pediatric sample**	
Protocol step	debris level	nuclei status	debris level	nuclei status
Trial 1-2				
After homogenization	+++	intact	+++	intact
After first lysis	+++	intact	+++	intact
After centrifugation	+++	intact	+++	intact
After second lysis	++	intact	++	intact
After second centrifugation	+	intact	++	intact
Trial 3				
After homogenization	+++	intact		
After first lysis	+++	intact		
Before filter	+++	intact, blebbing		
After filter	+++	intact, blebbing		
Trial 4-5				
Before myelin removal	+++	intact	+++	intact
After myelin removal	+	intact	+	intact
After second centrifugation	+	intact	+	intact

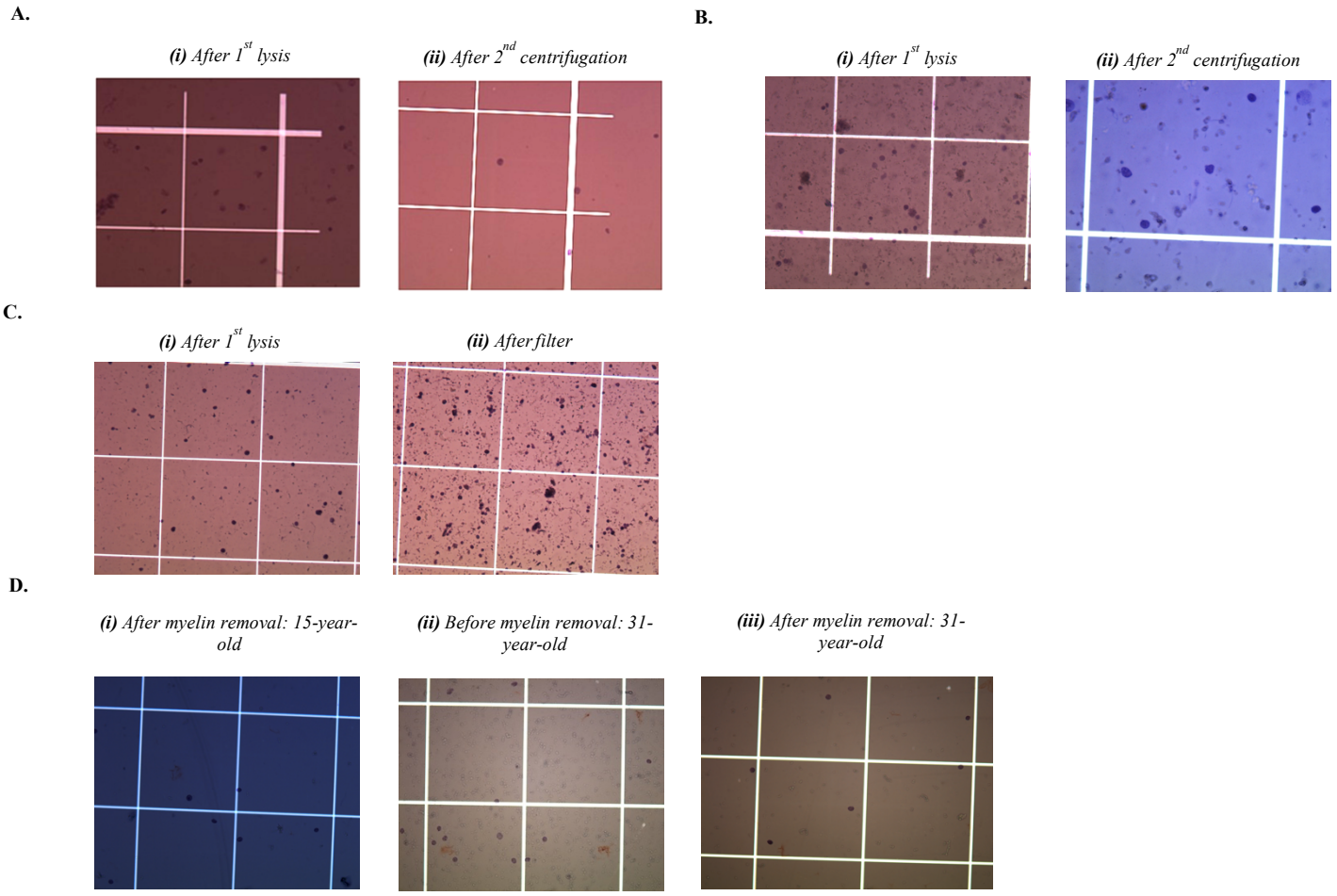


Figure 3.9: ***Microscope images showing the trypan blue stained nuclei for nuclei isolation trials.*** **(A)** shows the trypan blue stained nuclei from the 6-year-old and **(B)** 46-year-old samples for trials 1 and 2. **(C)** shows the trypan blue stained nuclei from the 46-year-old sample for trial 3. **(D)** shows the trypan blue stained nuclei from the 15-year-old and 31-year-old samples for trials 4 and 5.

3.4 ATAC-seq on human brain tissue

3.4.1 Analysing the dynamics of chromatin accessibility in pediatric and adult brain tissue using bulk ATAC-seq

ATAC-seq was performed on seven pediatric and three adult samples, using either fresh, frozen, or cryopreserved tissue (see **Table 2.1**). Prior to sequencing, the quality of each library was checked using the Agilent Bioanalyzer chip to assess the fragment size distribution (**Figure 3.10A**, **Supplementary figure 7)**. The cDNA fragment size distribution showed several peaks that increase in size, demonstrating the characteristic nucleosome periodicity of approximately $150 - 200\ bp$ (Buenrostro *et al.*, 2013). Most of the samples showed this fragment size distribution, indicating good quality libraries. The libraries were then sent for sequencing, generating Fastq files (**Supplementary file 11**).

3.4.2 ATAC-seq library quality control: pre- and post-alignment

The quality of the sequencing reads, and the presence of sequencing adapters was assessed using *FastQC*. The FASTQC reports (**Supplementary file 11**) visualize base quality scores, sequence length distribution, GC content, adapter content and sequence duplication levels. These results showed the presence of adapters, which were subsequently removed using *NGmerge* and the trimmed reads were re-assessed with *FastQC*. *FastQC* analysis showed that the samples were of sufficient quality and that the low-quality or adapter contaminated reads were successfully removed (**Supplementary file 11**).

After running *FastQC*, the trimmed reads were aligned to the human reference genome, GrCh38, using *Bowtie2* to generate BAM files. The alignment statistics (**Supplementary table 4)** show the total number of reads and unique alignment rate for each sample. Previous studies suggest that a unique mapping rate of approximately 80 % is standard for a successful ATAC-seq experiment (Yan *et al.*, 2020). An average of 82.29 % mappability and 23,255,805 million reads were obtained across the samples. Most of the samples show good alignment rates, with some exceptions: 23-month-old replicate 1 (68.25 %), both 31-year-old replicates (68.3 %), and 46-year-old replicate 1 (70.83 %) (**Supplementary table 4**).

After running *Bowtie2*, the sequence insert size metrics were assessed, using *Picard* tools. The sequencing insert size distribution shows a number of peaks which decrease in height (i.e., abundance) as the size of

the insert sequence increases (**Figure 3.10B, Supplementary figure 8**). The first large peak (around 50 bp) corresponds to highly fragmented open chromatin, the second peak (around 200 bp) corresponds where Tn5 inserted around a single nucleosome and the third peak (around 400 bp) corresponds to where Tn5 inserted around two adjacent nucleosomes showing more closed chromatin (**Figure 3.10B, Supplementary figure 8**). All the ATAC-seq libraries show a higher number of smaller insert sizes which is expected as the transpose will preferentially cut in regions of open chromatin. Overall, these results show good insert size distribution for all samples and replicates (**Supplementary figure 8**). However, the insert size distribution plots for the 23-month-old sample, indicated that the fresh replicates (T1 and T2) were not the best quality, and this was something to consider when performing the downstream analyses (**Supplementary figure 8**).

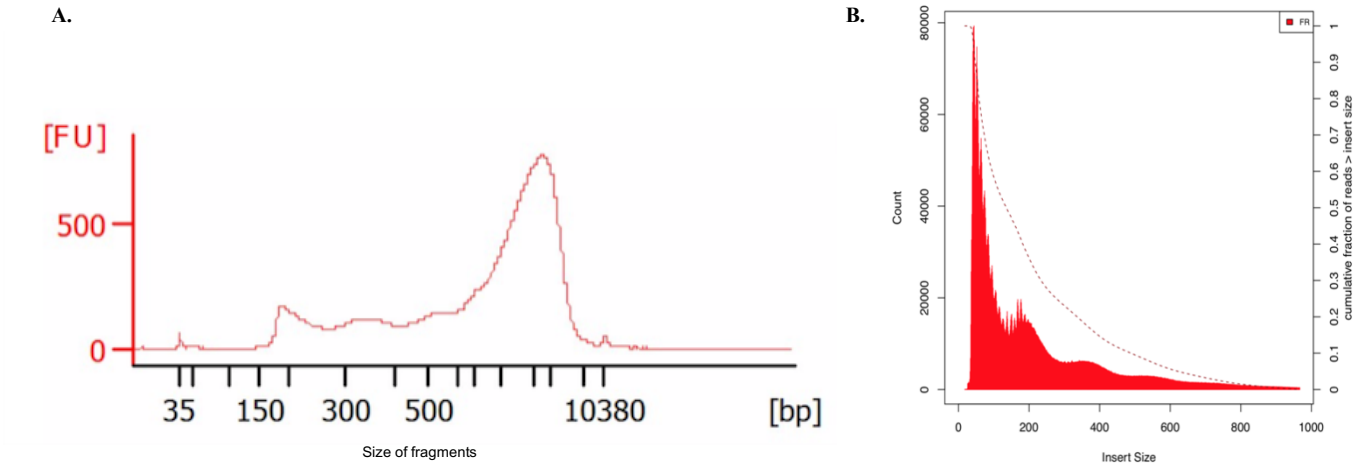


Figure 3.10: ***Quality control plots for the 19-month-old technical replicate 1.*** **(A)** Shows the Agilent Bioanalyzer trace indicating the fragment size distribution in the ATAC-seq library before sequencing was performed. **(B)** Shows the sequence insert size distribution after mapping the sequencing reads to the human genome for the same library using *Bowtie2*. bp, base pairs; FU, fluorescent units.

To further assess the quality of these libraries, they were visualised using the UCSC genome browser (http://genome-euro.ucsc.edu/index.html) (**Figure 3.11**). These coverage tracks were displayed at a specific gene locus, *Glial Fibrillary Acidic Protein* (*GFAP*), which is used as a marker for astrocytes (Baba *et al.*, 1997). This is a useful way of assessing the quality of the ATAC-seq datasets, as it allows the visualization of sequencing tracks at the location of genes of interest and facilitates visual comparisons with publicly available track sets, such as the GeneHancer track set (Fishilevich *et al.*, 2017) and the ENCODE regulation tracks (ENCODE Project, 2012; Moore *et al.*, 2020). The GeneHancer database is an extensive database

comprised of the genomic locations of putative regulatory elements, including transcription start sites, promoters and enhancers, as well as their target genes (Fishilevich *et al.*, 2017). The ENCODE regulation tracks represent histone modification data for seven human cell lines. The layered H3K4Me3 track shows a histone mark associated with promoters and the layered H3K4Me1 and H3K27Ac tracks show histone marks commonly associated with enhancers, while H3K4Me1 has also been associated with poised promoters (ENCODE Project, 2012; Bae and Lesch, 2020). These results showed that all the samples were of good quality as they all displayed sequence pileups in the promoter region of the *GFAP* gene (highlighted in blue) which is known to be expressed in the brain (**Figure 3.11**). This region was annotated as the GFAP TSS in the GeneHancer database and the sequence pile-up over the TSS is associated with H3K4Me1 signal (**Figure 3.11**). The 23-month-old sample was not removed from the final analyses because it showed sequence pileups in this expected region (**Figure 3.11**).

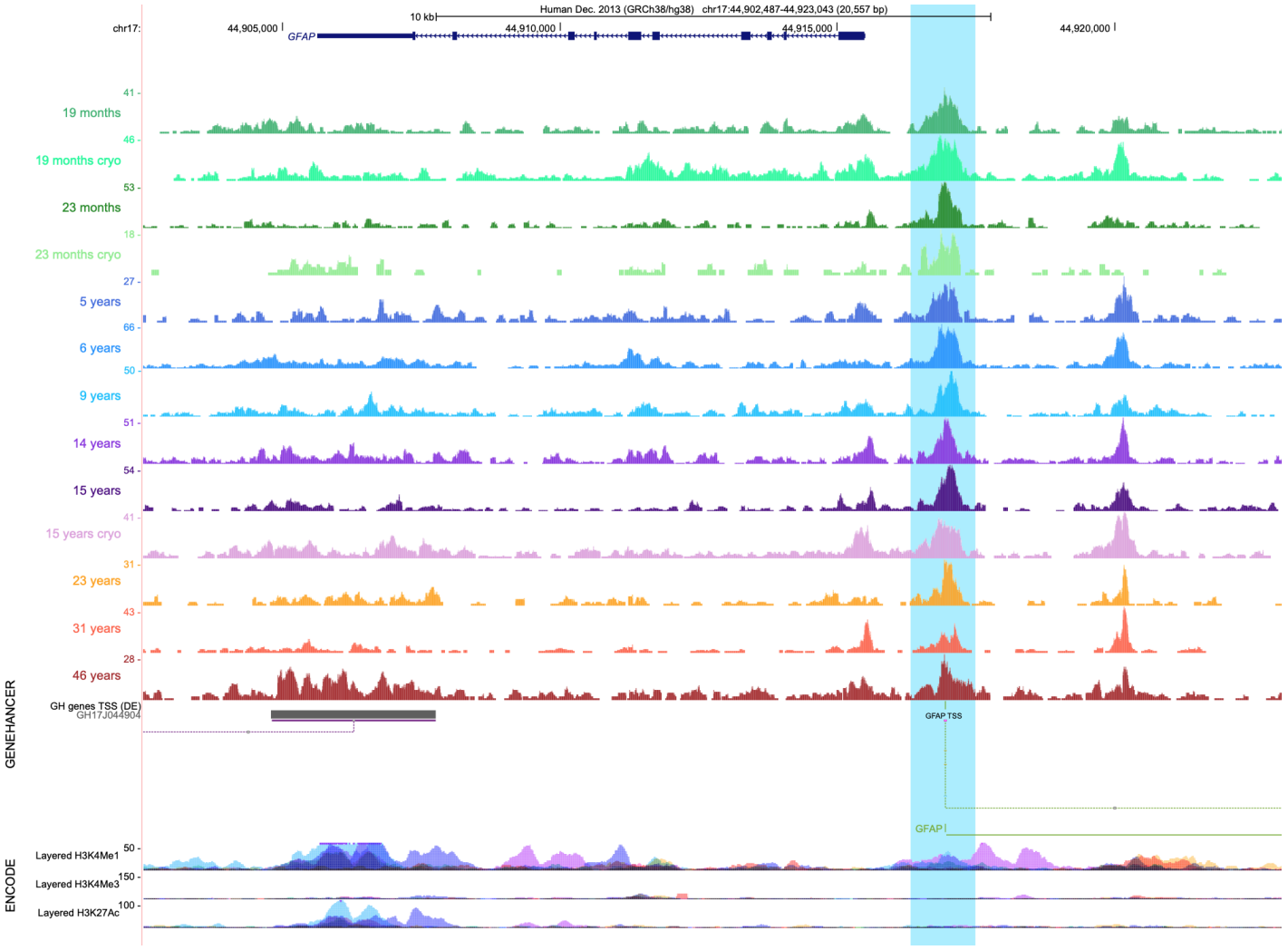


Figure 3.11: ***USCSC Genome Browser view of open chromatin tracks for each sample at the GFAP gene locus.*** Colour *coded* ATAC-seq coverage tracks from the 19-month-old, 23-month-old, 5-year-old, 6-year-old, 9-year-old, 14-year-old, 15-year-old, 23-year-old, 31-year-old, and 46-year-old are shown below the *GFAP* transcript annotation. Below the coverage tracks is the GeneHancer track set (Fishilevich *et al.*, 2017) detailing the regulatory elements and lastly the chromatin modification tracks: Layered H3K4Me1, Layered H3K4Me3 and Layered H3K27Ac (ENCODE project, 2012). The *GFAP* promoter region is highlighted in blue.

3.4.3 Preprocessing and filtering of aligned ATAC-seq data using the Reske *et. al* and Harvard FAS Informatics pipelines

To date, there are very few benchmark studies comparing different bioinformatic tools used to analyse ATAC-seq data. Here, two different peak-calling pipelines, developed by Reske *et. al* and Harvard FAS Informatics respectively, were tested to determine which set of tools are best suited for studying human brain tissue. These pipelines used the *MACS2* and *Genrich* peak-calling tools respectively.

Before peak-calling, the post-alignment filtering steps (see **Figure 2.9**) for both pipelines were performed to remove any reads of low mapping quality and those that were improperly paired. These steps also included removal of mitochondrial reads, PCR duplicates and ENCODE blacklisted regions. Lastly, the complexity of each library was determined using *preseqR* and *ATACQC*. The library with the lowest complexity value was 46-year-old T1, therefore all samples were normalised to match this sample.

3.4.4 Identification of significantly open chromatin regions (OCRs) using two different approaches

Peak-calling was performed on the filtered files for each sample using the Reske *et al.* pipeline (*MACS2*) and the Harvard FAS Informatics pipeline (*Genrich*) to identify significant OCRs (**Figure 3.12; Supplementary table 5**). In all samples, except 5-year-old T2, *Genrich* called more peaks than *MACS2* **(Figure 3.12; Supplementary table 5)**. This difference is likely due to the fact that *Genrich* calls narrow peaks, while *MACS2* was used to calls broad peaks (https://github.com/jsh58/Genrich). For most samples, the technical replicates, and cryopreserved replicates showed a similar number of *MACS2* peaks, except for the 5-year-old and 46-year-old replicates (**Figure 3.12**). Interestingly, there was a higher number of *Genrich* peaks for the cryopreserved samples compared to the fresh samples (**Figure 3.12**). This was not the case for *MACS2*, which yielded similar results for cryopreserved and fresh samples. Furthermore, for both peak-calling methods, there was a greater number of peaks called for the middle (6-year-old and 9-year-old), late (14-year-old and 15-year-old) and adulthood (23-year-old, 31-year-old, and 46-year-old) samples compared to the early childhood (19-month-old, 23-month-old, and 5-year-old) **(Figure 3.12; Supplementary table 5)**.

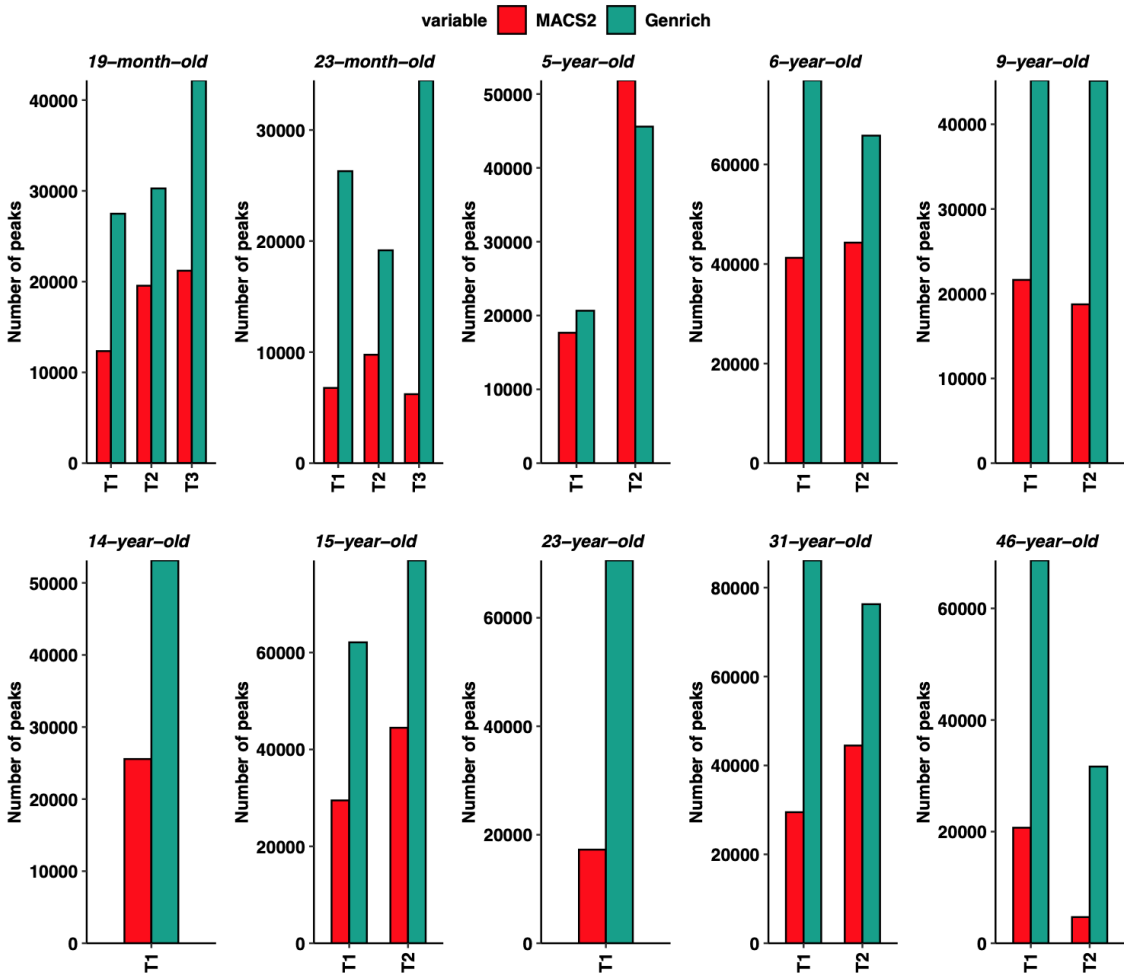


Figure 3.12: ***Total number of peaks called using the Genrich and MACS2 peak-calling methods.*** Bar plots showing the total number of peaks called using the peak caller *MACS2* (red) or *Genrich* (green) for each sample and replicate, where applicable. T1, technical replicate 1, T2, technical replicate 2 and T3, cryopreserved sample.

3.4.5 Peak annotation

After peak-calling with *MACS2* or *Genrich*, the peaks were grouped and analysed in 4 categories: 1) total peaks for each replicate, 2) peaks that were unique to each replicate, 3) peaks that showed any overlap (i.e., at least 1 base pair overlap) between replicates, when present and 4) peaks that showed at least 50 % overlap between replicates, when present, not including the cryopreserved replicates. These groups were annotated using *HOMER* (**Figure 3.13**). When focusing on the total peaks for each replicate, the assigned annotation categories were similar for *MACS2* and *Genrich* peaks, with the majority of peaks being assigned to intronic

and intergenic regions, as expected (**Figure 3.13**). The peaks that were unique to each technical replicate or cryopreserved replicate were mainly assigned to intronic and intergenic regions (unique peaks). When comparing the peaks that were common between the technical replicates for each sample (common peaks and 50 % overlap peaks) to the other peak categories, there is an increase in peaks that were assigned to promoter regions across all the samples (**Figure 3.13)**. The samples which had one replicate (14-year-old and 23-year-old samples) were also annotated and the majority of these peaks were assigned to intronic, and intergenic regions as expected (**Supplementary figure 9**).

The peak-sets that showed at least 50 % overlap between replicates from each sample were merged to generate separate consensus peak sets for each peak-calling approach to be used as a basis for DA analysis. The single peak files for the samples without replicates (14-year-old and 23-year-old) and the cryopreserved replicates (19-month-old and 23-month-old) were also included in the consensus peak-sets. The total number of peaks for the *MACS2* consensus peak-set was 67 046 and the total number of peaks for the *Genrich* consensus peak-set was 95 654. The consensus peak-sets were annotated using *HOMER*. As expected, the majority of peaks were assigned to intronic and intergenic regions (**Figure 3.14**). However, a greater percentage of peaks were annotated as promoter regions in the *MACS2* consensus peak-set compared to *Genrich* consensus peak-set (**Figure 3.14**). The peaks that showed any overlap between the two consensus peak-sets and peaks that showed at least 50 % overlap between the two consensus peak-sets were identified and annotated. The majority of overlapping peaks were assigned to intronic and intergenic regions (**Figure 3.14**). These consensus peak sets were then used as the basis for the DA analysis. In addition, the majority of unique peaks for both peak callers were assigned to intronic and intergenic regions.

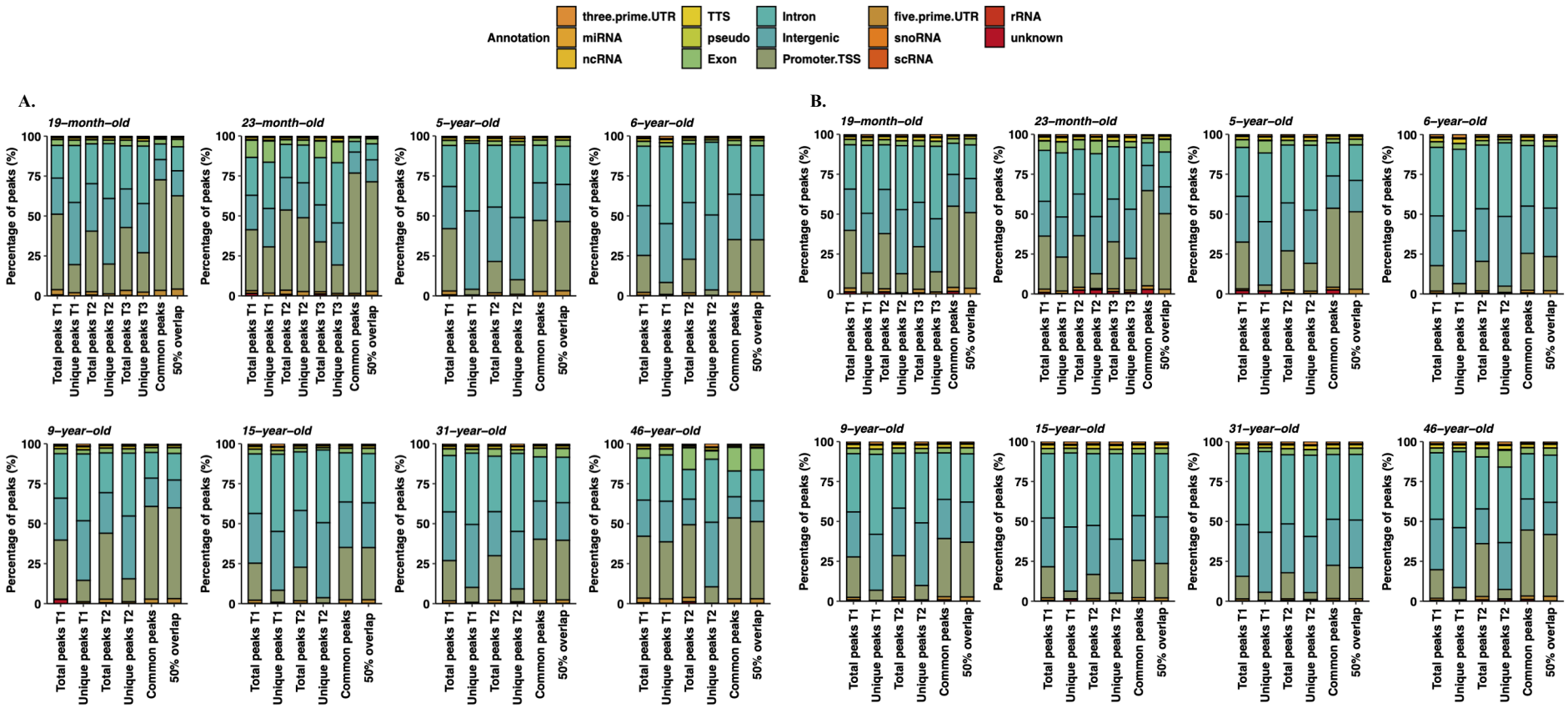


Figure 3.13: ***Peak annotations after performing peak-calling using MACS2 and Genrich peak callers (A-B).*** **(A)** Peak annotation for each sample using the *MACS2* peak caller. **(B)** Peak annotation for each sample using the *Genrich* peak caller. For both peak-callers, annotations are given for 1) the total peaks for each replicate, 2) peaks that were unique to each replicate, 3) peaks that showed any overlap (i.e., at least 1 base pair overlap) between replicates and 4) peaks that showed at least 50 % overlap between replicates. T1, technical replicate 1, T2, technical replicate 2 and T3, cryopreserved sample.

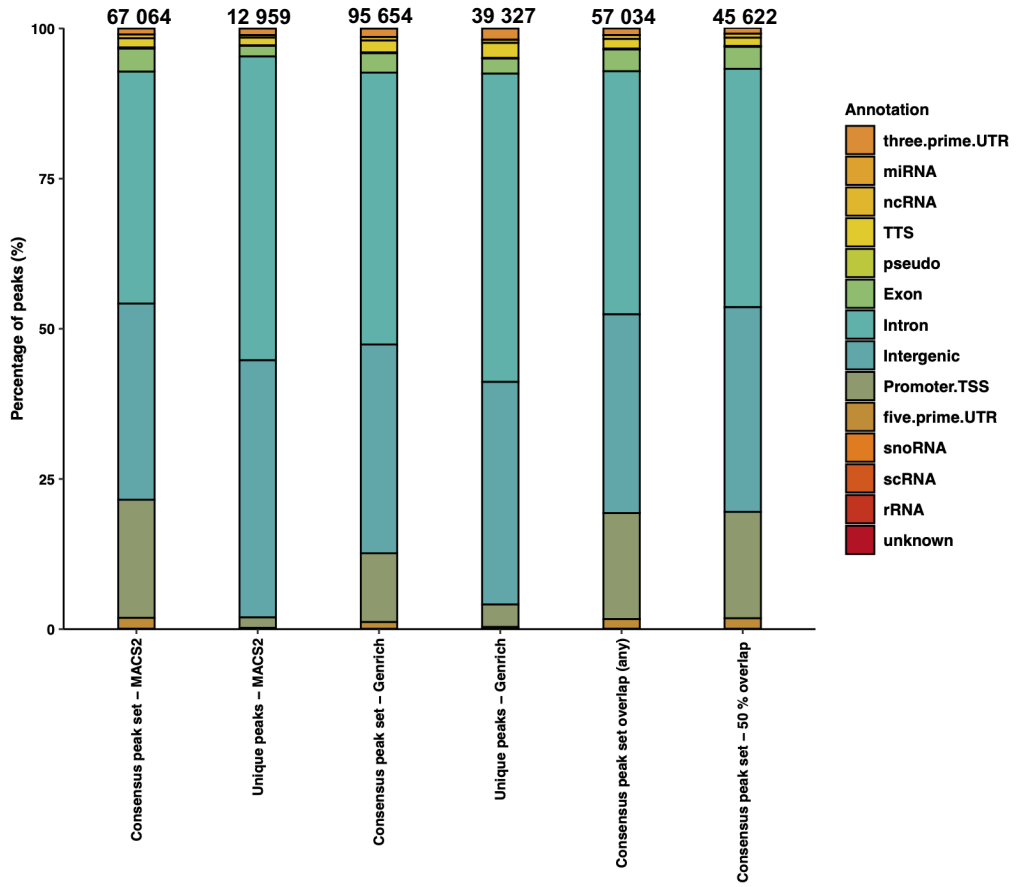


Figure 3.14: ***Peak annotations of the consensus peak sets and their overlapping peaks for each peak-calling method.*** Annotations are given for 1) the total peaks for each consensus peak set, 2) peaks that were unique to each peak-caller's consensus peak set, 3) peaks that showed any overlap (i.e., at least 1 base pair overlap) between consensus peaks sets and 4) peaks that showed at least 50 % overlap between consensus peaks sets.

3.4.6 Differential accessibility analysis

One of the main objectives of this study was to identify regions in the genome that are differentially active over the course of human brain maturation. These regions are likely to be gene regulatory elements that play an important role in directing gene expression as the brain matures. To do this, the ATAC-seq libraries were grouped into four developmental periods: 1) early childhood (19-month-old, 23-month-old, and 5-year-old), 2) middle childhood (6-year-old and 9-year-old), 3) late childhood (14-year-old and 15-year-old) and 4) adulthood (23-year-old, 31-year-old, and 46-year-old). DA analysis was performed, comparing the four different developmental periods to each other.

Currently, there are no differential accessibility tools developed specifically for ATAC-seq data analysis. However, there are several differential expression analysis tools, designed for ChIP-seq or RNA-seq data, that can be used. Some of these tools include, *DESeq2* (Love, Huber and Anders, 2014) and *Diffbind* (http://bioconductor.org/packages/devel/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf). The Reske *et. al* 2020 study demonstrated that the bioinformatic tools that are used can influence the DA analysis results. For this study, four different analysis approaches that utilised normalisation pipelines derived from Reske *et. al* were tested to determine how the peak-caller and normalisation method used can influence the DA results. Approach *I* utilised regions defined by the *MACS2* consensus peak set and a TMM normalisation method, while Approach *II* used the Loess normalisation method on the same peaks. Approach *III* utilised regions defined by the *Genrich* consensus peak set and a TMM normalisation method, while Approach *IV* used the Loess normalisation method on the same peaks. Significant DA peaks were identified by an FDR < 0.05 threshold (**Figure 3.15**). These patterns of differential accessibility were then visualised using MA plots, which show the difference between any two groups (i.e., the fold change) on the *y*-axis (M) and average intensity (i.e., the read count) on the *x*-axis (A).

The results from using different analysis approaches, were not similar. When assessing the changes in accessibility between early and middle childhood, Approach III yielded the most significantly DA peaks (**Figure 3.15i: Genrich|TMM**), with the majority of significant DA peaks showing increased accessibility in early childhood compared to middle childhood. On the other hand, Approach IV resulted in a majority of significant DA peaks with decreased accessibility (**Figure 3.15i: Genrich|Loess**). None of the four approaches yielded significantly DA peaks when assessing the changes in accessibility between early childhood and late childhood (**Figure 3.15ii)**. However, when comparing accessibility between early childhood and adulthood, both Approach I (**Figure 3.15iii: Macs2|TMM**) and Approach III (**Figure 3.15iii: Genrich|TMM**) showed a majority of significantly DA peaks with decreased accessibility in early childhood compared to adulthood. When comparing accessibility between middle childhood and late childhood (**Figure 3.15iv**) or between middle childhood and adulthood (**Figure 3.15v**), very few significant DA peaks were identified. Interestingly, the highest number of significantly DA peaks were identified in the final comparison between late childhood and adulthood using Approach I, with the majority of significantly DA peaks showing decreased accessibility in late childhood compared to adulthood (**Figure 3.15vi: Macs2|TMM**). The three other approaches yielded either no or very few significantly DA peaks for this comparison.

To further investigate the biological function of the significantly DA peaks, *HOMER* was used to annotate peaks as either promoter or distal regions (i.e., intronic or intergenic) (**Figure 3.16**). For all four analysis approaches, the majority of DA peaks were assigned to distal regions, indicating that they may represent distal enhancer regions. Overlap in the promoter or distal peaks called by Approach I and III for the early childhood vs adult comparison was also determined. There were 212 peaks that overlapped between the two approaches and the majority of these peaks were assigned to distal regions.

Overall, these results show that utilising different peak-calling and normalisation approaches will generate varying results. The number of significantly DA peaks identified ranged from 1 to 25 781 peaks, with the majority of peaks showing decreased accessibility in the younger age category. As expected, the more conservative loess normalisation method resulted in far fewer significantly DA peaks when compared to the TMM normalisation method. It is interesting to note that for many of the comparisons, all the approaches yielded few or no significant DA peaks.

1. Early childhood: 19-month-old, 23-month-old, 5-year-old

2. Middle childhood: 6-year-old, 9-year-old

3. Late childhood: 14-year-old, 15-year-old

4. Adulthood: 23-year-old, 31-year-old, 46-year-old

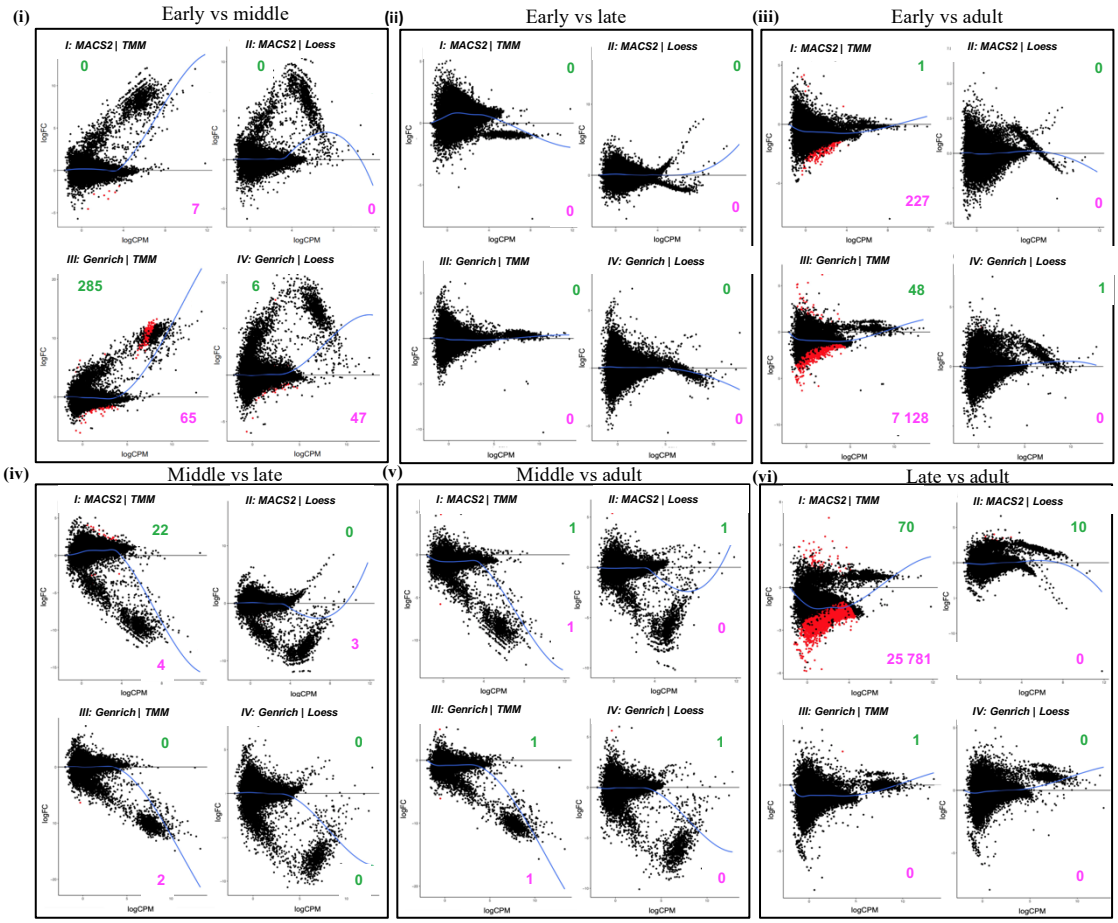


Figure 3.15: ***Differential accessibility distributions from each comparison analysed using two different peak calling methods and two different normalization methods.*** Within each panel, the top two MA plots show DA results using the *MACS2* peak set and the bottom two plots show DA results using the *Genrich* peak set. Use of either the TMM or loess-based normalization method is indicated. The panels show MA plots resulting from comparisons in accessibility between **(i)** early childhood and middle childhood, **(ii)** early childhood and late childhood, **(iii)** early childhood and adulthood, **(iv)** middle childhood and late childhood, **(v)** middle childhood and adulthood and **(vi)** late childhood and adulthood. The MA plot X-axis shows the average ATAC signal abundance at that region (log counts pers million [CPM]) and the Y-axis shows the log difference in ATAC signal between the two groups (log fold change [FC]). Black dots represent non-significant regions, and red dots represent significantly differentially accessible regions (FDR < 0.05). Blue lines represent loess fits for each distribution. The number of peaks that significantly increased or decreased in accessibility is indicated in green and magenta, respectively.

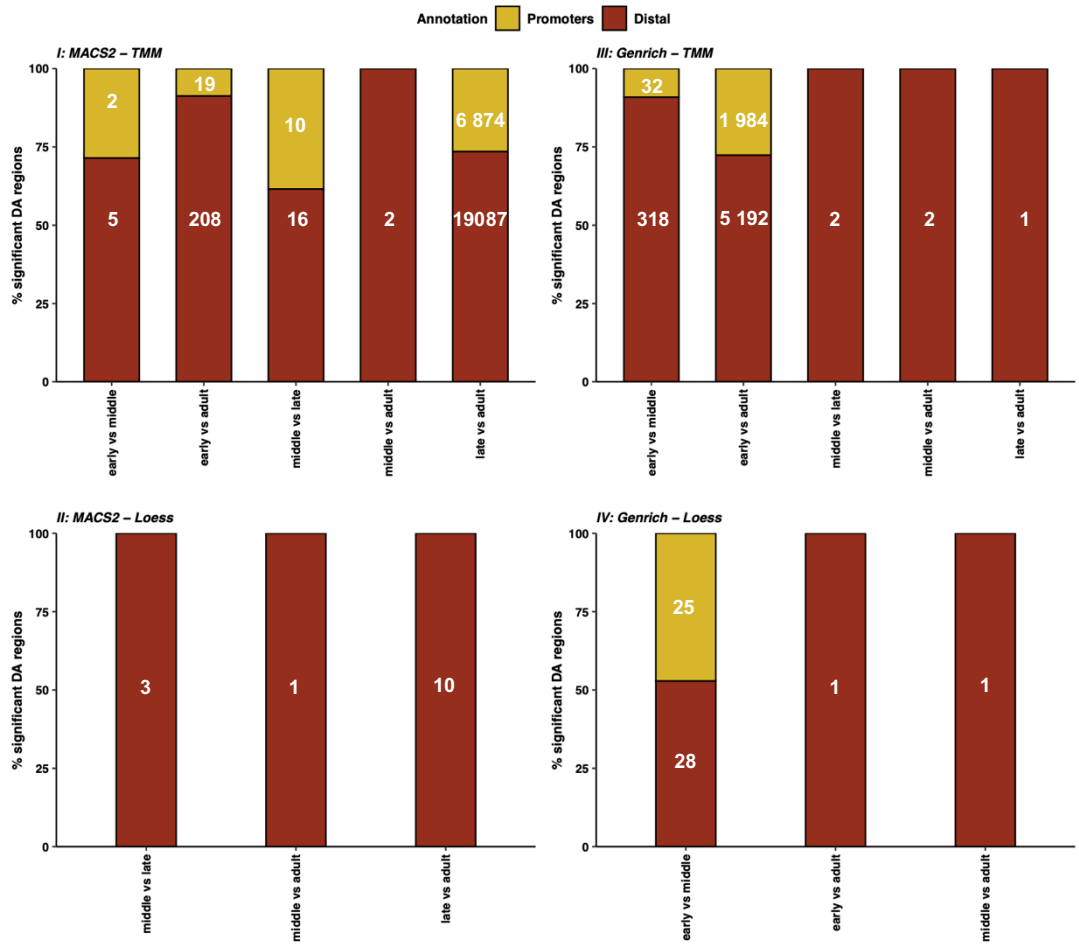


Figure 3.16: ***Annotation of significantly differential accessible peaks.*** Bar plots indicating the percentage of peaks annotated as located in either promoter or distal regions when using either **(i)** MACS-TMM, **(ii)** Genrich-TMM, **(iii)** MACS-Loess or **(iv)** Genrich-Loess approach. The distal category includes, intronic and intergenic peaks.

3.4.7 Functional analysis of significantly DA peaks

In addition to providing genomic annotations for peaks, *HOMER* assigns each peak to the nearest putative target gene. These gene associations were interrogated, focusing specifically on the comparisons that yielded the highest number of statistically DA peaks i.e., early childhood vs adulthood comparison (Approach I and III) and the late childhood vs adulthood comparison (Approach I).

GO analysis of these genes for each comparison revealed important terms associated with neurodevelopment and neurodevelopmental diseases and the top 10 enriched GO terms for each comparison

was visualised (**Figure 3.17**). However, term enrichment was often not significant (i.e., p.adjust > 0.05). GO terms associated with cell differentiation such as "oligodendrocyte differentiation" were enriched for the early childhood vs adulthood comparison with Approach I (**Figure 3.17Ai; Macs2|TMM**). Furthermore, terms associated with development and regulation of development such as "axon development", "positive regulation of neurogenesis" and "positive regulation of cell development" were also enriched for the early childhood vs adulthood comparison with Approach I (**Figure 3.17Ai; Macs2|TMM**). Terms associated with synapse function (e.g., "synapse organization", "macroautophagy", "vacuole organization" and "establishment of protein localization to the membrane") and terms associated with axon development were enriched for the late childhood vs adulthood comparison (**Figure 3.17B; Macs2|TMM**). GO analysis for the 212 overlapping peaks between Approach I and III was also performed with all of the enriched genes being associated to DA peaks that decreased in accessibility in early childhood compared to adulthood (**Figure 3.17C**). This analysis also clearly shows enriched terms associated to oligodendrocytes.

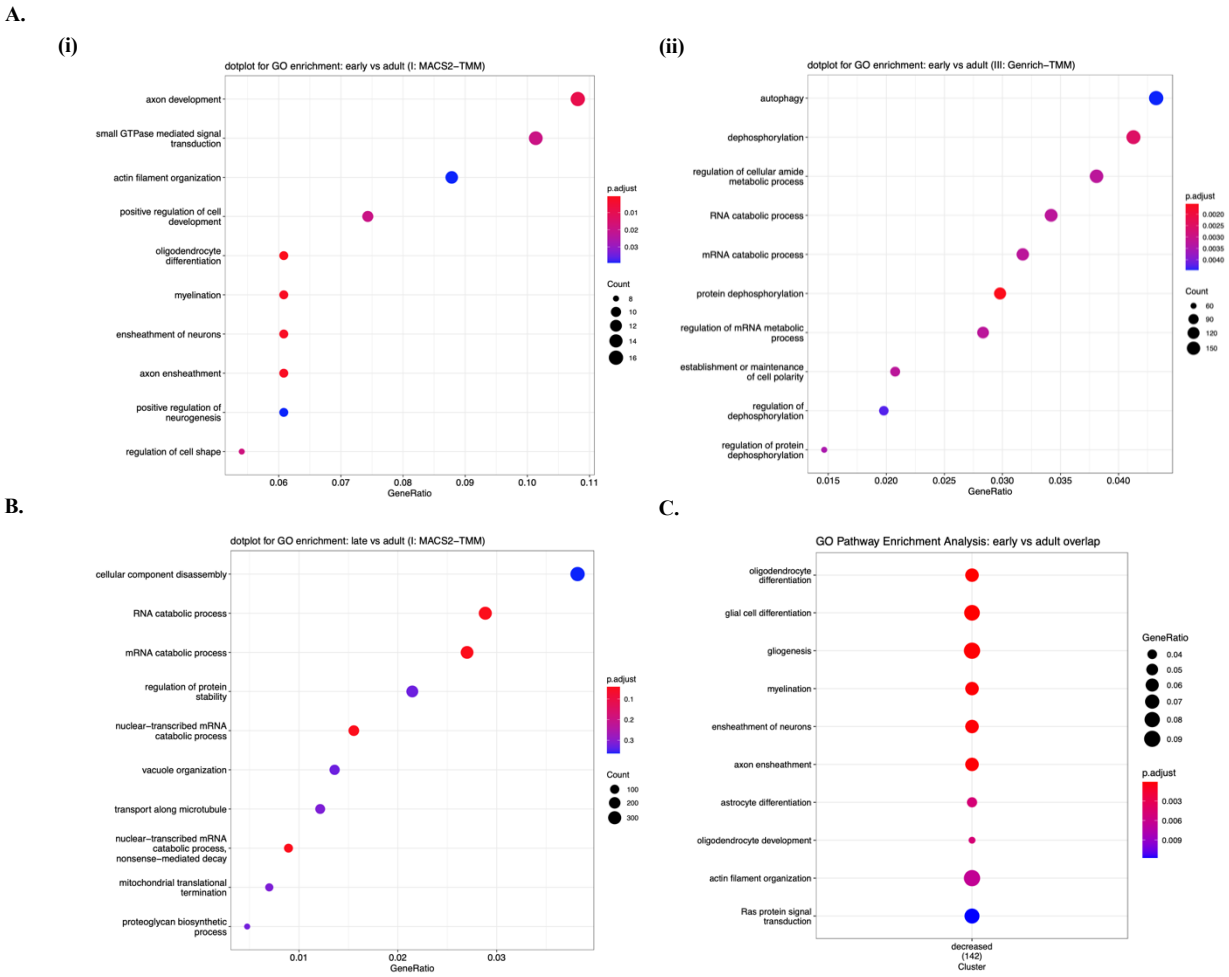


Figure 3.17: ***GO enrichment analysis on significant DA genes for the early childhood vs adulthood and late childhood vs adulthood comparisons.*** **(A)** GO Pathway Enrichment analysis results for Approach I **(i)** and III **(ii)**, for the early childhood vs adulthood comparison. **(B)** GO pathway Enrichment analysis for Approach I for the late childhood vs adulthood comparison. **(C)** GO enrichment analysis on significant DA genes that showed overlap for the early childhood vs adulthood peaks called by Approach I and III. The number in brackets on the x-axis shows the total number of genes that were enriched and that decreased in accessibility. Each dot represents the enrichment term (red, high enrichment; blue, low enrichment). The size of each dot corresponds to the GeneRatio (number of genes belonging to a particular cluster/total number of genes).

Genes that played an interesting role in neurodevelopment were also identified and visualised using the UCSC genome browser. All of the genes highlighted below were also the genes associated with the overlapping peaks between Approach I and III. Significant DA peaks were also compared with publicly available GeneHancer track set (Fishilevich *et al.*, 2017), the ENCODE regulation tracks (ENCODE Project, 2012; Moore *et al.*, 2020) and the snATAC-seq chromatin accessibility profiles for seven major cell types of the brain (Morabito *et al.*, 2021). The snATAC-seq datasets were obtained from prefrontal cortex of postmortem brain tissue (Morabito *et al.*, 2021). The Phosphofurin Acidic Cluster Sorting Protein 2 (*PACS2*) gene promoter was identified through DA analysis for the early childhood vs adulthood comparison using Approach I (*MACS2-TMM*) (**Figure 3.18**). This gene encodes for a multifunctional sorting protein which is critical for pathway traffic regulation and regulation of mitochondria-associated membranes (MAMs) (Li *et al.*, 2020). More importantly, *PACS2* has been shown to play a role in the pathogenesis of several neurodegenerative diseases such as Alzheimer's and Parkinson's disease (Hedskog *et al.*, 2013). Furthermore, a missense mutation in this gene has been associated with early infantile epileptic encephalopathy and facial dysmorphism (Terrone *et al.*, 2020). This gene promoter was shown to be decreased in accessibility in early childhood compared to adulthood and this region was annotated as a promoter in the GeneHancer database (**Figure 3.18**). In addition, this region also showed overlap with the snATAC-seq data with peaks the astrocytes, excitatory neurons, microglia, and oligodendrocytes. This region was highly accessible in oligodendrocytes, and this was consistent with the snRNA-seq data where *PACS2* was expressed in many cell types but was highly expressed in oligodendrocytes (**Supplementary figure 3.10**).

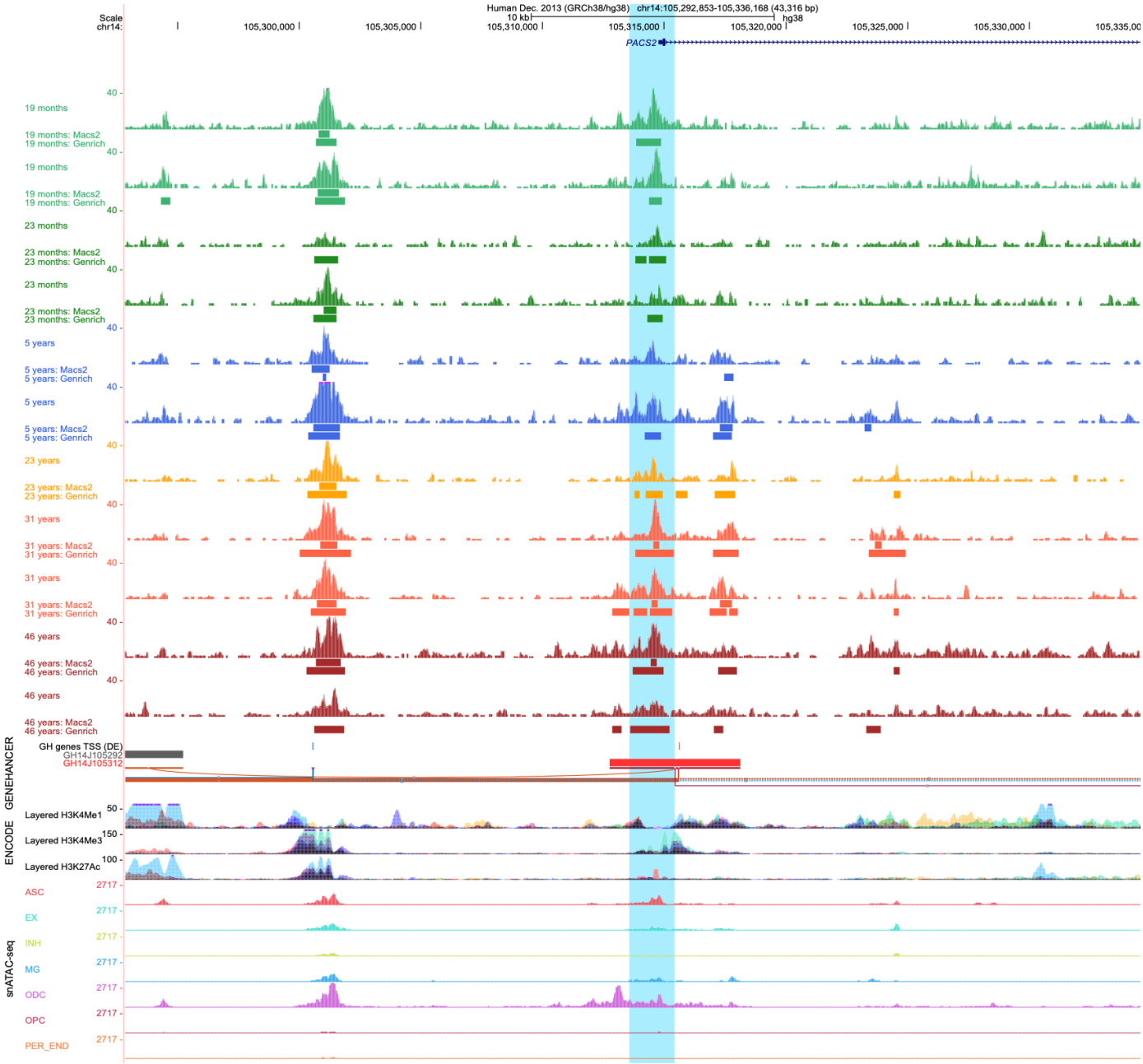


Figure 3.18: ***UCSC Genome Browser view of open chromatin tracks for each sample at the PACS2 gene locus.*** Colour coded ATAC-seq coverage tracks from the 19-month-old, 23-month-old, 5-year-old, 6-year-old, 9-year-old, 14-year-old, 15-year-old, 23-year-old, 31-year-old, and 46-year-old are shown below the *PACS2* transcript annotation. Bars below coverage plots indicate peak regions identified with *MACS2* and *Genrich*. Below the coverage tracks is the GeneHancer track set (Fishilevich *et al.*, 2017) detailing the regulatory elements and the chromatin modification tracks: Layered H3K4Me1, Layered H3K4Me3 and Layered H3K27Ac (ENCODE project, 2012). Below the GeneHancer and ENCODE track sets are the snATAC-seq chromatin accessibility profiles for the seven major cell types of the brain (Morabito *et al.*, 2021). ASC, astrocytes; EX, excitatory neurons; INH, inhibitory neurons; MG, microglia; ODC, oligodendrocytes; OPC, oligodendrocyte precursor cells; PER_END, pericytes/endothelial cells. The *PACS2* region of interest is highlighted in blue.

OPALIN/*TMEM10* was detected through DA analysis for the early childhood vs adulthood comparisons using Approach I (*MACS2-TMM*). This region was shown to decrease in accessibility in early childhood compared to adulthood. As mentioned previously, this is a transmembrane protein involved in oligodendrocyte differentiation and it is specifically expressed in myelinating oligodendrocytes (de Faria *et al.*, 2019). There was no overlap with the GeneHancer dataset, however, there was a clear overlap with the oligodendrocyte peaks in the snATAC-seq dataset (**Figure 3.19**) and could indicate that this is an alternate promoter. The snRNA-seq data showed that *OPALIN* was highly expressed in oligodendrocytes and microglia **(Supplementary figure 3.10**).

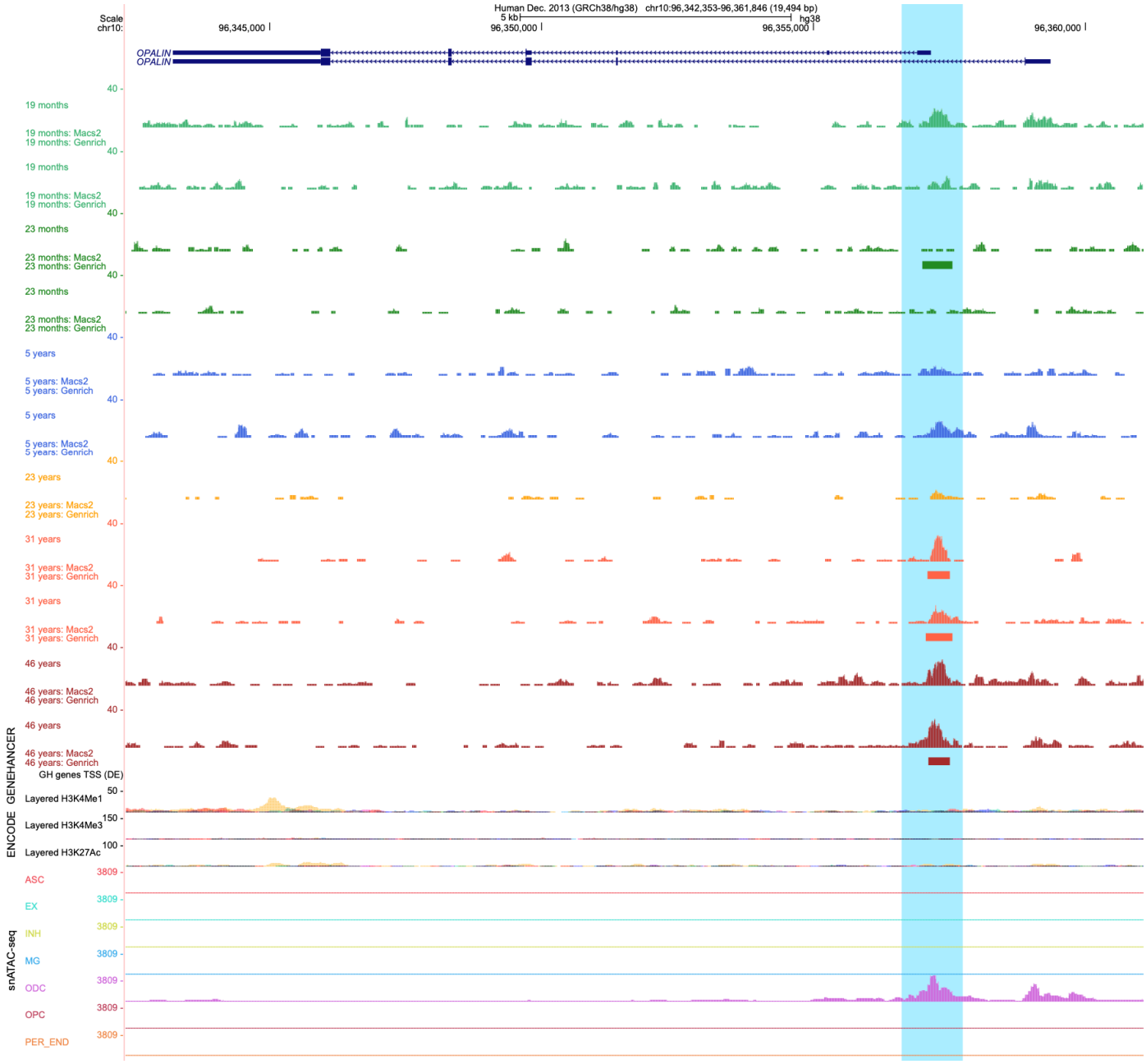


Figure 3.19: ***UCSC Genome Browser view of open chromatin tracks for each sample at the OPALIN gene locus.*** Colour coded ATAC-seq coverage tracks from the 19-month-old, 23-month-old, 5-year-old, 6-year-old, 9-year-old, 14-year-old, 15-year-old, 23-year-old, 31-year-old, and 46-year-old are shown below the *OPALIN* transcript annotation. Bars below coverage plots indicate peak regions identified with *MACS2* and *Genrich*. Below the coverage tracks is the GeneHancer track set (Fishilevich *et al.*, 2017) detailing the regulatory elements and the chromatin modification tracks: Layered H3K4Me1, Layered H3K4Me3 and Layered H3K27Ac (ENCODE project, 2012). Below the GeneHancer and ENCODE track sets are the snATAC-seq chromatin accessibility profiles for the seven major cell types of the brain (Morabito *et al.*, 2021). ASC, astrocytes; EX, excitatory neurons; INH, inhibitory neurons; MG, microglia; ODC, oligodendrocytes; OPC, oligodendrocyte precursor cells; PER_END, pericytes/endothelial cells. The *OPALIN* region of interest is highlighted in blue.

CNTN2 or Contactin 2 is a protein-coding gene that has been associated with epilepsy and has previously been shown to be overexpressed in gliomas (Rickman *et al.*, 2001; Stogmann *et al.*, 2013). This gene has also been shown to be expressed in post-mitotic and mature oligodendrocytes and *CNTN2* plays an important role in oligodendrocyte branching (Zoupi *et al.*, 2018). *CNTN2* was also visualised using the UCSC genome browser and there was a clear overlap with an enhancer in the GeneHancer dataset and with peaks in the snATAC-seq datasets. This region was shown to decrease in accessibility in early childhood compared to adulthood. The snATAC-seq dataset showed that the *CNTN2* region was accessible in astrocytes, excitatory neurons, inhibitory neurons, and oligodendrocytes. However, this region was specifically enriched in oligodendrocytes as there was higher levels of snATAC-seq signal at the *CNTN2* locus in this cell type (**Figure 3.20**). Furthermore, this gene was greatly expressed in oligodendrocytes compared to the other cell types in our snRNA-seq data (**Supplementary figure 3.10**).

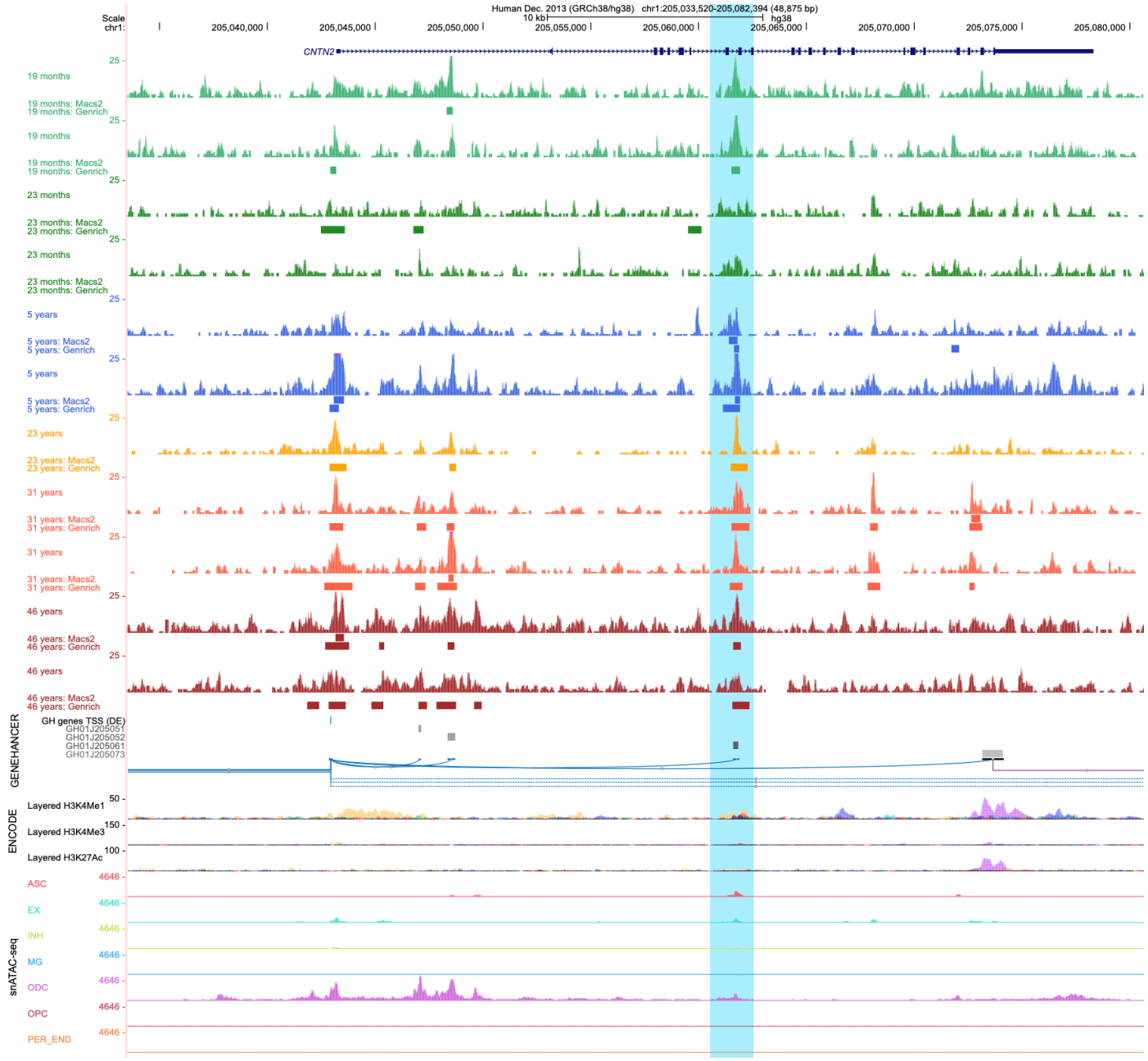


Figure 3.20: ***UCSC Genome Browser view of open chromatin tracks for each sample at the CNTN2 gene locus.*** Colour coded ATAC-seq coverage tracks from the 19-month-old, 23-month-old, 5-year-old, 6-year-old, 9-year-old, 14-year-old, 15-year-old, 23-year-old, 31-year-old, and 46-year-old are shown below the *CNTN2* transcript annotation. Bars below coverage plots indicate peak regions identified with *MACS2* and *Genrich*. Below the coverage tracks is the GeneHancer track set (Fishilevich *et al.*, 2017) detailing the regulatory elements and the chromatin modification tracks: Layered H3K4Me1, Layered H3K4Me3 and Layered H3K27Ac (ENCODE project, 2012). Below the GeneHancer and ENCODE track sets are the snATAC-seq chromatin accessibility profiles for the seven major cell types of the brain (Morabito *et al.*, 2021). ASC, astrocytes; EX, excitatory neurons; INH, inhibitory neurons; MG, microglia; ODC, oligodendrocytes; OPC, oligodendrocyte precursor cells; PER_END, pericytes/endothelial cells. The *CNTN2* region of interest is highlighted in blue.

Finally, Tubulin Beta 4A class Iva (*TUBB4A*) was identified through DA analysis for the early childhood vs adulthood comparison using Approach I (*MACS2-TMM*) (**Figure 3.21**). *TUBB4A* was shown to decrease in accessibility in early childhood compared to adulthood. This gene encodes the tubulin beta 4A protein, which is required to form microtubules (Curiel *et al.*, 2017; Sase *et al.*, 2020). Previous studies have shown that mutations in this gene leads to a range of neurological disorders such as leukodystrophy (Simons *et al.*, 2013; Curiel *et al.*, 2017; Sase *et al.*, 2020). There was no overlap with the GeneHancer dataset but there was overlap with the snATAC-seq dataset, where it was accessible in all the cell types but not in the endothelial cells and was highly enriched in oligodendrocytes. This peak was assigned to an intergenic region, suggesting that it could be an enhancer (**Figure 3.21**).

Figure 3.21: ***UCSC Genome Browser view of open chromatin tracks for each sample at the TUBB4A gene locus.*** Colour coded ATAC-seq coverage tracks from the 19-month-old, 23-month-old, 5-year-old, 6-year-old, 9-year-old, 14-year-old, 15-year-old, 23-year-old, 31-year-old, and 46-year-old are shown below the *TUBB4A* transcript annotation. Bars below coverage plots indicate peak regions identified with *MACS2* and *Genrich*. Below the coverage tracks is the GeneHancer track set (Fishilevich *et al.*, 2017) detailing the regulatory elements and the chromatin modification tracks: Layered H3K4Me1, Layered H3K4Me3 and Layered H3K27Ac (ENCODE project, 2012). Below the GeneHancer and ENCODE track sets are the snATAC-seq chromatin accessibility profiles for the seven major cell types of the brain (Morabito *et al.*, 2021). ASC, astrocytes; EX, excitatory neurons; INH, inhibitory neurons; MG, microglia; ODC, oligodendrocytes; OPC, oligodendrocyte precursor cells; PER_END, pericytes/endothelial cells. The *TUBB4A* region of interest is highlighted in blue.

Bulk ATAC-seq datasets were successfully generated from brain tissue samples from a broad range of ages. The data was analysed using different bioinformatic pipelines, yielding varying results. However, these analyses identified several interesting genes that were related to neurodevelopment and showed overlap with the snRNA-seq data. Furthermore, GO analysis results revealed enriched terms that were associated with neurodevelopment.

Chapter 4 - Discussion

The overall aim of this research project was to characterise the molecular profiles of pediatric and adult cells to gain a better understanding of how the brain cell atlas changes as we age. The aim was to identify major cell types in pediatric brain samples and identify important genes being expressed. In addition, a dataset of putative enhancers and promoters controlling their expression was also generated. To achieve this, advanced technology was utilised such as snRNA-seq using the 10x Genomics platform and ATAC-seq technology. Importantly, this project also aimed to establish the methods for these tools in the Hockman lab, especially the human brain tissue processing steps.

4.1 Cell-type specific gene expression analysis using SnRNA-seq

4.1.1 Sample quality

A lot of work has been done applying snRNA-seq to human brain tissue samples and several sample preparation methods have been developed for snRNA-seq (Lake *et al.*, 2016, 2018; 10x, 2017; Habib *et al.*, 2017). Additionally, a great deal of effort has been taken to optimise these nuclei isolation protocols for the use of fresh or flash-frozen human tissue (10x, 2017; Habib *et al.*, 2017; Hu *et al.*, 2017). These studies have assessed the effect of using these nuclei isolation protocols on the sample quality and downstream analyses results. Habib *et al.* demonstrated that using their nuclei isolation protocol, on frozen adult human brain tissue, resulted in the generation of high-quality nuclei. In addition, they also showed that they were able to generate high-quality libraries and identify cell types, cell subtypes and rare cell types of the human brain (Habib *et al.*, 2017). This project aimed to optimise their nuclei isolation protocol for use in the Hockman lab. This protocol allowed us to prepare ante-mortem brain samples that were obtained during surgery to remove epileptic tissue.

Habib *et al.* recommends that if tissue is to be preserved rather than used fresh, it should be flash frozen before storing at −80 ℃ (Habib *et al.*, 2017). In this study, the 15-year-old sample was flash frozen in liquid nitrogen immediately after arriving in the lab, however, the 14-year-old sample was slow-frozen by placing in the −80 ℃ freezer after surgery. This slow freezing process could have resulted in RNA degradation in the 14-year-old sample and ultimately may have affected the quality of the sequencing libraries generated and therefore the downstream analysis steps. Following nuclei isolation, the total number of nuclei for each sample was determined by staining with Hoechst and counting on a hemocytometer.

During this step, it was also possible to assess the level of cell debris in the sample. Two technical replicate (T) tissue samples were prepared and analysed for the 14-year-old sample, while a single tissue sample was prepared for the 15-year-old sample. A higher concentration of nuclei was obtained for 14-year-old T2 compared to T1. During the nuclei isolation protocol, the 14-year-old T1 nuclei suspension was left on ice while the tissue fragment used for 14-year-old T2 was being homogenised. During this time, a proportion of nuclei from T1 could have degraded, resulting in a lower number of total nuclei. Additionally, there was a large amount of cell debris present in both the 14-year-old and 15-year-old samples, which ultimately affected the data analysis and interpretation of the results. The presence of cell debris could indicate that the time the nuclei spent in lysis buffer might have been too long and caused the cells to over lyse. It is also possible that the size of cell strainers used was incorrect, resulting in excess cell debris passing through the cell strainers.

The possible effects of the high levels of cell debris could be seen in the initial quality checking of the snRNA-seq libraries after running the Cell Ranger *Count* command. More than 50 % of nuclei were lost for the 14-year-old sample, particularly for T2 and the fraction of reads in each nucleus was below the expected amount (70 %), specifically for the 14-year-old sample (45.1 % and 66.3 %), which is an indication of high levels of ambient RNA (10X Genomics, 2020). Another snRNA-seq pipeline and protocol was developed by Grindberg *et al.* where they applied their protocol to nuclei and whole cells from mouse brain tissue (Grindberg *et al.*, 2013). The results obtained from the nuclei and whole cells were compared to each other and they concluded that their protocol was able to resolve high quality nuclei and general expression differences between genes were conserved between whole cells and nuclei. Grindberg *et al.* highlighted the possibility that using nuclei might introduce sources of technical variation, such as increased mRNA degradation during the cell lysis step (Grindberg *et al.*, 2013). This could explain the high levels of ambient RNA seen for the 14-year-old sample.

Additional QC and filtering steps were performed using the *Seurat* pipeline, confirming that the 14-yearold libraries were of lower quality than the 15-year-old libraries. These results showed high levels of genes mapping to the mitochondrial genome in the 14-year old sample, which has previously been shown to be an indication of cell stress and associated with low quality or dying cells (Zhao, 2002). When visualising the total number of UMI and genes per sample, a single peak was expected. The total number of UMIs and genes should be above 500, which was seen for all the samples. The single peak indicates a good quality library as you would expect the majority of nuclei to have the same or similar number of UMIs and genes. This pattern was seen for the 15-year-old libraries but not for the 14-year-old libraries, which showed a range of peaks. For the 14-year-old sample, this could be due to the overall low nuclei count and the high mitochondrial read count. After the initial QC was performed, any identified doublets were removed from

each individual replicate with 14-year-old T1 library having a higher number of identified doublets compared to the 14-year-old T2. The 15-year-old replicates all had a similar number of identified doublets, with 15-year-old T2 having the highest number of doublets. In addition, the 15-year-old sample yielded a higher percentage of doublets than the 14-year-old sample, but this could be due to the greater number of total nuclei obtained for the 15-year-old sample. In this study, a doublet rate between 0.66 and 3.67 % was observed and previous studies have obtained higher doublet rates between $2.1 - 7.5$ % (Gaublomme *et al.*, 2019; Nagy *et al.*, 2020; Thrupp *et al.*, 2020). These results demonstrated just how important it was to perform the doublet removal step, otherwise these identified doublets would have remained in the final dataset and affected the ability to assign cell types to specific clusters (Kang *et al.*, 2018; DePasquale *et al.*, 2019; McGinnis, Murrow and Gartner, 2019). Overall, more nuclei were filtered out from the 14-year-old libraries (11.26 % – 11.71 %) compared to the 15-year-old libraries (7.6 % - 8.28 %).

4.1.2 Clustering of nuclei

After filtering low quality nuclei and doublets, the snRNA-seq libraries were normalised and integrated for cluster analysis. When using *Seurat* to cluster data, it is recommended to use a resolution parameter of 0.4 to 1.4 for smaller datasets and increasing this number for larger datasets (Stuart *et al.*, 2019) (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html). For the clustering analysis, the lowest resolution of 0.4 was used, resulting in 23 clusters. If a higher resolution was used, a greater number of clusters would have been identified and this would have also allowed for the identification of potential subtypes of the major cell types. However, because of the lower quality of the 14-year-old sample and the lower number of total nuclei, increasing the resolution parameter and generating a higher number of clusters would affect the ability to annotate these clusters.

Clustering did not result in any clusters that were specific to a particular sample, however the majority of nuclei in clusters 6 and 8 were from 14-year-old T2. These clusters were identified as cell debris clusters due to their high mitochondrial gene expression levels, corroborating the QC results which showed a high mitochondrial gene read count for 14-year-old T2. Finally, several clusters were largely made up of nuclei from the 15-year-old sample compared to the 14-year-old sample. This could be a result of the lower quality of the 14-year-old sample and less nuclei being recovered. Overall, these findings indicate that the 14-yearold sample libraries were inferior in quality when compared to the 15-year old sample libraries. It is important to note that the 15-year-old sample was processed after additional nuclei isolation test runs were performed and it is likely that efficiency of the process was improved through these test runs.

4.1.3 Nuclei preparation and optimisation

Based on the results obtained, it was clear that additional filtering and lysis optimisations needed to be performed to improve the nuclei isolation protocol and, in particular, to reduce the amount of debris in the resulting nuclei preparation. In total, four additional nuclei isolation tests were conducted on pediatric and adult human brain tissue. The first two sets of nuclei isolation tests were performed using the standard protocol used for the 14-year-old and 15-year-old samples in order to assess the level of cell debris at each step in the protocol. Even though the results showed that there was a reduction in cell debris for the pediatric sample, this was accompanied by a loss in nuclei. Therefore, further nuclei isolation tests needed to be conducted to identify a method that reduced the amount of cell debris without losing nuclei. The third test used the Gaublomme *et al.* nuclei isolation protocol, which utilised a different lysis buffer, different cell filters and a shorter lysis time. For this reason, it was expected that this protocol would reduce the level of cell debris. Collectively, these tests showed that cell debris was still present at high level and that using cell strainers of different sizes as used in the Gaublomme *et al.* protocol did not reduce the level of cell debris (**Figure 3.11**). The final test used an additional myelin removal step with the standard nuclei isolation protocol to reduce the level of myelin and any cell debris particles (Mok, 2019). Several studies have also used a myelin removal step in their nuclei isolation protocol (Li *et al.*, 2019; Olah *et al.*, 2020; Claes *et al.*, 2021). These results demonstrated that incorporating a myelin removal step did indeed remove a large amount of cell debris, resulting in a cleaner nuclei preparation.

For future analyses, optimisations of the nuclei isolation tests should be performed before each experiment. This is especially important if brain tissue from a specific brain region or age group is used for the first time. Finally, when processing the samples, it is important to work efficiently to avoid over lysis of cells. An additional myelin removal step should also be performed to improve the overall quality of the nuclei preparation.

4.1.4 Classification of cell types

A combination of automated annotation, using the *SCSA* tool (Cao, Wang and Peng, 2020), and manual annotation was used to determine the cellular identity of the clusters. *SCSA* was used to automatically assign cell types using a combination of differentially expressed genes from multiple databases (Cao, Wang and Peng, 2020). The manual annotation method was beneficial to verify the automated annotation results and to determine the identity of the majority of 'unknown' clusters that could not be annotated by *SCSA*. Using the manual annotation method to identify the excitatory and inhibitory neuronal sub-clusters was more of a challenge so it was important to also use the automatic annotation approach as this tool utilises multiple datasets resulting in more accurate annotations. Two unknown clusters remained after using this combined

approach. In the future it would be advantageous to utilise additional automated cell type assignment tools such as *SCINA* (Zhang *et al.*, 2019), *scCATCH* (Shao *et al.*, 2020) or *CellAssign* (A. W. Zhang *et al.*, 2019) and compare it to the *SCSA* annotation results as well as the manual annotation results. In *Seurat*, the *TransferData* function can also be used to transfer previously published cell type annotations onto the clusters, which would assist in the annotation (Stuart *et al.*, 2019). Interestingly, it seems that previous studies have primarily used manual annotation and specific markers to assign cell types to identified clusters (Habib *et al.*, 2017; Lake *et al.*, 2018; Zhong *et al.*, 2018; Polioudakis *et al.*, 2019). More recent studies have used a combination of manual and automated approaches to identify cell types (Hodge *et al.*, 2019; Smajic *et al.*, 2020; Bakken *et al.*, 2021). Using a low resolution of 0.4 during the cell clustering in *Seurat* could have also affected the ability to assign certain cell types as different resolution parameters results in a different number of clusters with varying compositions. To improve these results in the future, testing a range of resolutions would be beneficial as well as using additional cell-type specific markers from the literature or databases.

The majority of clusters generated from our data were identified as one of the major cell types of the brain, consistent with previously published data, where scRNA-seq or snRNA-seq analysis was performed (Darmanis *et al.*, 2015; Nowakowski *et al.*, 2017). For example, Darmanis *et al*., performed scRNA-seq on adult and fetal human brain. They successfully grouped the single cells into six major cell types of the brain: neurons, OPCs, astrocytes, oligodendrocytes, microglia, and vascular cells (Darmanis *et al.*, 2015). Ultimately, this showed that even with a small sample size, identifying the major cell types of the brain is possible and demonstrates the power of single cell analysis. More recent studies are now utilising this power to not only identify all the broad cell types, but also identify subtypes of neuronal and non-neuronal cells (Lake *et al.*, 2018; Hodge *et al.*, 2019).

Oligodendrocytes (15, 147) and astrocytes (6, 642) made up the largest proportion of snRNA-seq datasets in this study. This was not expected as previous studies have shown that the majority of nuclei in the datasets, usually belonged to neuronal clusters (Habib *et al.*, 2017; Fan *et al.*, 2018; Polioudakis *et al.*, 2018; Hodge *et al.*, 2019; Nagy *et al.*, 2020). This could be a result of the type of samples that were used in this study, with the 14-year-old and 15-year-old brain tissue samples consisting of white matter as well as grey matter. This was done to ensure a wide diversity of cell types and potentially rare neurons were captured, particularly in the white matter. Habib *et al.* obtained adult brain tissue samples from the prefrontal cortex but also from the hippocampus. The hippocampus has been shown to be made up of several distinct types of neurons and could be why the majority of their nuclei were identified as neurons in their study (Habib *et al.*, 2017).

4.1.5 Differential expression analysis

DE analysis was performed on two different samples which come from two different brain regions, temporal, and frontal regions. This was done to identify cell-type specific changes in expression between the two brain regions. As mentioned before, this study focused on establishing the snRNA-seq workflow to understand brain maturation. In the absence of samples from patients of distinct ages, this comparison between the two brain regions from patients of similar age was performed as a proof of principle and to demonstrate the feasibility of the DE analysis pipeline.

Interestingly, the temporal lobe showed the greatest number of enriched genes (19, 886 genes) compared to the frontal lobe (17, 812 genes). Previous studies have demonstrated that brain region and age contribute more to the global differences in gene expression compared to other variables such as sex or inter-individual variation (Kang *et al.*, 2011). This spatio-temporal regulation occurs primarily during prenatal development (Johnson *et al.*, 2009; Kang *et al.*, 2011; Li *et al.*, 2018). In particular, Li *et al.* performed bulk RNA-seq on postmortem human brains. The ages of these samples ranged from 8 PCW to 40 years old (Li *et al.*, 2018). The bulk RNA-seq studies revealed that the regional differences in gene expression decreased during the important late fetal – early infancy period, meaning that these cortical regions are most similar to each other during this period. These regional differences in gene expression begin to increase during the later stages of development (late childhood – adolescence) (Li *et al.*, 2018). However, the gene expression profiles of the temporal and frontal lobes are still more similar to each other than they were during the early stages of development. In this study, the DE analysis revealed a large number of DE genes, corroborating their findings that regional differences in gene expression can be detected in adolescence and expanding this regional analysis to a single cell level. (Li *et al.*, 2018). For future analysis, it would be interesting to use additional brain tissue samples from different developmental periods and brain regions to verify these results at the single cell level.

When assessing DE genes in specific cell clusters in this study, the excitatory neurons showed the greatest number of DE genes while the endothelial cells showed the lowest number of DE genes. Overall, the trend showed that the cell types with a larger number of total nuclei, showed a greater number of DE genes. Certainly, other studies have shown that cell types with a greater number of nuclei, generate a larger number of DE genes (Nagy *et al.*, 2020). For example, in this study they obtained a greater number of excitatory neurons compared to inhibitory neurons and glial cell types with the excitatory neurons showing a greater number of DE genes (Nagy *et al.*, 2020). Several genes enriched in the temporal lobe (in all the cell types) were identified as long non-coding RNAs (lncRNAs), including *LINC01473*, *MEG3* and *LINC00472*. LncRNAs are defined as a class of linear transcripts with at least 200 nucleotides in length (Djebali *et al.*,

2012). Previous studies have demonstrated that lncRNAs are highly expressed in the human brain (approximately 40 %) and play an important role in the regulation of normal brain development (Derrien *et al.*, 2012; Roberts, Morris and Wood, 2014). LncRNAs have demonstrated spatio-temporal expression which further confirms their importance in brain maturation. One particular study identified lncRNAs that were regionally differentially expressed across brain development, with postnatal and adulthood showing a greater number of DE lncRNAs compared to prenatal development, specifically in the cerebellar cortex (Zhang *et al.*, 2017). Possibly, the lncRNAs identified from this analysis are regionally differentially expressed because they play an important role in the overall development and function of specific brain regions such as the temporal lobe. This is consistent with this study, where a large number of lncRNAs were enriched. For example, maternally expressed 3 (*MEG3*) was enriched in the temporal lobe for excitatory neurons, inhibitory neurons, astrocytes, endothelial cells, and microglia. *MEG3* encodes a lncRNA that is expressed in many normal tissues such as the pituitary and human liver tissue. This gene has also been shown to function as a tumor suppressor gene (Zhou, Zhang and Klibanski, 2012). In addition, *MEG3* has been shown to be involved in the regulation of neuronal synapse plasticity (Tan *et al.*, 2017). Interestingly, *MEG3* was also found to be upregulated in the nucleus accumbens of heroin users, showing a possible role for lncRNAs in addictive behaviours (Michelhaugh *et al.*, 2011). This is consistent with the functional enrichment analysis results, where several terms related to addiction were enriched in the temporal lobe, specifically for excitatory and inhibitory neurons. Perhaps, *MEG3* is regionally differentially expressed during adolescence because of the temporally distinct trajectories of each brain region. It has been shown that brain regions such as the primary motor cortex, frontal and occipital lobes mature first, while the temporal lobe matures later (Gogtay *et al.*, 2004).

Several genes enriched in the frontal lobe were also lncRNAs such as *TTTY14* and protein kinase genes such as neurogranin (*NRGN)*. *NRGN* was first shown to be expressed in pyramidal cells of the cortex and hippocampus (Represa *et al.*, 1990). More recently, it has been shown to be expressed in endothelial cells (Cheriyan *et al.*, 2020). *NRGN* was enriched in the frontal lobe for inhibitory neurons, OPCs and microglia. This gene has been identified as a calmodulin (CaM)-binding protein and plays an important role in regulating the CaM-Ca^{2+}-dependent enzymes involved in synaptic plasticity, synaptic regeneration, and long-term potentiation (Huang, 2004; Zhong *et al.*, 2009). One important gene that was enriched in the frontal lobe for excitatory neurons and inhibitory neurons was *KCNAB1*. *KCNAB1* encodes for a specific voltage-gated potassium channel (*Kvbeta1*) and the functions of these potassium channels include neuronal excitability, neurotransmitter release and smooth muscle contraction (Leicher *et al.*, 1996). The *LGI1* gene encodes for a protein that has been shown to control the activities of the *Kvbeta1* potassium channel. In addition, mutations in the *LGI1* gene cause changes in the inactivation of the *Kvbeta1* potassium channel (Schulte *et al.*, 2006). Several mutations in the *LGI1* gene have been identified that has been shown to cause

autosomal dominant lateral temporal lobe epilepsy (ADLTE) (Kalachikov *et al.*, 2002; Morante-Redolat, 2002). Therefore, *KCNAB1* was identified as a susceptibility gene for lateral temporal epilepsy (LTE) because it interacts with the *LGI1* gene (Busolin *et al.*, 2011). Not much is known about the direct association between *KCNAB1* and LTE, however, one particular study identified a novel mutation within the *KCNAB1* gene in a patient with early infantile epileptic encephalopathy (EIEE) (Zhang *et al.*, 2015). *KCNAB1* could be enriched in the frontal lobe based on the overall functions of this region such as voluntary movement and executive functions such as decision-making (Chayer and Freedman, 2001). It could also indicate that the frontal lobe tissue sample does not represent normal brain tissue and was more affected by the epilepsy background of the patient.

The most notable difference between GO, KEGG and DO terms across the two brain regions was that several of the temporal lobe enriched terms were more distinct and related to specific cell-type functions such as “phagocytosis” in microglia or “regulation of post-synaptic membrane potential” in neurons. This was not observed for the frontal lobe where the functional enrichment terms were similar across all the cell types. Again, this may be due to the fact that the temporal lobe samples were of better quality than the frontal lobe samples. Furthermore, the higher number of lower quality nuclei in the frontal lobe could have made the cell-type specific gene expression profiles less distinct. However, the terms enriched in the temporal lobe are not significant as none of them have a p value < 0.05. The DO terms such as epilepsy syndrome and focal epilepsy were enriched in all cell types (except for the endothelial cells) for the frontal lobe. Interestingly, these terms were not enriched in any cell types for the temporal lobe. Again, this could show that the frontal lobe samples (14 T1 and T2) were more affected by the epilepsy background of the patient and perhaps does not represent normal brain tissue. This also could be why the 14-year-old sample was of inferior quality compared to the 15-year-old sample.

These results showed that snRNA-seq libraries were successfully generated for the pediatric brain. In addition, a pipeline was successfully applied to analyse the resulting snRNA-seq data, and a pilot DE analysis step was conducted where the gene expression profiles between the same cell types found in two different brain regions were compared. However, it is important that additional samples that are of consistent quality are used for future analyses to determine if the trends seen in the DE analysis are in fact significant. Finally, this analyses also needs to be repeated with samples from different age groups to identify how gene expression changes over the course of brain maturation.

4.2 Exploring chromatin dynamics during brain maturation using ATAC-seq

4.2.1 Sample quality

As mentioned above, the brain tissue samples used for this study were obtained from patients with neurological conditions such as temporal or frontal lobe epilepsy and Sturge Weber syndrome (**Table 2.1**). This was something to consider when assessing the quality of samples and interpreting the results. The use of cryopreserved tissue fragments was also tested and compared to using fresh tissue fragments for the ATAC-seq library generation. Post sequencing QC was performed using *Picard* tools, where most of the samples showed good insert size distribution. All of the samples showed a distribution of peaks that decreased in height as the size of the insert sequence increased, which corresponds to where the Tn5 transpose was inserted. Most of the samples also showed a higher number of smaller insert sizes, which is expected and corresponds to the highly fragmented open chromatin. However, this was not seen for the 23-month-old samples (fresh and cryopreserved). This sample could have been over-tagmented, where the transpose gains access and cuts parts of the closed chromatin rather than being restricted to regions of open chromatin. Importantly, the 19-month-old cryopreserved sample showed good insert distribution, again, demonstrating that cryopreserved samples can yield similar results to fresh or frozen samples. Previous studies have shown that cryopreservation yields high quality data that is comparable with data obtained from fresh cells or tissue (Milani *et al.*, 2016; Corces *et al.*, 2017; Fujiwara *et al.*, 2019). The quality of these samples was also assessed by using the UCSC genome browser and the coverage tracks were displayed at the *GFAP* locus. This demonstrated that the samples were of sufficient quality as they all displayed sequence pile ups at the expected region of the *GFAP* gene, including the 23-month-old samples. In light of this, all samples were included in the downstream analysis.

4.2.2 OCR identification and annotation

Peak calling was performed to identify regions of significantly open chromatin across all ATAC-seq libraries using *MACS2* and *Genrich* and revealed varying results. Apart from the 5-year-old ATAC-seq libraries, more peaks were called using *Genrich* compared to *MACS2*. As mentioned, this might be because *Genrich* calls narrow peaks, while *MACS2* was used to call broad peaks. To our knowledge, there are currently no benchmark studies comparing the functionality of these two peak-callers and other peak-callers in general. *Genrich* would be the preferred tool to use as it has a mode dedicated to ATAC-seq data (https://github.com/jsh58/Genrich), whereas *MACS2* does not. However, *MACS2* is the most commonly used tool to date, and it is the default peak caller of the ENCODE ATAC-seq pipeline (Yan *et al.*, 2020) (https://github.com/ENCODE-DCC/atac-seq-pipeline).

For future analysis, it would be beneficial to also call narrow peaks with MACS2 to determine if the same or similar number of total peaks are obtained as with *Genrich*. A greater number of *Genrich* peaks were identified for the cryopreserved samples compared to the fresh samples. Perhaps, cryopreservation caused the chromatin to degenerate and lead to small regions being exposed. This could be why more narrow peaks were called when using *Genrich*.

The identified peaks were grouped into four categories and annotated using *HOMER*. The majority of peaks were assigned to intronic and intergenic regions for most of the categories. This was expected as *cis*-regulatory elements are known to be found in these regions. The consensus peak sets from both the *MACS2* and *Genrich* analyses were also annotated, where the *MACS2* consensus peak set displayed a greater percentage of promoter regions compared to the *Genrich* consensus peak set. Another study conducted RNA-seq and ATAC-seq to generate a transcriptomic atlas of the rhesus macaque brain (Yin *et al.*, 2020). They used *Genrich* to call peaks and *ChIPseeker* was used to annotate these peaks. They found that the majority of peaks enriched in promoter regions were region-specific and peaks that were highly enriched in enhancer regions were conserved ATAC peaks (shared by multiple brain regions). In addition, they also generated a consensus peak set comprised of all the identified peaks from the different brain regions, where the majority of these peaks were enriched in promoter regions (Yin *et al.*, 2020). This was not seen for our analysis, where the *Genrich* consensus peak set displayed a greater percentage of intronic and intergenic regions. However, Yin *et al*. performed their peak annotation using a different tool (*ChIPseeker*) where they defined the distance to the nearest TSS differently compared to the tool used in this study (*HOMER*) (Yin *et al.*, 2020). In their study, Yin *et al.,* defined the distance to the TSS to be between -2,500 kb to 2,500 kb, whereas in this study the default range defined by *HOMER* was used (-1kb to 100kb) (Heinz *et al.*, 2010; Yin *et al.*, 2020)

4.2.3 Differential accessibility analysis over the course of brain maturation

The purpose of the DA analysis was to identify regions in the genome that are differentially accessible over the course of brain maturation. For this analysis, four different approaches were tested. These approaches combined the two different peak calling tools with two different normalisation methods (TMM or Loess). This approach was taken because Reske *et al.* 2020 had demonstrated that the bioinformatic tools that are utilised to analyse ATAC-seq data can influence the DA analysis results (Reske, Wilson and Chandler, 2020). Indeed, the analysis showed varying results depending on what approach was used. Between the two normalisation methods used, the loess normalisation has been shown to be the more conservative compared to the TMM normalisation method (Reske, Wilson and Chandler, 2020). Certainly, the results showed that the loess normalisation method yielded fewer significant DA peaks compared to the TMM normalisation method. The total number of DA regions identified using the TMM normalisation method was between 1 – 25, 781 DA regions and 1 – 7, 128 DA regions were identified using the loess normalisation method.

In addition, Approach I yielded the highest number of DA regions for the comparison between late childhood and adulthood. These differences in the results could indicate that the DA peaks identified with TMM normalisation were actually technical noise which were eliminated with the more conservative loess normalisation approach. However, DA peaks identified with TMM normalisation could have also been true biological signal. Approach I and Approach III for the early childhood vs adulthood comparison yielded some overlapping peaks, providing some evidence that these peaks may truly be DA. In addition, many of these peaks were in fact associated with genes that play a role in brain functioning and maturation.

One of those genes was *PACS2*, which has been shown to be expressed in the brain and plays an important role in cellular homeostasis (Thomas *et al.*, 2017) (https://www.proteinatlas.org/ENSG00000179364PACS2). It does this by controlling the function of a structure called the mitochondria-associated membrane (MAM). MAMs and specifically *PACS2* are involved in Ca^{2+} transfer, mitochondrial dynamics, apoptosis, and autophagy (Mendes *et al.*, 2005; Simmen *et al.*, 2005; Li *et al.*, 2020). Certain neurological diseases such as Alzheimer's and Parkinson's diseases are characterised by damage to these cellular processes. *PACS2* has also been shown to play a role in the pathogenesis of Alzheimer's and Parkinson's disease (Hedskog *et al.*, 2013). Furthermore, a missense mutation in this gene has been associated with early infantile epileptic encephalopathy (Terrone *et al.*, 2020).

A few of these significantly DA peaks were associated with genes that play a role in oligodendrocyte function such as *OPALIN*. Oligodendrocytes are important for myelin assembly, ensheathment of axons and critical for maintaining axonal integrity (Li *et al.*, 1994). As mentioned previously, this gene is a transmembrane protein and has been shown to be involved in oligodendrocyte differentiation, where it was shown to be expressed by myelinating oligodendrocytes (Li *et al.*, 1994; de Faria *et al.*, 2019). This was expected since oligodendrocytes made up the largest proportion of the identified cell types in the snRNAseq datasets in this study and the expression of these genes were enriched in the oligodendrocyte cluster. In addition, GO terms such as "oligodendrocyte differentiation" and "myelination" were enriched in the early childhood vs adult comparison. For future analysis, increasing the sample size for each age category and testing additional normalisation methods would be required to verify and improve these results. *EdgeR* was used to conduct DA analysis, but it would be interesting to test other popular tools such as *DESeq2* (Love, Huber and Anders, 2014) and *limma* (Ritchie *et al.*, 2015) using different p-value and fold-change cut offs. The majority of significantly DA peaks for both the *Genrich* and *MACS2* peak-sets showed decreased accessibility in the younger age category, indicating that these are genomic regions that are closed in childhood and then became open in the adult.

These regions could be turned off and inaccessible during early brain development but become activated in adulthood. Perhaps these regions become activated to govern more specialised functions that have not completely developed during the earlier stages of brain development. The majority of these DA regions were also annotated as intronic or intergenic regions, thus could represent enhancer regions. De la Torre-Ubieta *et al.* performed ATAC-seq on embryonic human cortical brain tissue to study cortical neurogenesis. Here, they also utilised *MACS2* to call peaks and *DESeq2* to perform DA analysis, comparing two distinct regions, the germinal zone (GZ) and the cortical plate (CP). In their study, they identified DA peaks that were enriched in both promoter and enhancer regions (TorreUbieta *et al.*, 2018).

GO analysis of the genes associated with these peaks revealed terms associated with neurodevelopment and neurodevelopmental diseases. Interestingly, the "early childhood vs adult" comparison resulted in mostly development terms ("axon development", "positive regulation of neurogenesis", "oligodendrocyte differentiation" and "myelination"), whereas the "late childhood vs adult" comparison resulted in mostly functional terms ("vacuole organisation", "proteoglycan biosynthetic process" and "transport along microtubule"). This was unexpected as the developmental terms were associated with regions that were closed in childhood compared to adulthood. Perhaps the role of these developmental genes in the later stages of brain development are still unknown.

Finally, for some of the comparisons across age categories, all four analysis approaches yielded few or no significant DA peaks. The lack of significantly DA regions could be due to biological variability in human samples which may make the chances of finding a consistently DA region less likely than when using an inbred mouse line or human cell lines for this type of analysis. Again, an increased sample size might be required to provide increased power to detect more significant DA peaks in samples with high biological variability. Alternatively, the majority of the chromatin in the brain may already be primed for brain maturation and function, resulting in very few changes being necessary to regulate gene expression as the brain matures (Colantuoni *et al.*, 2011; Kang *et al.*, 2011; Li *et al.*, 2018). These studies used human brain tissue samples from a broad range of ages and spanning several developmental periods (from embryonic development to late adulthood). They found that the most significant changes in gene expression occur during embryonic and fetal development until it reaches a plateau during late infancy. Perhaps, the majority of changes in chromatin accessibility will also occur during the late-fetal transition corresponding to the gene expression changes, where a decrease in chromatin accessibility changes will also be observed after this period.

It is also possible that subtle or cell type-specific changes in chromatin accessibility could have also been masked because the bulk ATAC-seq method was utilised. Single cell ATAC-seq analysis would be required in the future to verify these results (Ziffra et al., 2019; Domcke et al., 2020; Trevino et al., 2021). For example, Ziffra *et al*., applied scATAC-seq to human forebrain samples at midgestation. Firstly, using the scATAC-seq data they were able to identify the major cell types of the brain such as radial glia, excitatory neurons, and interneurons (Ziffra *et al.*, 2019). Furthermore, they were able to identify thousands of loci that undergo significant cell-type specific changes in accessibility during corticogenesis (Ziffra *et al.*, 2019). For example, a *EOMES* enhancer was identified which is the cell-type specific marker for intermediate progenitor cells (IPCs). This enhancer becomes highly accessible during the early stages of cortical neurogenesis and then the enhancer shows decreased accessibility during the later stages (Ziffra *et al.*, 2019).

Overall, the analysis showed that bulk ATAC-seq libraries were successfully generated from pediatric and adult brain tissue samples and different bioinformatic pipelines were utilised to analyse the ATAC-seq data, yielding varying results. This analysis needs to be repeated with additional samples and replicates for each sample to verify the DA results. Furthermore, scATAC-seq would be useful to corroborate these results and to also investigate cell type-specific changes in gene regulation during brain development, especially in conjunction with snRNA-seq.

4.3 Conclusion

Considerable work has been done to uncover the molecular underpinnings of brain maturation. However, there is still a lot we do not know about this complex process. Through testing and optimisation of several nuclei isolation protocols, a standard nuclei isolation protocol was generated that successfully reduced the amount of cell debris associated with processing human brain tissue. These experiments also allowed for the successful generation of snRNA-seq libraries for pediatric brain tissue from two different regions of the brain. Furthermore, pilot DE analysis of the snRNA-seq datasets, highlighted several differentially expressed genes between the temporal and frontal lobes. These established protocols can now be utilised to analyse additional brain tissue samples spanning different developmental periods to obtain a better understanding of brain maturation. This study also showed the successful generation of bulk ATAC-seq libraries for pediatric and adult brain tissue. Testing of bioinformatic pipelines to analyse the bulk ATAC-seq allowed for the generation of a consensus list of putative enhancers and promoters, which showed overlap with the snRNA-seq data.

References:

10x Genomics. (2018) 'What is Cell Ranger?' Available at: https://support.10xgenomics.com/single-cell-geneexpression/software/pipelines/latest/what-is-cell-ranger (Accessed: 29 July 2019).

10x Genomics. (2017) '10x Genomics® Sample Preparation Demonstrated Protocols Isolation of Nuclei for Single Cell RNA Sequencing'. Available at: https://assets.ctfassets.net/an68im79xiti/6FhJX6yndYy0OwskGmMc8I/e2677be827e82cd954ecfb8b30278e5e/CG000124_SamplePrepDemonstratedProtocol_-_Nuclei_RevD.pdf (Accessed: 6 April 2022).

10x Genomics. (2019) 'Chromium Next GEM Single Cell 3′ Reagent Kits v3.1', 1000152(D), pp. 1–3. Available at: www.10xgenomics.com/trademarks (Accessed: 19 February 2021).

10x Genomics. (2020) 'Technical note - Interpreting Cell Ranger Web Summary Files for Single Cell Expression Assay_CG000329', pp. 1–9. Available at: https://assets.ctfassets.net/an68im79xiti/163qWiQBTVi2YLbskJphQX/e90bb82151b1cdab6d7e9b6c845e6130/CG000329_TechnicalNote_InterpretingCellRangerWebSummaryFiles_RevA.pdf (Accessed: 26 March 2022).

Abrahams, B. S. *et al.* (2007) 'Genome-wide analyses of human perisylvian cerebral cortical patterning', *Proceedings of the National Academy of Sciences*, 104(45), pp. 17849–17854. doi: 10.1073/pnas.0706128104.

Amemiya, H. M., Kundaje, A. and Boyle, A. P. (2019) 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome', *Scientific Reports*, 9(1), pp. 1–5. doi: 10.1038/s41598-019-45839-z.

Amezquita, R. A. *et al.* (2020) 'Orchestrating single-cell analysis with Bioconductor', *Nature Methods*, 17(2), pp. 137–145. doi: 10.1038/s41592-019-0654-x.

Anderson, S. A. (2002) 'Distinct Origins of Neocortical Projection Neurons and Interneurons In Vivo', *Cerebral Cortex*, 12(7), pp. 702–709. doi: 10.1093/cercor/12.7.702.

Avanzini, G. and Franceschetti, S. (2003) 'Cellular biology of epileptogenesis', *Lancet Neurology*, 2(1), pp. 33–42. doi: 10.1016/S1474-4422(03)00265-5.

Baba, H. *et al.* (1997) 'GFAP Gene Expression during Development of Astrocyte', *Developmental Neuroscience*, 19(1), pp. 49–57. doi: 10.1159/000111185.

Bae, S. and Lesch, B. J. (2020) 'H3K4me1 Distribution Predicts Transcription State and Poising at Promoters', *Frontiers in Cell and Developmental Biology*. Frontiers Media S.A., 8, p. 289. doi: 10.3389/fcell.2020.00289.

Baek, S., Goldstein, I. and Hager, G. L. (2017) 'Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity', *Cell Reports*. ElsevierCompany., 19(8), pp. 1710–1722. doi: 10.1016/j.celrep.2017.05.003.

Bakken, T. E. *et al.* (2021) 'Comparative cellular analysis of motor cortex in human, marmoset and mouse', *Nature*, 598(7879), pp. 111–119. doi: 10.1038/s41586-021-03465-8.

Beghi, E. *et al.* (2019) 'Global, regional, and national burden of epilepsy, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016', *The Lancet Neurology*, 18(4), pp. 357–375. doi: 10.1016/S1474-4422(18)30454-X.

Bernstein, B. E. *et al.* (2010) 'The NIH roadmap epigenomics mapping consortium', *Nature Biotechnology*. Nature Publishing Group, 28(10), pp. 1045–1048. doi: 10.1038/nbt1010-1045.

Boulting, G. L. *et al.* (2021) 'Activity-dependent regulome of human GABAergic neurons reveals new patterns of gene regulation and neurological disease heritability', *Nature Neuroscience*. Springer US, 24(3), pp. 437–448. doi: 10.1038/s41593-020-00786-1.

Boyle, A. P. *et al.* (2008) 'High-Resolution Mapping and Characterization of Open Chromatin across the Genome', *Cell*, 132(2), pp. 311–322. doi: 10.1016/j.cell.2007.12.014.

Buenrostro, J. *et al.* (2016) 'ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide', pp. 1–10. doi: 10.1002/0471142727.mb2129s109.ATAC-seq.

Buenrostro, J. D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*. Nature

Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 10(12), pp. 1213–1218. doi: 10.1038/nmeth.2688.

Busolin, G. *et al.* (2011) 'Association of intronic variants of the KCNAB1 gene with lateral temporal epilepsy', *Epilepsy Research*. Elsevier B.V., 94(1–2), pp. 110–116. doi: 10.1016/j.eplepsyres.2011.01.010.

Butler, A. *et al.* (2018) 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology*, 36(5), pp. 411–420. doi: 10.1038/nbt.4096.

Bystron, I., Blakemore, C. and Rakic, P. (2008) 'Development of the human cerebral cortex: Boulder Committee revisited', *Nature Reviews Neuroscience*, 9(2), pp. 110–122. doi: 10.1038/nrn2252.

Cai, Y. *et al.* (2020) 'Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis', *EBioMedicine*. Elsevier B.V., 53, p. 102686. doi: 10.1016/j.ebiom.2020.102686.

Cajal, S. R. y (1906) 'The structure and connexions of neurons. In Nobel Lectures Physiology or Medicine 1901-1921', pp. 220–253. Available at: https://www.nobelprize.org/uploads/2018/06/cajal-lecture.pdf (Accessed 6 March 2022).

Calo, E. and Wysocka, J. (2013) 'Modification of Enhancer Chromatin: What, How, and Why?', *Molecular Cell*. Elsevier Inc., 49(5), pp. 825–837. doi: 10.1016/j.molcel.2013.01.038.

Cao, J. *et al.* (2019) 'The single-cell transcriptional landscape of mammalian organogenesis', *Nature*, 566(7745), pp. 496–502. doi: 10.1038/s41586-019-0969-x.

Cao, Y., Wang, X. and Peng, G. (2020) 'SCSA: A cell type annotation tool for single-cell RNA-seq data', *Frontiers in Genetics*, 11(May), pp. 1–8. doi: 10.3389/fgene.2020.00490.

Carroll, S. B. (2003) 'Genetics and the making of Homo sapiens', *Nature*, 422(6934), pp. 849–857. doi: 10.1038/nature01495.

Carullo, N. V. N. and Day, J. J. (2019) 'Genomic enhancers in brain health and disease', *Genes*, 10(1), pp. 1–20. doi: 10.3390/genes10010043.

Cauli, B. *et al.* (1997) 'Molecular and Physiological Diversity of Cortical Nonpyramidal Cells', *The Journal of Neuroscience*, 17(10), pp. 3894–3906. doi: 10.1523/JNEUROSCI.17-10-03894.1997.

Chatterjee, S. (2011) 'Brain tuberculomas, tubercular meningitis, and post-tubercular hydrocephalus in children', *Journal of Pediatric Neurosciences*. Medknow Publications, 6(3), p. 96. doi: 10.4103/18171745.85725.

Chayer, C. and Freedman, M. (2001) 'Frontal lobe functions', *Current Neurology and Neuroscience Reports*, 1(6), pp. 547–552. doi: 10.1007/s11910-001-0060-4.

Cherian, A. and Thomas, S. (2011) 'Central nervous system tuberculosis'. African health sciences. doi: 10.4314/ahs.v11i1.65007.

Cheriyan, V. T. *et al.* (2020) 'Neurogranin regulates eNOS function and endothelial activation', *Redox Biology*. Elsevier B.V., 34(February), p. 101487. doi: 10.1016/j.redox.2020.101487.

Chiang, S. S. *et al.* (2014) 'Treatment outcomes of childhood tuberculous meningitis: a systematic review and meta-analysis', *The Lancet Infectious Diseases*, 14(10), pp. 947–957. doi: 10.1016/S1473-3099(14)70852-7.

Claes, C. *et al.* (2021) 'Plaque-associated human microglia accumulate lipid droplets in a chimeric model of Alzheimer's disease', *Molecular Neurodegeneration*. Molecular Neurodegeneration, 16(1), p. 50. doi: 10.1186/s13024-021-00473-0.

Colantuoni, C. *et al.* (2011) 'Temporal dynamics and genetic control of transcription in the human prefrontal cortex', *Nature*. Nature Publishing Group, 478(7370), pp. 519–523. doi: 10.1038/nature10524.

Cooper, J. A. (2008) 'A mechanism for inside-out lamination in the neocortex', *Trends in Neurosciences*, 31(3), pp. 113–119. doi: 10.1016/j.tins.2007.12.003.

Corces, M. R. *et al.* (2017) 'An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues', *Nature Methods*, 14(10), pp. 959–962. doi: 10.1038/nmeth.4396.

Curiel, J. *et al.* (2017) 'TUBB4A mutations result in specific neuronal and oligodendrocytic defects that closely match clinically distinct phenotypes', *Human Molecular Genetics*, 26(22), pp. 4506–4518. doi: 10.1093/hmg/ddx338.

Darmanis, S. *et al.* (2015) 'A survey of human brain transcriptome diversity at the single cell level', *Proceedings of the National Academy of Sciences of the United States of America*, 112(23), pp. 7285–7290. doi: 10.1073/pnas.1507125112.

Davidson, E. H., McClay, D. R. and Hood, L. (2003) 'Regulatory gene networks and the properties of the developmental process', *Proceedings of the National Academy of Sciences*, 100(4), pp. 1475–1480. doi: 10.1073/pnas.0437746100.

DeFelipe, J. *et al.* (2013) 'New insights into the classification and nomenclature of cortical GABAergic interneurons', *Nature Reviews Neuroscience*. Nature Publishing Group, 14(3), pp. 202–216. doi: 10.1038/nrn3444.

DeFelipe, J. and Fariñas, I. (1992) 'The pyramidal neuron of the cerebral cortex: Morphological and chemical characteristics of the synaptic inputs', *Progress in Neurobiology*, 39(6), pp. 563–607. doi: 10.1016/0301-0082(92)90015-7.

Deng, C., Daley, T. and Smith, A. (2015) 'Applications of species accumulation curves in large-scale biological data analysis', *Quantitative Biology*, 3(3), pp. 135–144. doi: 10.1007/s40484-015-0049-7.

Deng, Q. *et al.* (2014) 'Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells', *Science*, 343(6167), pp. 193–196. doi: 10.1126/science.1245316.

DePasquale, E. A. K. *et al.* (2019) 'DoubletDecon: Deconvoluting Doublets from Single-Cell RNA Sequencing Data', *Cell Reports*, 29(6), pp. 1718-1727.e8. doi: 10.1016/j.celrep.2019.09.082.

DePasquale, E. A. K. *et al.* (2020) 'Protocol for Identification and Removal of Doublets with DoubletDecon', *STAR Protocols*. The Author(s), 1(2), p. 100085. doi: 10.1016/j.xpro.2020.100085.

Derrien, T. *et al.* (2012) 'The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression', *Genome Research*, 22(9), pp. 1775–1789. doi: 10.1101/gr.132159.111.

Ding, J. *et al.* (2019) 'Systematic comparative analysis of single cell RNA-sequencing methods', *bioRxiv*, p. 632216. doi: 10.1101/632216.

Djebali, S. *et al.* (2012) 'Landscape of transcription in human cells', *Nature*, 489(7414), pp. 101–108. doi: 10.1038/nature11233.

Domcke, S. *et al.* (2020) 'A human cell atlas of fetal chromatin accessibility', *Science*, 370(6518). doi: 10.1126/science.aba7612.

Ecker, J. R. *et al.* (2017) 'The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas', *Neuron*. Elsevier Inc., 96(3), pp. 542–557. doi: 10.1016/j.neuron.2017.10.007.

ENCODE Project (2011) 'A User's Guide to the Encyclopedia of DNA Elements (ENCODE)', *PLOS Biology*. Public Library of Science, 9(4), p. e1001046. Available at: https://doi.org/10.1371/journal.pbio.1001046.

ENCODE Project (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

Erokhin, M. *et al.* (2015) 'Eukaryotic enhancers: Common features, regulation, and participation in diseases', *Cellular and Molecular Life Sciences*. Springer Basel, 72(12), pp. 2361–2375. doi: 10.1007/s00018-015-1871-9.

Fan, X. *et al.* (2018) 'Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis', *Cell Research*. Springer US, 28(7), pp. 730–745. doi: 10.1038/s41422-018-0053-3.

de Faria, O. *et al.* (2019) 'TMEM10 Promotes Oligodendrocyte Differentiation and is Expressed by Oligodendrocytes in Human Remyelinating Multiple Sclerosis Plaques', *Scientific Reports*, 9(1), p. 3606. doi: 10.1038/s41598-019-40342-x.

Farina, K. L. *et al.* (2003) 'Two ZBP1 KH domains facilitate β-actin mRNA localization, granule formation, and cytoskeletal attachment', *Journal of Cell Biology*, 160(1), pp. 77–87. doi: 10.1083/jcb.200206003.

Fiers, M. W. E. J. *et al.* (2018) 'Mapping gene regulatory networks from single-cell omics data', *Briefings in Functional Genomics*, 17(4), pp. 246–254. doi: 10.1093/bfgp/elx046.

Fishilevich, S. *et al.* (2017) 'GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards', *Database*, 2017, pp. 1–17. doi: 10.1093/database/bax028.

Fu, G. *et al.* (2021) 'A potential association of RNF219 - AS1 with ADHD: Evidence from categorical analysis of clinical phenotypes and from quantitative exploration of executive function and white matter microstructure endophenotypes', *CNS Neuroscience & Therapeutics*, 27(5), pp. 603–616. doi: 10.1111/cns.13629.

Fujiwara, S. *et al.* (2019) 'High Quality ATAC-Seq Data Recovered from Cryopreserved Breast Cell Lines and Tissue', *Scientific Reports*. Springer US, 9(1), pp. 1–11. doi: 10.1038/s41598-018-36927-7.

Fullard, J. F. *et al.* (2017) 'Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci', *Human Molecular Genetics*, 26(10), pp. 1942–1951. doi: 10.1093/hmg/ddx103.

Fullard, J. F. *et al.* (2018) 'An atlas of chromatin accessibility in the adult human brain', *Genome Research*, 28(8), pp. 1243–1252. doi: 10.1101/gr.232488.117.
Gaspar, J. M. (2018a) 'Improved peak-calling with MACS2', *bioRxiv*, pp. 1–16. doi: 10.1101/496521.

Gaspar, J. M. (2018b) 'NGmerge: Merging paired-end reads via novel empirically-derived models of sequencing errors', *BMC Bioinformatics*. BMC Bioinformatics, 19(1), pp. 1–9. doi: 10.1186/s12859-0182579-2.

Gaublomme, J. T. *et al.* (2019) 'Nuclei multiplexing with barcoded antibodies for single-nucleus genomics', *Nature Communications*, 10(1), pp. 1–8. doi: 10.1038/s41467-019-10756-2.

Gensert, J. M. and Goldman, J. E. (1997) 'Endogenous Progenitors Remyelinate Demyelinated Axons in the Adult CNS', *Neuron*, 19(1), pp. 197–203. doi: 10.1016/S0896-6273(00)80359-1.

Geschwind, D. H. and Rakic, P. (2013) 'Cortical evolution: Judge the brain by its cover', *Neuron*. Elsevier Inc., 80(3), pp. 633–647. doi: 10.1016/j.neuron.2013.10.045.

Gogtay, N. *et al.* (2004) 'Dynamic mapping of human cortical development during childhood through early adulthood', *Proceedings of the National Academy of Sciences*, 101(21), pp. 8174–8179. doi: 10.1073/pnas.0402680101.

Grindberg, R. V. *et al.* (2013) 'RNA-sequencing from single nuclei', *Proceedings of the National Academy of Sciences of the United States of America*, 110(49), pp. 19802–19807. doi: 10.1073/pnas.1319700110.

Gross, D. S. and Garrard, W. T. (1988) 'Nuclease Hypersensitive Sites in Chromatin', *Annual Review of Biochemistry*, 57(1), pp. 159–197. doi: 10.1146/annurev.bi.57.070188.001111.

Guo, Q. *et al.* (2022) 'Single-cell transcriptomic landscape identifies the expansion of peripheral blood monocytes as an indicator of HIV-1-TB co-infection', *Cell Insight*. The Authors, 1(1), p. 100005. doi: 10.1016/j.cellin.2022.100005.

Habib, N. *et al.* (2017) 'Massively parallel single-nucleus RNA-seq with DroNc-seq', *Nature Methods*. Nature Publishing Group, 14(10), pp. 955–958. doi: 10.1038/nmeth.4407.

Hawrylycz, M. J. *et al.* (2012) 'An anatomically comprehensive atlas of the adult human brain transcriptome', *Nature*, 489(7416), pp. 391–399. doi: 10.1038/nature11405.

Hedskog, L. *et al.* (2013) 'Modulation of the endoplasmic reticulum-mitochondria interface in Alzheimer's disease and related models', *Proceedings of the National Academy of Sciences of the United States of America*, 110(19), pp. 7916–7921. doi: 10.1073/pnas.1300677110.

Heinz, S. *et al.* (2010) 'Simple Combinations of Lineage-Determining Transcription Factors Prime cisRegulatory Elements Required for Macrophage and B Cell Identities', *Molecular Cell*. Elsevier Inc., 38(4), pp. 576–589. doi: 10.1016/j.molcel.2010.05.004.

Heller, M. J. (2002) 'DNA Microarray Technology: Devices, Systems, and Applications', *Annual Review of Biomedical Engineering*, 4(1), pp. 129–153. doi: 10.1146/annurev.bioeng.4.020702.153438.

Heng, T. S. P. *et al.* (2008) 'The immunological genome project: Networks of gene expression in immune cells', *Nature Immunology*, 9(10), pp. 1091–1094. doi: 10.1038/ni1008-1091.

Henry, A. M. and Hohmann, J. G. (2012) 'High-resolution gene expression atlases for adult and Developing Mouse Brain and Spinal Cord', *Mammalian Genome*, 23(9–10), pp. 539–549. doi: 10.1007/s00335-0129406-2.

Hodge, R. D. *et al.* (2019) 'Conserved cell types with divergent features in human versus mouse cortex', *Nature*. Springer US, 573(7772), pp. 61–68. doi: 10.1038/s41586-019-1506-7.

Hu, H. *et al.* (2001) 'Presynaptic Ca 2+ -Activated K + Channels in Glutamatergic Hippocampal Terminals and Their Role in Spike Repolarization and Regulation of Transmitter Release', *The Journal of Neuroscience*, 21(24), pp. 9585–9597. doi: 10.1523/JNEUROSCI.21-24-09585.2001.

Hu, P. *et al.* (2017) 'Dissecting Cell-Type Composition and Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel Single-Nucleus RNA-Seq', *Molecular Cell*. Elsevier Inc., 68(5), pp. 1006-1015.e7. doi: 10.1016/j.molcel.2017.11.017.

Huang, K.-P. (2004) 'Neurogranin/RC3 Enhances Long-Term Potentiation and Learning by Promoting Calcium-Mediated Signaling', *Journal of Neuroscience*, 24(47), pp. 10660–10669. doi: 10.1523/JNEUROSCI.2213-04.2004.

Hwang, B., Lee, J. H. and Bang, D. (2018) 'Single-cell RNA sequencing technologies and bioinformatics pipelines', *Experimental and Molecular Medicine*. Springer US, 50(8). doi: 10.1038/s12276-018-0071-8.

Jäkel, S. and Dimou, L. (2017) 'Glial cells and their function in the adult brain: A journey through the history of their ablation', *Frontiers in Cellular Neuroscience*, 11(February), pp. 1–17. doi: 10.3389/fncel.2017.00024.

Johnson, M. B. *et al.* (2009) 'Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis', *Neuron*, 62(4), pp. 494–509. doi: 10.1016/j.neuron.2009.03.027.

Julius, M. H., Masuda, T. and Herzenberg, L. A. (1972) 'Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter.', *Proceedings of the National Academy of Sciences of the United States of America*, 69(7), pp. 1934–1938. doi: 10.1073/pnas.69.7.1934.

Kalachikov, S. *et al.* (2002) 'Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features', *Nature Genetics*, 30(3), pp. 335–341. doi: 10.1038/ng832.

Kang, H. J. *et al.* (2011) 'Spatio-temporal transcriptome of the human brain', *Nature*, 478(7370), pp. 483–489. doi: 10.1038/nature10523.

Kang, H. M. *et al.* (2018) 'Multiplexed droplet single-cell RNA-sequencing using natural genetic variation', *Nature Biotechnology*. Nature Publishing Group, 36(1), pp. 89–94. doi: 10.1038/nbt.4042.

Kimelberg, H. K. and Nedergaard, M. (2010) 'Functions of Astrocytes and their Potential As Therapeutic Targets', *Neurotherapeutics*, 7(4), pp. 338–353. doi: 10.1016/j.nurt.2010.07.006.

Ko, E. A. *et al.* (2008) 'Physiological roles of K+ channels in vascular smooth muscle cells', *Journal of Smooth Muscle Research*, 44(2), pp. 65–81. doi: 10.1540/jsmr.44.65.

Kornack, D. R. and Rakic, P. (1998) 'Changes in cell-cycle kinetics during the development and evolution of primate neocortex', *Proceedings of the National Academy of Sciences of the United States of America*, 95(3), pp. 1242–1246. doi: 10.1073/pnas.95.3.1242.

Krishnaswami, S. R. *et al.* (2016) 'Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons', *Nature Protocols*. Nature Publishing Group, 11(3), pp. 499–524. doi: 10.1038/nprot.2016.015.

Kubbies, M. (1990) 'Flow cytometric recognition of clastogen induced chromatin damage in G0/G1 lymphocytes by non-stoichiometric Hoechst fluorochrome binding', *Cytometry*, 11(3), pp. 386–394. doi: 10.1002/cyto.990110309.

Kuhn, S. *et al.* (2019) 'Oligodendrocytes in Development, Myelin Generation and Beyond', *Cells*. doi: 10.3390/cells8111424.

Kundaje, A. *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*. The Author(s), 518, p. 317. Available at: https://doi.org/10.1038/nature14248.

Lacar, B. *et al.* (2016) 'Nuclear RNA-seq of single neurons reveals molecular signatures of activation', *Nature Communications*, 7(1), p. 11022. doi: 10.1038/ncomms11022.

Lake, B. B. *et al.* (2016) 'Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain', *Science*, 352(6293), pp. 1586–1590.

Lake, B. B. *et al.* (2018) 'Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain', *Nature Biotechnology*, 36(1), pp. 70–80. doi: 10.1038/nbt.4038.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 9, p. 357. Available at: https://doi.org/10.1038/nmeth.1923.

Latt, S. A. *et al.* (1975) 'Recent developments in the detection of deoxyribonucleic acid synthesis by 33258 Hoechst fluorescence.', *Journal of Histochemistry & Cytochemistry*, 23(7), pp. 493–505. doi: 10.1177/23.7.1095650.

Leicher, T. *et al.* (1996) 'Structural and Functional Characterization of Human Potassium Channel Subunit β1 (KCNA1B)', *Neuropharmacology*, 35(7), pp. 787–795. doi: 10.1016/0028-3908(96)00133-5.

Li, C. *et al.* (1994) 'Myelination in the absence of myelin-associated glycoprotein', *Nature*, 369(6483), pp. 747–750. doi: 10.1038/369747a0.

Li, C. *et al.* (2020) 'PACS-2: A key regulator of mitochondria-associated membranes (MAMs)', *Pharmacological Research*, 160(139). doi: 10.1016/j.phrs.2020.105080.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, M. *et al.* (2018) 'Integrative functional genomic analysis of human brain development and neuropsychiatric risks', *Science*, 362(6420). doi: 10.1126/science.aat7615.

Li, Q. *et al.* (2019) 'Developmental Heterogeneity of Microglia and Brain Myeloid Cells Revealed by Deep Single-Cell RNA Sequencing', *Neuron*. Elsevier Inc., 101(2), pp. 207-223.e10. doi: 10.1016/j.neuron.2018.12.006.

Lodato, S. and Arlotta, P. (2015) 'Generating Neuronal Diversity in the Mammalian Cerebral Cortex', *Annual Review of Cell and Developmental Biology*, 31(1), pp. 699–720. doi: 10.1146/annurev-cellbio100814-125353.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lun, A. T. L. and Smyth, G. K. (2015) 'Csaw: A Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows', *Nucleic Acids Research*, 44(5), p. e45. doi: 10.1093/nar/gkv1191.

Luthi-Carter, R. and Cha, J. H. J. (2003) 'Mechanisms of transcriptional dysregulation in Huntington's disease', *Clinical Neuroscience Research*, 3(3), pp. 165–177. doi: 10.1016/S1566-2772(03)00059-8.

Ma, S. and Zhang, Y. (2020) 'Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq', *Molecular Biomedicine*. Molecular Biomedicine, 1(1), p. 9. doi: 10.1186/s43556-020-00009-w.

MacNeil, L. T. and Walhout, A. J. M. (2011) 'Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression', *Genome Research*, 21(6), p. 999. doi: 10.1101/gr.097378.109.21.

Macosko, E. Z. *et al.* (2015) 'Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets', *Cell*. Elsevier, 161(5), pp. 1202–1214. doi: 10.1016/j.cell.2015.05.002.

Mantione, K. J. *et al.* (2014) 'Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq', *Medical Science Monitor Basic Research*, 20, pp. 138–142. doi: 10.12659/msmbr.892101.

Marioni, J. C. *et al.* (2019) 'RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays', *Genome Research*, 18(9), pp. 1509–1517. doi: 10.1101/gr.079558.108.

Martínez-Cerdeño, V. and Noctor, S. C. (2018) 'Neural progenitor cell terminology', *Frontiers in Neuroanatomy*, 12(December), pp. 1–8. doi: 10.3389/fnana.2018.00104.

McGinnis, C. S., Murrow, L. M. and Gartner, Z. J. (2019) 'DoubletFinder: Doublet Detection in SingleCell RNA Sequencing Data Using Artificial Nearest Neighbors', *Cell Systems*. Elsevier Inc., 8(4), pp. 329337.e4. doi: 10.1016/j.cels.2019.03.003.

McKenzie, A. T. *et al.* (2018) 'Brain Cell Type Specific Gene Expression and Co-expression Network Architectures', *Scientific Reports*. Springer US, 8(1), pp. 1–19. doi: 10.1038/s41598-018-27293-5.

Mendes, C. C. P. *et al.* (2005) 'The Type III Inositol 1,4,5-Trisphosphate Receptor Preferentially Transmits Apoptotic Ca2+ Signals into Mitochondria', *Journal of Biological Chemistry*. Â© 2005 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 280(49), pp. 40892–40900. doi: 10.1074/jbc.M506623200.

Meng, G. and Mei, H. (2019) 'Transcriptional dysregulation study reveals a core network involving the progression of Alzheimer's disease', *Frontiers in Aging Neuroscience*, 11(MAY), pp. 1–16. doi: 10.3389/fnagi.2019.00101.

Michelhaugh, S. K. *et al.* (2011) 'Mining Affymetrix microarray data for long non-coding RNAs: altered expression in the nucleus accumbens of heroin abusers', *Journal of Neurochemistry*, 116(3), pp. 459–466. doi: 10.1111/j.1471-4159.2010.07126.x.

Milani, P. *et al.* (2016) 'Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells', *Scientific Reports*. Nature Publishing Group, 6(May), pp. 1–10. doi: 10.1038/srep25474.

Mok, S. (2019) *Isolation of Nuclei from Adult Human Brain Tissue for 10x Genomics Platform*, *Proticols.io*. Available at: https://www.protocols.io/view/isolation-of-nuclei-from-adult-human-brain-tissue-j8nlk56k1l5r/v1 (Accessed: 29 March 2022).

Molyneaux, B. J. *et al.* (2007) 'Neuronal subtype specification in the cerebral cortex', *Nature Reviews Neuroscience*, 8(6), pp. 427–437. doi: 10.1038/nrn2151.

Moore, J. E. *et al.* (2020) 'Expanded encyclopedias of DNA elements in the human and mouse genomes', *Nature*, 583(7818), pp. 699–710. doi: 10.1038/s41586-020-2493-4.

Morabito, S. *et al.* (2021) 'Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease', *Nature Genetics*. Springer US, 53(8), pp. 1143–1155. doi: 10.1038/s41588-02100894-z.

Morante-Redolat, J. M. (2002) 'Mutations in the LGI1/Epitempin gene on 10q24 cause autosomal dominant lateral temporal epilepsy', *Human Molecular Genetics*, 11(9), pp. 1119–1128. doi: 10.1093/hmg/11.9.1119.

Nagy, C. *et al.* (2020) 'Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons', *Nature Neuroscience*. Springer US, 23(6), pp. 771–781. doi: 10.1038/s41593-020-0621-y.

Nowakowski, T. J. *et al.* (2017) 'Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex', *Science*, 358(6368), pp. 1318–1323. doi: 10.1126/science.aap8809.

Ofengeim, D. *et al.* (2017) 'Single-Cell RNA Sequencing: Unraveling the Brain One Cell at a Time', *Trends in Molecular Medicine*. Elsevier Ltd, 23(6), pp. 563–576. doi: 10.1016/j.molmed.2017.04.006.

Olah, M. *et al.* (2020) 'Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease', *Nature Communications*. Springer US, 11(1), p. 6129. doi: 10.1038/s41467-02019737-2.

Ou, J. *et al.* (2018) 'ATACseqQC: A Bioconductor package for post-alignment quality assessment of ATAC-seq data', *BMC Genomics*. BMC Genomics, 19(1), pp. 1–13. doi: 10.1186/s12864-018-4559-3.

Owolabi, L. F. *et al.* (2020) 'Prevalence of active epilepsy, lifetime epilepsy prevalence, and burden of epilepsy in Sub-Saharan Africa from meta-analysis of door-to-door population-based surveys', *Epilepsy and Behavior*. Elsevier Inc., 103, p. 106846. doi: 10.1016/j.yebeh.2019.106846.

Parikshak, N. N. *et al.* (2013) 'Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism', *Cell*. Elsevier Inc., 155(5), pp. 1008–1021. doi: 10.1016/j.cell.2013.10.031.

Pfisterer, U. *et al.* (2020) 'Identification of epilepsy-associated neuronal subtypes and gene expression underlying epileptogenesis', *Nature Communications*, 11(1), p. 5038. doi: 10.1038/s41467-020-18752-7.

Picelli, S. *et al.* (2013) 'Smart-seq2 for sensitive full-length transcriptome profiling in single cells', *Nature Methods*, 10(11), pp. 1096–1100. doi: 10.1038/nmeth.2639.

Polioudakis, D. *et al.* (2018) 'A single cell transcriptomic analysis of human neocortical development Authors': *bioRxiv*, pp. 1–26. doi: https://doi.org/10.1101/401885.

Polioudakis, D. *et al.* (2019) 'A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation', *Neuron*. Elsevier Inc., 103(5), pp. 785-801.e8. doi: 10.1016/j.neuron.2019.06.011.

Pollen, A. A. *et al.* (2014) 'Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex', *Nature Biotechnology*, 32(10), pp. 1053–1058. doi: 10.1038/nbt.2967.

Poulain, F. E. and Sobel, A. (2010) 'The microtubule network and neuronal morphogenesis: Dynamic and coordinated orchestration through multiple players', *Molecular and Cellular Neuroscience*. Elsevier Inc., 43(1), pp. 15–32. doi: 10.1016/j.mcn.2009.07.012.

Qiu, X. *et al.* (2017) 'Single-cell mRNA quantification and differential analysis with Census', *Nature Methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 14, p. 309. Available at: https://doi.org/10.1038/nmeth.4150.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Rakic, P. and Lombroso, P. J. (1998) 'Development of the cerebral cortex: I. Forming the cortical structure', *Journal of the American Academy of Child and Adolescent Psychiatry*. The American Academy of Child and Adolescent Psychiatry, 37(1), pp. 116–117. doi: 10.1097/00004583-199801000-00026.

Regev, A. *et al.* (2017) 'The Human Cell Atlas', *bioRxiv*. doi: http://dx.doi.org/10.1101/121202.

Regev, A. *et al.* (2018) 'The Human Cell Atlas White Paper', *eLife*, 6. Available at: https://www.humancellatlas.org/files/NIH_reponse_regev.pdf.

Represa, A. *et al.* (1990) 'Neurogranin: immunocytochemical localization of a brain-specific protein kinase C substrate', *The Journal of Neuroscience*, 10(12), pp. 3782–3792. doi: 10.1523/JNEUROSCI.10-1203782.1990.

Reske, J. J., Wilson, M. R. and Chandler, R. L. (2020) 'ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation', *Epigenetics and Chromatin*. BioMed Central, 13(1), pp. 1–17. doi: 10.1186/s13072-020-00342-y.

Rich, A. R. and McCordock, H. (1944) 'The Pathogenesis of Tuberculosis'.

Richmond, T. J. and Davey, C. A. (2003) 'The structure of DNA in the nucleosome core', *Nature*, 423(6936), pp. 145–150. doi: 10.1038/nature01595.

Rickman, D. S. *et al.* (2001) 'The gene for the axonal cell adhesion molecule TAX-1 is amplified and aberrantly expressed in malignant gliomas', *Cancer Research*, 61(5), pp. 2162–2168.

Ripke, S. *et al.* (2013) 'Genome-wide association analysis identifies 13 new risk loci for schizophrenia', *Nature Genetics*, 45(10), pp. 1150–1159. doi: 10.1038/ng.2742.

Ritchie, M. E. *et al.* (2015) 'limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*, 43(7), pp. e47–e47. doi: 10.1093/nar/gkv007.

Roberts, T. C., Morris, K. V. and Wood, M. J. A. (2014) 'The role of long non-coding RNAs in neurodevelopment, brain function and neurological disease', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1652), p. 20130507. doi: 10.1098/rstb.2013.0507.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2009) 'edgeR: A Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26(1), pp. 139–140. doi: 10.1093/bioinformatics/btp616.

Robinson, M. D. and Oshlack, A. (2010) 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biology*, 11(3), p. R25. doi: 10.1186/gb-2010-11-3-r25.

Rubenstein, John L.R. (2011) 'Annual research review: Development of the cerebral cortex: Implications for neurodevelopmental disorders', *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 52(4), pp. 339–355. doi: 10.1111/j.1469-7610.2010.02307.x.

Sanguinetti, G. and Walker, J. M. (2019) *Gene Regulatory Networks*. 1st edn. Edited by G. Sanguinetti and V. A. Huynh-Thu. New York, NY: Springer New York (Methods in Molecular Biology). doi: 10.1007/9781-4939-8882-2.

Sase, S. *et al.* (2020) 'TUBB4A mutations result in both glial and neuronal degeneration in an H-ABC leukodystrophy mouse model', *eLife*, 9, pp. 1–28. doi: 10.7554/eLife.52986.

Scharer, C. D. *et al.* (2016) 'ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells', *Scientific Reports*. Nature Publishing Group, 6(May), pp. 1–9. doi: 10.1038/srep27030.

Schulte, U. *et al.* (2006) 'The Epilepsy-Linked Lgi1 Protein Assembles into Presynaptic Kv1 Channels and Inhibits Inactivation by Kvβ1', *Neuron*, 49(5), pp. 697–706. doi: 10.1016/j.neuron.2006.01.033.

See, P. *et al.* (2018) 'A Single-Cell Sequencing Guide for Immunologists', *Frontiers in Immunology*, 9(OCT), pp. 1–13. doi: 10.3389/fimmu.2018.02425.

Seo, J. H. *et al.* (2014) 'Oligodendrocyte Precursor Cells Support Blood-Brain Barrier Integrity via TGF-β Signaling', *PLoS ONE*. Edited by C. V. Borlongan, 9(7), p. e103174. doi: 10.1371/journal.pone.0103174.

Shao, X. *et al.* (2020) 'scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data', *iScience*, 23(3). doi: 10.1016/j.isci.2020.100882.

Shinnar, S. and Pellock, J. M. (2002) 'Update on the epidemiology and prognosis of pediatric epilepsy', *Journal of Child Neurology*. doi: 10.1177/08830738020170010201.

Silbereis, J. C. *et al.* (2016) 'The Cellular and Molecular Landscapes of the Developing Human Central Nervous System', *Neuron*. Elsevier Ltd, 89(2), p. 248. doi: 10.1016/j.neuron.2015.12.008.

Simmen, T. *et al.* (2005) 'PACS-2 controls endoplasmic reticulum–mitochondria communication and Bidmediated apoptosis', *The EMBO Journal*, 24(4), pp. 717–729. doi: 10.1038/sj.emboj.7600559.

Simons, C. *et al.* (2013) 'A De Novo Mutation in the β-Tubulin Gene TUBB4A Results in the Leukoencephalopathy Hypomyelination with Atrophy of the Basal Ganglia and Cerebellum', *The American Journal of Human Genetics*, 92(5), pp. 767–773. doi: 10.1016/j.ajhg.2013.03.018.

Simons, M. and Trajkovic, K. (2006) 'Neuron-glia communication in the control of oligodencrocyte function and myelin biogenesis', *Journal of Cell Science*, 119(21), pp. 4381–4389. doi: 10.1242/jcs.03242.

Sjöstedt, E. *et al.* (2020) 'An atlas of the protein-coding genes in the human, pig, and mouse brain', *Science*, 367(6482). doi: 10.1126/science.aay5947.

Smajic, S. *et al.* (2020) 'Single-cell sequencing of the human midbrain reveals glial activation and a neuronal state specific to Parkinson's disease', *medRxiv*, p. 2020.09.28.20202812. doi: https://doi.org/10.1101/2020.09.28.20202812.

Song, L. *et al.* (2021) 'STAB: A spatio-temporal cell atlas of the human brain', *Nucleic Acids Research*. Oxford University Press, 49(D1), pp. D1029–D1037. doi: 10.1093/nar/gkaa762.

Starke, J. R. (1999) 'Tuberculosis of the central nervous system in children', *Seminars in Pediatric Neurology*, 6(4), pp. 318–331. doi: https://doi.org/10.1016/S1071-9091(99)80029-1.

Stiles, J. and Jernigan, T. L. (2010) 'The basics of brain development', *Neuropsychology Review*, 20(4), pp. 327–348. doi: 10.1007/s11065-010-9148-4.

Stogmann, E. *et al.* (2013) 'Autosomal recessive cortical myoclonic tremor and epilepsy: association with a mutation in the potassium channel associated gene CNTN2', *Brain*, 136(4), pp. 1155–1160. doi: 10.1093/brain/awt068.

Stuart, T. *et al.* (2018) 'Comprehensive integration of single cell data', *bioRxiv*, pp. 1–24.

Stuart, T. *et al.* (2019) 'Comprehensive Integration of Single-Cell Data', *Cell*. Elsevier Inc., 177(7), pp. 1888-1902.e21. doi: 10.1016/j.cell.2019.05.031.

Su, Y. *et al.* (2017) 'Neuronal activity modifies the chromatin accessibility landscape in the adult brain', *Nature Neuroscience*, 20(3), pp. 476–483. doi: 10.1038/nn.4494.

Takaku, M. *et al.* (2016) 'GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler', *Genome Biology*. Genome Biology, 17(1), p. 36. doi: 10.1186/s13059-016-0897-0.

Tan, M. C. *et al.* (2017) 'The Activity-Induced Long Non-Coding RNA Meg3 Modulates AMPA Receptor Surface Expression in Primary Cortical Neurons', *Frontiers in Cellular Neuroscience*, 11(May), pp. 1–12. doi: 10.3389/fncel.2017.00124.

Tang, F. *et al.* (2009) 'mRNA-Seq whole-transcriptome analysis of a single cell', *Nature Methods*, 6(5), pp. 377–382. doi: 10.1038/nmeth.1315.

Taylor, D. M. *et al.* (2019) 'The Pediatric Cell Atlas: Defining the Growth Phase of Human Development at Single-Cell Resolution', *Developmental Cell*, 49(1), pp. 10–29. doi: 10.1016/j.devcel.2019.03.001.

Tebbenkamp, A. T. N. *et al.* (2014) 'The developmental transcriptome of the human brain: implications for neurodevelopmental disorders', *Current opinion in neurology*, 27(2), pp. 149–156. doi: 10.1097/WCO.0000000000000069.

Terrone, G. *et al.* (2020) 'A further contribution to the delineation of epileptic phenotype in PACS2-related syndrome', *Seizure*. Elsevier, 79(February), pp. 53–55. doi: 10.1016/j.seizure.2020.05.001.

Thomas, G. *et al.* (2017) 'Caught in the act – protein adaptation and the expanding roles of the PACS proteins in tissue homeostasis and disease', *Journal of Cell Science*, 130(11), pp. 1865–1876. doi: 10.1242/jcs.199463.

Thrupp, N. *et al.* (2020) 'Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans', *CellReports*. ElsevierCompany., 32(13), p.108189. doi: 10.1016/j.celrep.2020.108189.

Thurman, R. E. *et al.* (2012a) 'The accessible chromatin landscape of the human genome', *Nature*. Nature Publishing Group, 489(7414), pp. 75–82. doi: 10.1038/nature11232.

Torre-Ubieta, L. de la *et al.* (2018) 'The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis', *Cell*. Elsevier Inc., 172(1–2), pp. 289-295.e18. doi: 10.1016/j.cell.2017.12.014.

Trapnell, C. *et al.* (2014) 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nature Biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 32, p. 381. Available at: https://doi.org/10.1038/nbt.2859.

Trevino, A. E. *et al.* (2021) 'Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution', *Cell*, 184(19), pp. 5053-5069.e23. doi: 10.1016/j.cell.2021.07.039.

Trinh, L. A. *et al.* (2017) 'Biotagging of Specific Cell Populations in Zebrafish Reveals Gene Regulatory Logic Encoded in the Nuclear Transcriptome', *Cell Reports*, 19(2), pp. 425–440. doi: 10.1016/j.celrep.2017.03.045.

Tung, P.-Y. *et al.* (2017) 'Batch effects and the effective design of single-cell gene expression studies', *Scientific Reports*. Nature Publishing Group, 7(1), p. 39921. doi: 10.1038/srep39921.

Typlt, M. *et al.* (2013) 'Mice with Deficient BK Channel Function Show Impaired Prepulse Inhibition and Spatial Learning, but Normal Working and Spatial Reference Memory', *PLoS ONE*. Edited by V. Ceña, 8(11), p. e81270. doi: 10.1371/journal.pone.0081270.

Wang, J. *et al.* (2017) 'Epilepsy-associated genes', *Seizure*, 44, pp. 11–20. doi: 10.1016/j.seizure.2016.11.030.

Van Well, G. T. J. *et al.* (2009) 'Twenty years of pediatric tuberculous meningitis: A retrospective cohort study in the western cape of south africa', *Pediatrics*, 123(1). doi: 10.1542/peds.2008-1353.

Werling, D. M. *et al.* (2020) 'Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex', *Cell Reports*, 31(1). doi: 10.1016/j.celrep.2020.03.053.

Wilkinson, R. J. *et al.* (2017) 'Tuberculous meningitis', *Nature Reviews Neurology*. Nature Publishing Group, 13(10), pp. 581–598. doi: 10.1038/nrneurol.2017.120.

Wu, T. *et al.* (2021) 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data', *The Innovation*. Elsevier Inc., 2(3), p. 100141. doi: 10.1016/j.xinn.2021.100141.

Yan, F. *et al.* (2020) 'From reads to insight: a hitchhiker's guide to ATAC-seq data analysis', *Genome Biology*. Genome Biology, 21(1), p. 22. doi: 10.1186/s13059-020-1929-3.

Yin, S. *et al.* (2020) 'Transcriptomic and open chromatin atlas of high-resolution anatomical regions in the rhesus macaque brain', *Nature Communications*. Springer US, 11(1), p. 474. doi: 10.1038/s41467-02014368-z.

Yip, K. Y. *et al.* (2012) 'Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors', *Genome Biology*, 13(9), p. R48. doi: 10.1186/gb-2012-13-9-r48.

Yu, G. *et al.* (2012) 'ClusterProfiler: An R package for comparing biological themes among gene clusters', *OMICS A Journal of Integrative Biology*, 16(5), pp. 284–287. doi: 10.1089/omi.2011.0118.

Yuan, H. *et al.* (2019) 'CancerSEA: a cancer single-cell state atlas', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D900–D908. doi: 10.1093/nar/gky939.

Yuste, R. *et al.* (2020) 'A community-based transcriptomics classification and nomenclature of neocortical cell types', *Nature Neuroscience*, 23(12), pp. 1456–1468. doi: 10.1038/s41593-020-0685-8.

Zeisel, A. *et al.* (2015) 'Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq', *science*, 347(6226), pp. 0–5.

Zhang *et al.* (2019) 'SCINA: Semi-Supervised Analysis of Single Cells in Silico', *Genes*, 10(7), p. 531. doi: 10.3390/genes10070531.

Zhang, A. W. *et al.* (2019) 'Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling', *Nature Methods*, 16(10), pp. 1007–1015. doi: 10.1038/s41592-019-0529-1.

Zhang, X.-Q. *et al.* (2017) 'Spatial-temporal transcriptional dynamics of long non-coding RNAs in human brain', *Human Molecular Genetics*, 26(16), pp. 3202–3211. doi: 10.1093/hmg/ddx203.

Zhang, X. *et al.* (2019) 'CellMarker: a manually curated resource of cell markers in human and mouse', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D721–D728. doi: 10.1093/nar/gky900.

Zhang, Yujia *et al.* (2015) 'Gene Mutation Analysis in 253 Chinese Children with Unexplained Epilepsy and Intellectual/Developmental Disabilities', *PLOS ONE*. Edited by M. S. Shapiro, 10(11), p. e0141782. doi: 10.1371/journal.pone.0141782.

Zhao, Q. (2002) 'A mitochondrial specific stress response in mammalian cells', *The EMBO Journal*, 21(17), pp. 4411–4419. doi: 10.1093/emboj/cdf445.

Zheng, G. X. Y. *et al.* (2017) 'Massively parallel digital transcriptional profiling of single cells', *Nature Communications*. Nature Publishing Group, 8. doi: 10.1038/ncomms14049.

Zhong, L. *et al.* (2009) 'Neurogranin enhances synaptic strength through its interaction with calmodulin', *EMBO Journal*. Nature Publishing Group, 28(19), pp. 3027–3039. doi: 10.1038/emboj.2009.236.

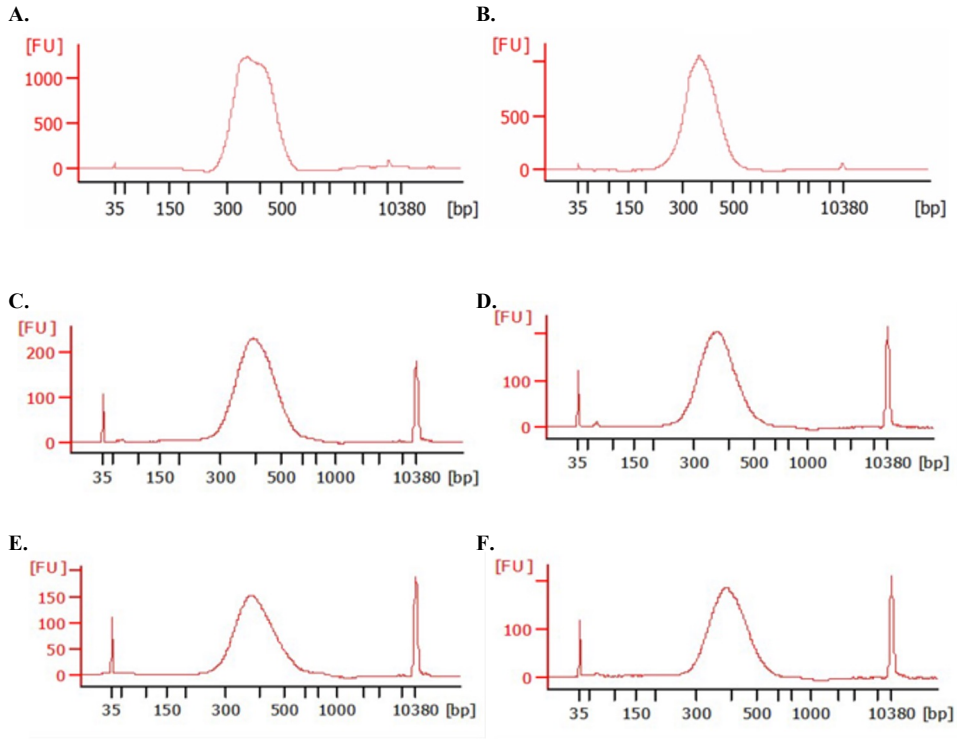
Zhong, S. *et al.* (2018) 'A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex', *Nature*, 555(7697), pp. 524–528. doi: 10.1038/nature25980.

Zhou, Y., Zhang, X. and Klibanski, A. (2012) 'MEG3 noncoding RNA: a tumor suppressor', *Journal of Molecular Endocrinology*, 48(3), pp. R45–R53. doi: 10.1530/JME-12-0008.

Zhu, Z. *et al.* (2016) 'Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets', *Nature Genetics*, 48(5), pp. 481–487. doi: 10.1038/ng.3538.

Ziffra, R. S. *et al.* (2019) 'Single cell epigenomic atlas of the developing human brain and organoids', *bioRxiv*, 11, p. 2019.12.30.891549. doi: https://doi.org/10.1101/2019.12.30.891549.

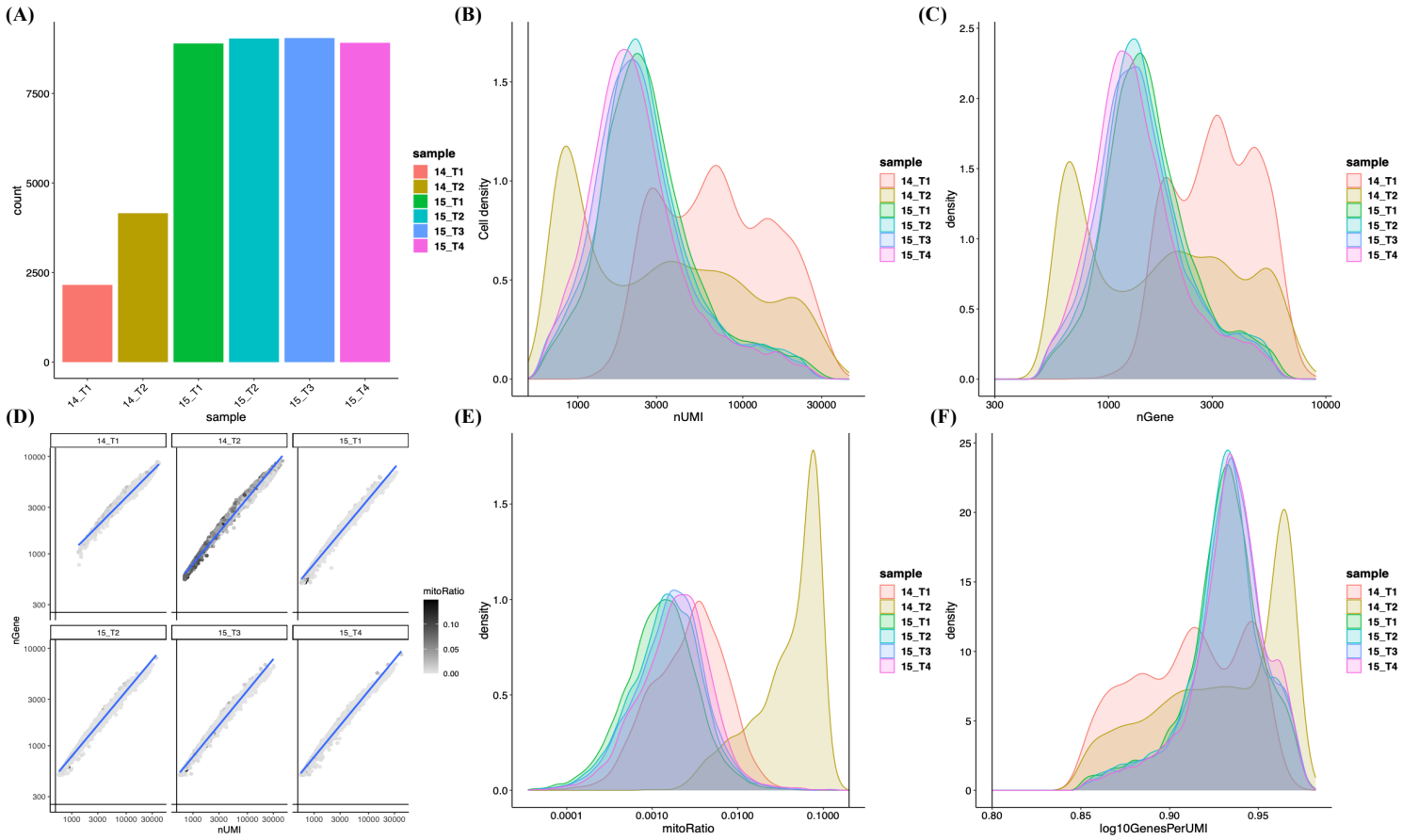
Zoupi, L. *et al.* (2018) 'The function of contactin-2/TAG-1 in oligodendrocytes in health and demyelinating pathology', *Glia*, 66(3), pp. 576–591. doi: 10.1002/glia.23266.



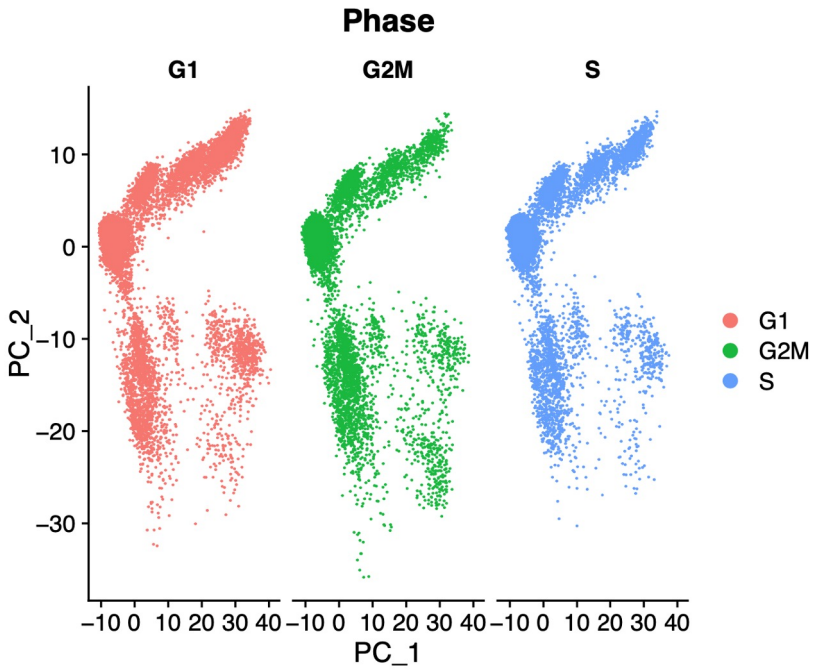
Supplementary figure 1: Post cDNA library construction fragment distribution. After preparing the snRNA-seq libraries, bioanalyzer QC plots were generated to check the fragment size distribution for 14-year-old technical replicate 1 **(A)**, 2 **(B)**, and the 15-year-old technical replicate 1 **(C)**, 2 **(D)**, 3 **(E)** and 4 **(F)** respectively.

Supplementary table 1: Cell ranger summary statistics for the 14-year-old and 15-year-old samples. Table outlining the Cell Ranger count results for each replicate. T, Technical replicate.

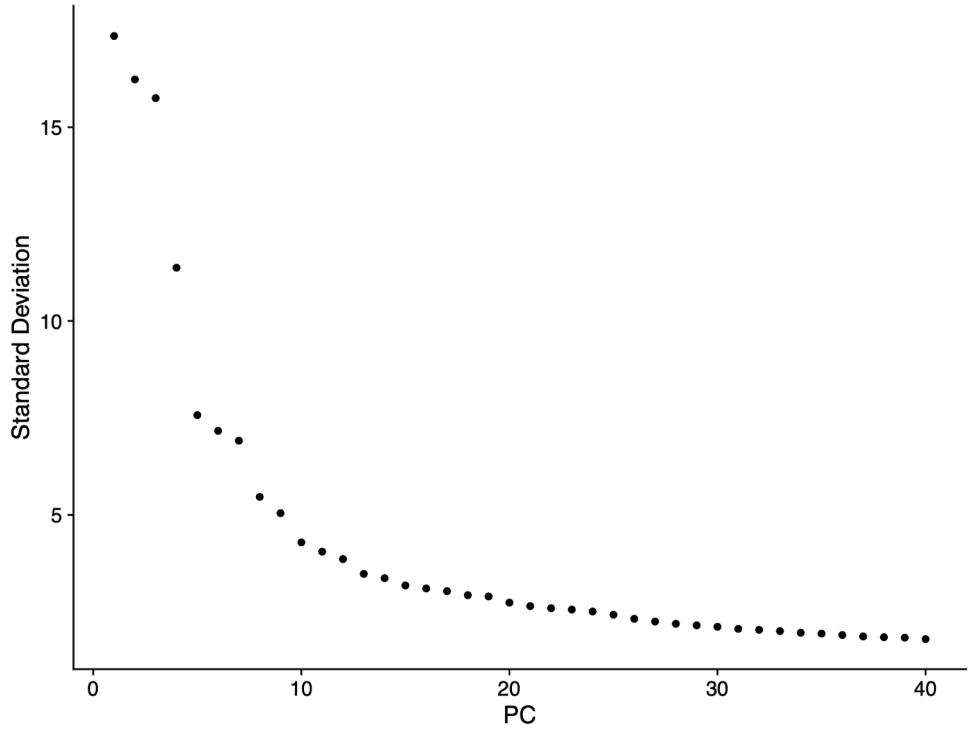
	14 T1	14 T2	15 T1	15 T2	15 T3	15 T4
Estimated number of nuclei	2 354	4 517	8 985	9 082	8 972	8 791
Total number of reads	275 606 347	268 194 186	180 011 299	175 466 165	155 018 682	161 442 044
Mean reads per nuclei	117 080	59 374	20 035	19 320	17 278	18 364
Median genes per nuclei	3 360	1 956	1 190	1 139	1 106	1 048
Median UMI counts per nuclei	7 542	3 491	2 037	1 942	1 860	1 735
Valid barcodes	97.80%	98%	98%	98.10%	98.20%	98.10%
Valid UMIs	99.80%	99.70%	100%	100%	100%	100%
Fraction of reads in nuclei	45.20%	66.30%	60.20%	61.30%	59%	60.70%
Reads mapped to genome	89.90%	85.90%	77.70%	77.10%	78.20%	78.00%
Reads mapped confidently to genome	84.90%	79.90%	75.30%	74.90%	75.90%	75.70%
Reads mapped to intergenic regions	7.30%	9.20%	7.90%	8.00%	8.10%	8.10%
Reads mapped to intronic regions	0%	0%	0%	0%	0%	0%
Reads mapped to exonic regions	77.60%	70.70%	67.40%	66.90%	67.90%	67.60%
Reads mapped to transcriptome	68.30%	63.90%	55.80%	56.50%	57.10%	56.70%
Reads mapped antisense to gene	5.20%	3.60%	7.90%	6.90%	7.10%	7.20%



Supplementary figure 2: Visualization of QC metrics after initial pre-processing of snRNA-seq datasets (before doublet removal). QC plots showing total number of nuclei per technical replicate **(A)**, UMI counts per nuclei **(B)**, distribution of genes detected per nuclei **(C)**, correlation between genes detected and number of UMIs **(D)**, distribution of mitochondrial gene expression detected per nuclei **(E)** and genes detected per UMI **(F)** log10 genes per UMI.



Supplementary figure 3: PCA plot of PC1 and PC2 for the merged sample. PCA was performed on the scaled data, and the top 2 PCs were plotted and coloured by different cell cycle phases. G1 is represented in red, G2M represented in green, and S-phase is represented in blue.



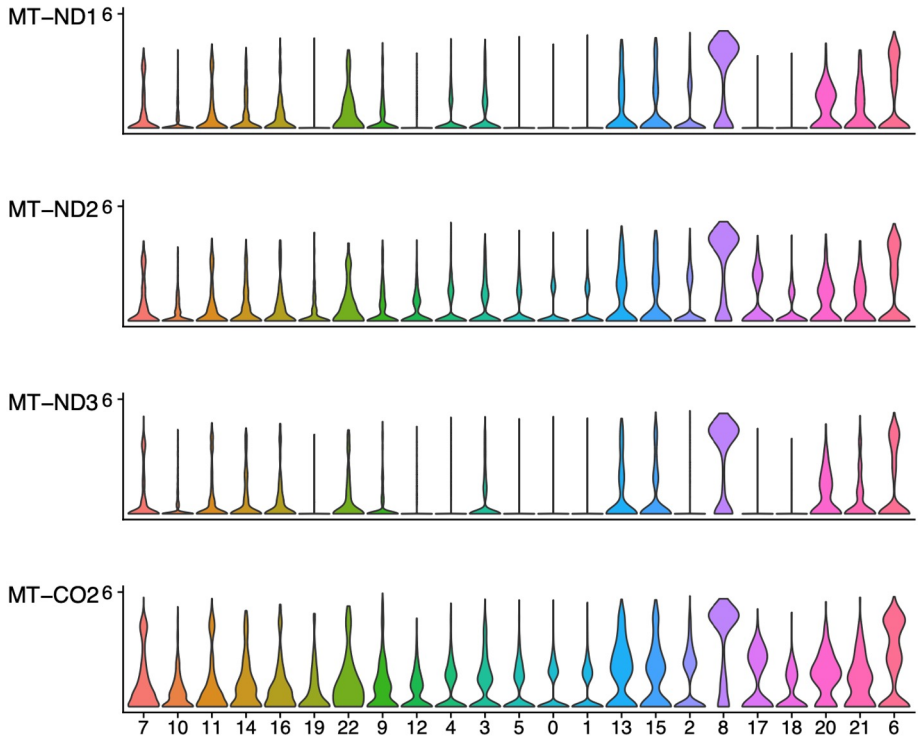
Supplementary figure 4: Elbow plot showing standard deviations for each PC.

Supplementary table 2: Total number of nuclei split by cluster and replicate for each sample. The percentage of nuclei per cluster derived from each replicate is also shown.

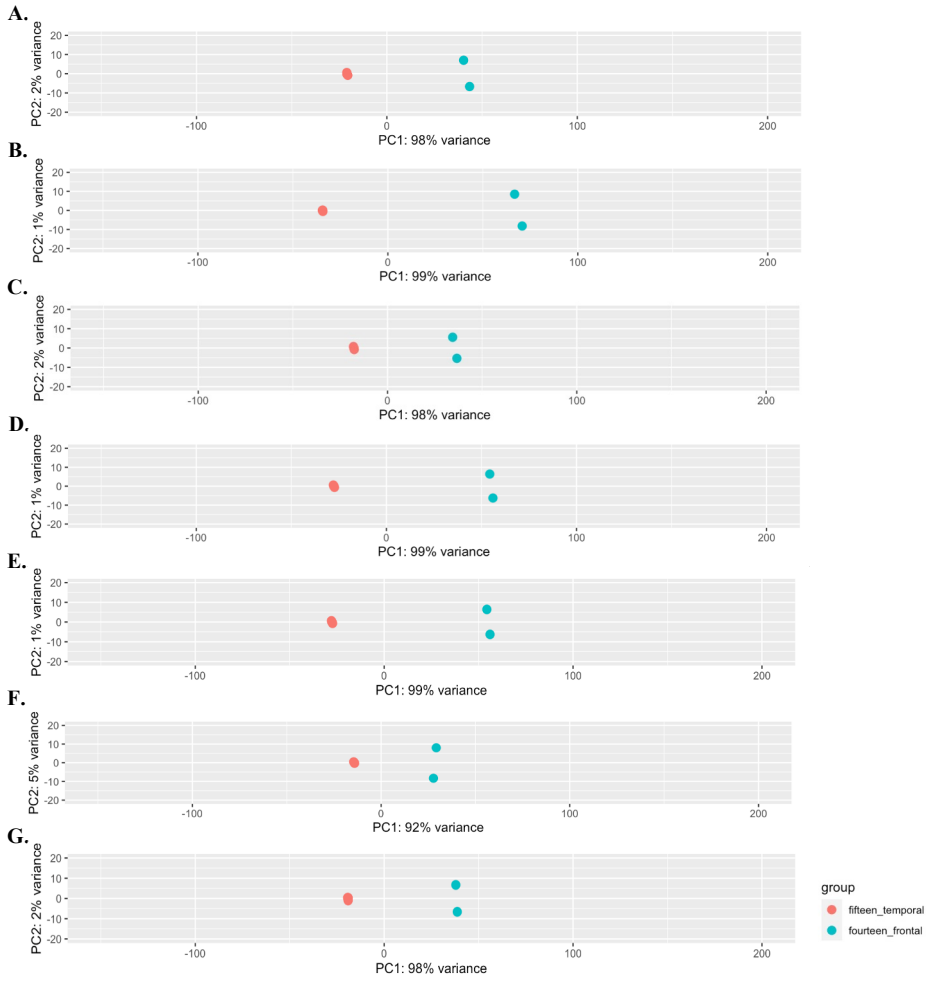
	14-T1	14-T2	15-T1	15-T2	15-T3	15-T4
0	206 (2.24 %)	179 (1.95 %)	2336 (25.41 %)	2253 (24.51 %)	1941 (21.11 %)	2279 (24.79 %)
1	131 (2.2 %)	150 (2.52 %)	1441 (24.21 %)	1445 (24.28 %)	1376 (23.12 %)	1409 (23.67 %)
2	211 (4.92 %)	317 (7.39 %)	926 (21.59 %)	932 (21.73 %)	947 (22.08 %)	956 (22.29 %)
3	373 (10.83 %)	496 (14.41 %)	607 (17.63 %)	625 (18.15 %)	680 (19.75 %)	662 (19.23 %)
4	176 (5.54 %)	198 (6.23 %)	705 (22.19 %)	671 (21.12 %)	721 (22.69 %)	706 (22.22 %)
5	91 (3.36 %)	96 (3.55 %)	627 (23.15 %)	644 (23.78 %)	667 (24.63 %)	583 (21.53 %)
6	24 (1.07 %)	695 (30.86 %)	308 (13.68 %)	302 (13.41 %)	625 (27.75 %)	298 (13.23%)
7	257 (14.5 %)	410 (23.14 %)	245 (13.83 %)	285 (16.08 %)	299 (16.87 %)	276 (15.58 %)
8	6 (0.53 %)	750 (66.67 %)	84 (7.47 %)	97 (8.62 %)	95 (8.44 %)	93 (8.27 %)
9	128 (11.99 %)	139 (13.01 %)	207 (19.38 %)	214 (20.04 %)	189 (17.7 %)	191 (17.88 %)
10	70 (10.67 %)	48 (7.32 %)	124 (18.9 %)	152 (23.17 %)	154 (23.48 %)	108 (16.46 %)
11	86 (13.56 %)	135 (21.29 %)	77 (12.15 %)	119 (18.77 %)	110 (17.35 %)	107 (16.88 %)
12	31 (4.89 %)	31 (4.89 %)	111 (17.51 %)	155 (24.45 %)	156 (24.61 %)	150 (23.66 %)
13	28 (4.7 %)	73 (12.25 %)	140 (23.49 %)	135 (22.65 %)	109 (18.29 %)	111 (18.62 %)
14	61 (11.32 %)	107 (19.85 %)	85 (15.77 %)	94 (17.44 %)	93 (17.25 %)	99 (18.37 %)
15	10 (2.7 %)	48 (12.97 %)	78 (21.08 %)	82 (22.16 %)	77 (20.81 %)	75 (20.27 %)
16	66 (23.57 %)	45 (16.07 %)	50 (17.86 %)	43 (15.36 %)	41 (14.64 %)	35 (12.5 %)
17	5 (1.81 %)	4 (1.44 %)	56 (20.22 %)	84 (30.32 %)	62 (22.38 %)	66 (23.83 %)
18	0 (0 %)	3 (1.21 %)	53 (21.46 %)	63 (25.51 %)	50 (20.24 %)	78 (31.58 %)
19	15 (7.43 %)	8 (3.96 %)	48 (23.76 %)	47 (23.27 %)	43 (21.29 %)	41 (20.3 %)
20	84 (23.57 %)	20 (23.57 %)	16 (23.57 %)	10 (23.57 %)	23 (23.57 %)	17 (23.57 %)
21	9 (7.5 %)	17 (14.17 %)	21 (17.5 %)	16 (13.33 %)	33 (27.5 %)	24 (20 %)
22	21 (19.27 %)	19 (17.43 %)	17 (15.6 %)	24 (22.02 %)	14 (12.84 %)	14 (12.84 %)

Supplementary table 3: SCSA automatic annotation results showing the assigned cell type for each cluster.

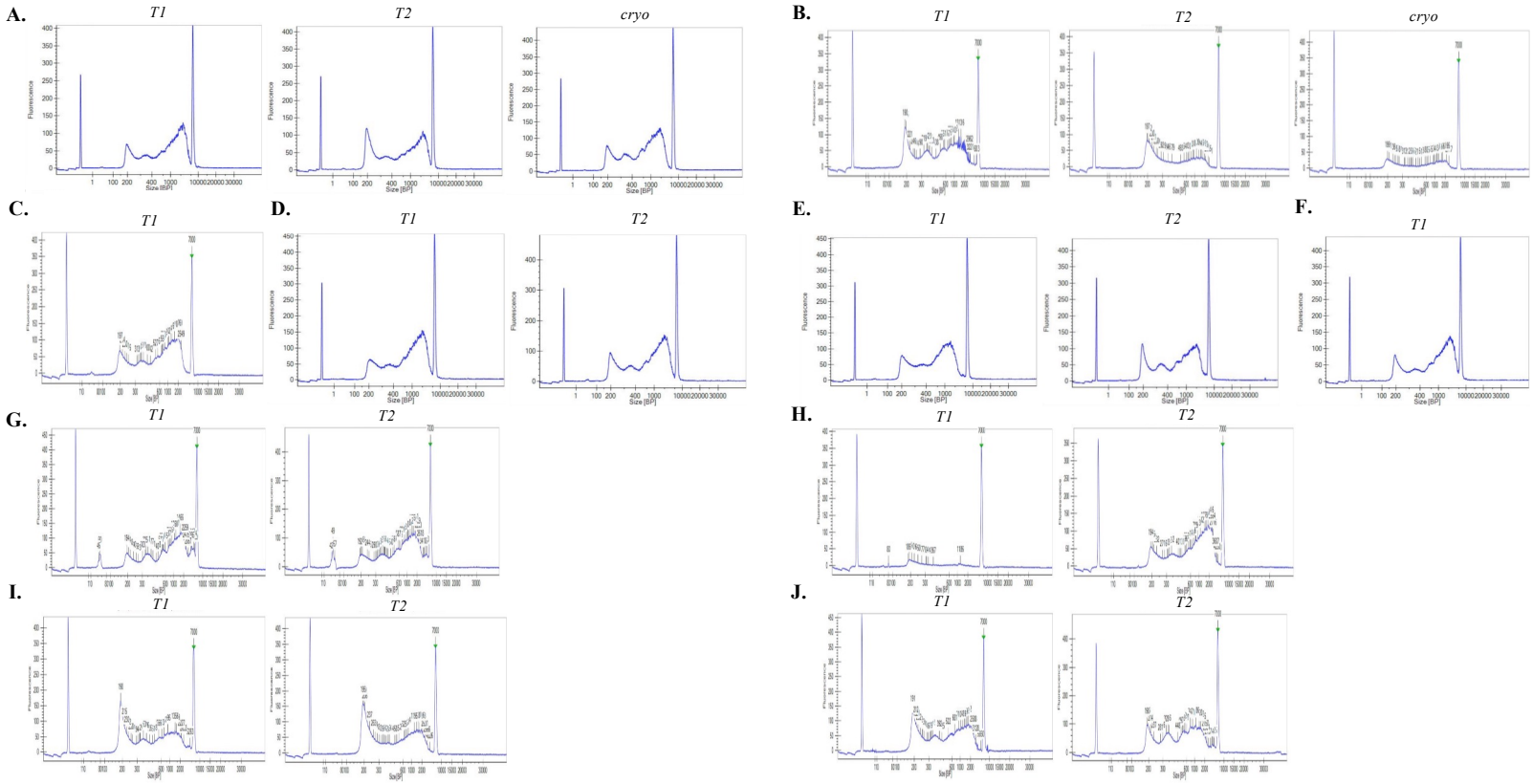
Cluster	Potential Cell type
0	Oligodendrocyte
1	Oligodendrocyte
2	Excitatory/Astrocyte
3	Astrocyte
4	Oligodendrocyte/OPC
5	Astrocyte
6	Neuron
7	Neuron
8	Unknown
9	Excitatory
10	Neuron
11	Neuron
12	Excitatory
13	Endothelial
14	Excitatory
15	Astrocyte
16	Neuron
17	Unknown
18	Astrocyte
19	Neuron
20	Microglia
21	Astrocyte
22	Excitatory



Supplementary figure 5: Gene expression data allowed identification of cell debris clusters in the integrated dataset. Violin plots showing the expression of mitochondrial genes, *MT-ND1*, *MT-ND2, MT-ND3,* and *MT-CO2* for each cluster. The cluster identity is shown on the x-axis and the expression level of each gene is shown on the y-axis.



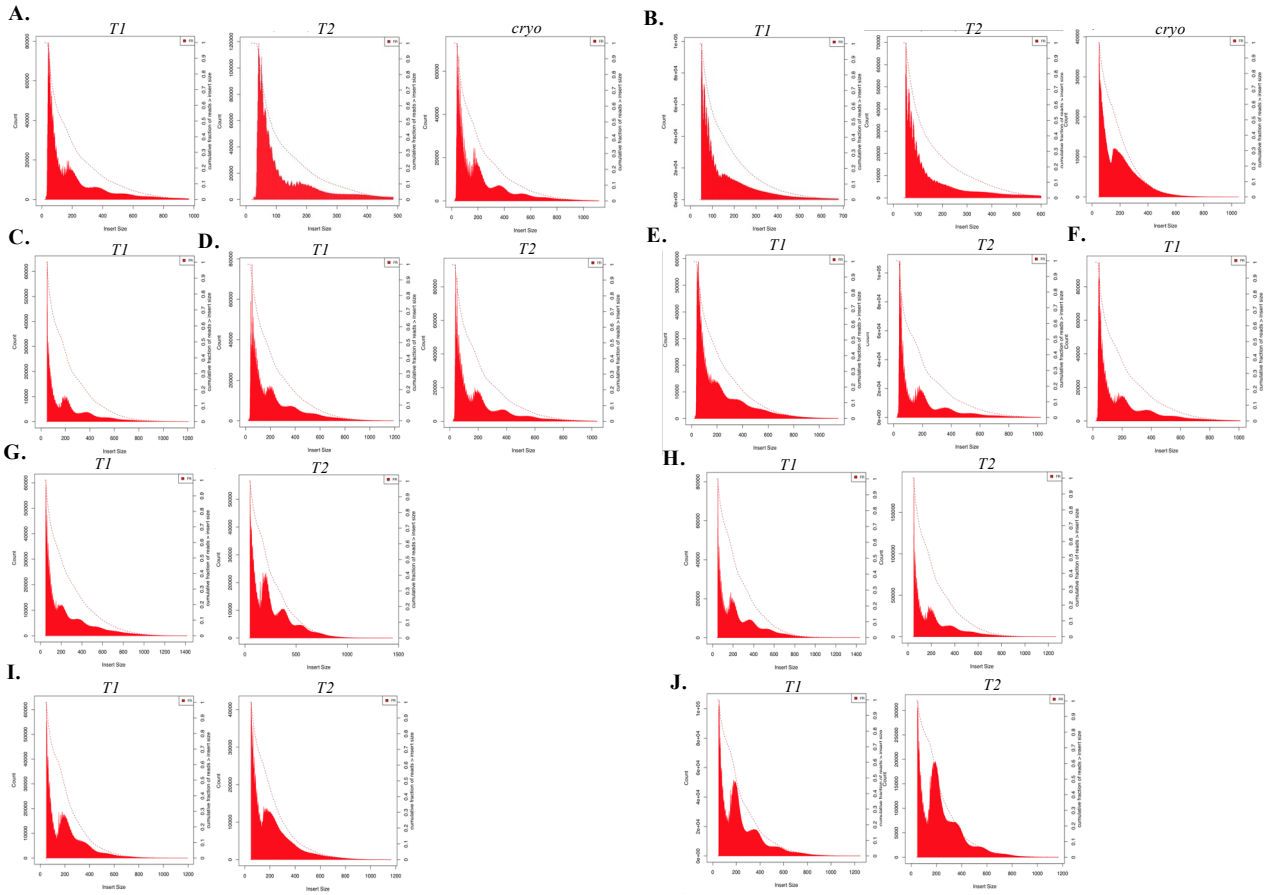
Supplementary figure 6: PCA plots for each cell type (A-G). (A) inhibitory neurons, **(B)** excitatory neurons, **(C)** OPCs, **(D)** astrocytes, **(E)** oligodendrocytes, **(F)** endothelial cells and **(G)** microglia.



Supplementary figure 7: *Quality of DNA libraries after performing PCR amplification (A-H).* ATAC-seq libraries were amplified and then the sample was run on the Agilent High Sensitivity Bioanalyzer to obtain the fragment size distribution and check the quality of the ATAC-seq libraries. **(A)** 19-month-old, **(B)** 23-month-old, **(C)** 5-year-old, **(D)** 6-year-old, **(E)** 9-year-old, **(F)** 14-year-old, **(G)** 15-year-old, **(H)** 23-year-old, **(I)** 31-year-old, and **(J)** 46-year-old. Cryo, cryopreserved sample; T1, technical replicate 1; T2, technical replicate 2. The X-axis shows the fragment size (base pairs), and the Y-axis shows the fluorescence units.

Supplementary table 4: *Bowtie 2* mapping results. Fastq files containing the reads were mapped to the human reference genome and BAM files were generated for each sample using *Bowtie 2*. The total number of reads and alignment rate for each sample is shown.

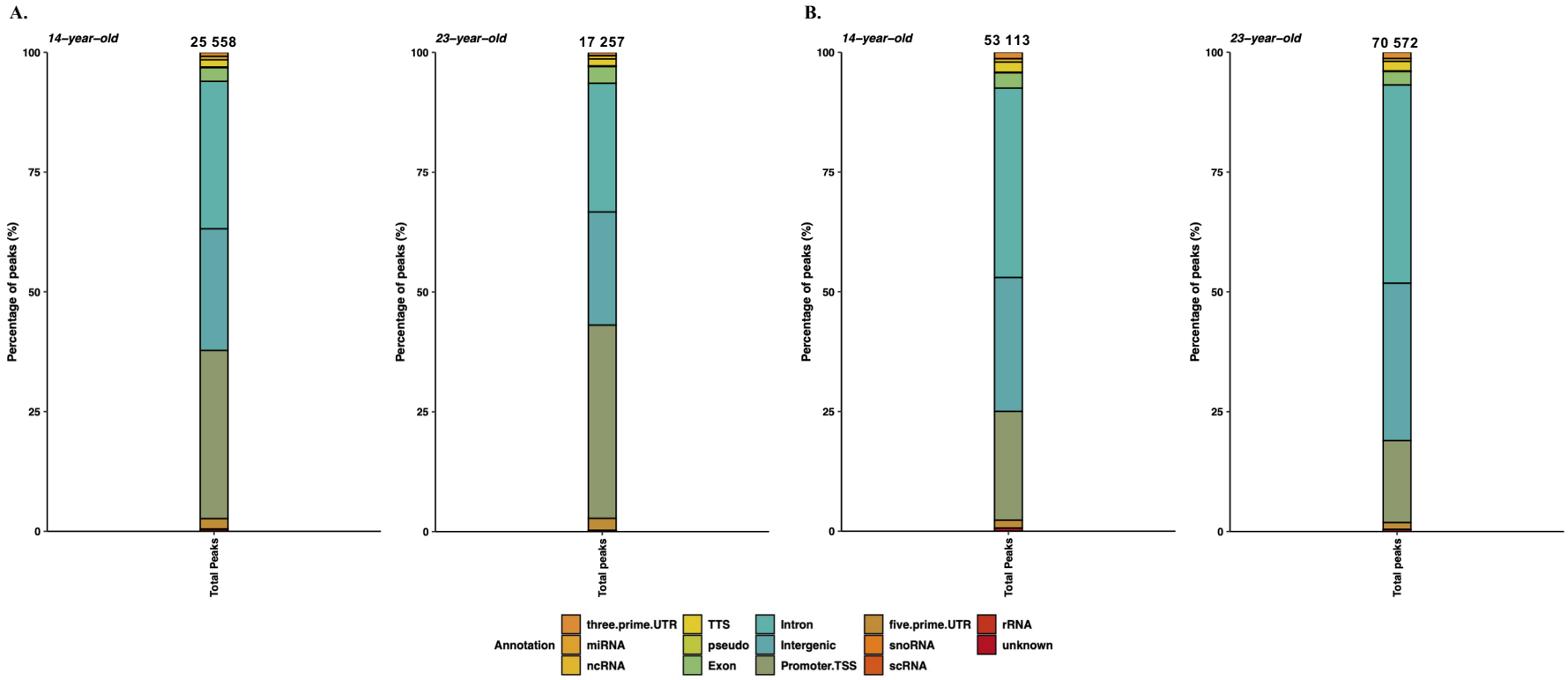
Sample	Number of reads	Alignment rate (%)
19-month-old fresh T1	23257372	97.31
19-month-old fresh T2	20619534	96.57
19-month-old cryo	25654011	92.84
9-year-old fresh T1	26722556	79.5
9-year-old fresh T2	20639308	91.87
6-year-old fresh T1	22170449	86.62
6-year-old fresh T2	20324490	76.09
14-year-old frozen	21253951	87.91
15-year-old fresh	17940036	88.18
15-year-old cryo	13052031	91.4
31-year-old fresh T1	23890004	68.3
31-year-old fresh T2	35683075	68.3
46-year-old fresh T1	13743548	70.83
46-year-old fresh T2	31525216	76.93
23-month-old fresh T1	33265509	68.25
23-month-old fresh T2	34788293	77.66
23-month-old cryo	14848753	85.89
5-year-old fresh T1	24690178	74.95
5-year-old fresh T2	14351984	85.61
23-year-old fresh	26695892	80.75



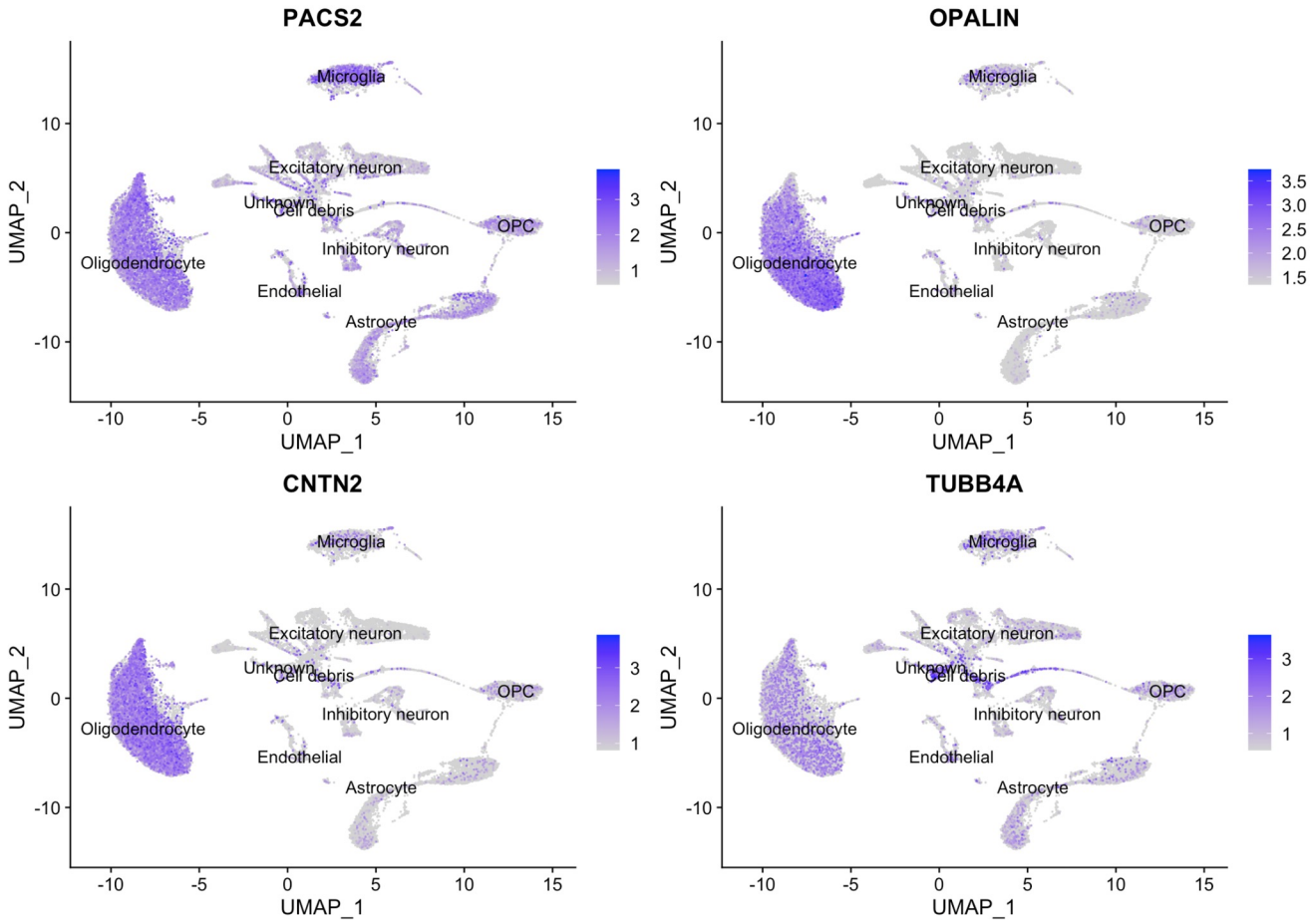
Supplementary figure 8: *Sequence insert size distribution plots after performing read alignment using Bowtie2.* After sequencing, the reads were aligned to the human reference genome (Grh38) using Bowtie2 (See Methods). Then, the distribution of sequence insert sizes was assessed using Picard tools to check the quality of the aligned data (A-J). **(A)** 19-month-old, **(B)** 23-month-old **(C)** 5-year-old, **(D)** 6-year-old, **(E)** 9-year-old,**(F)** 14-year-ol,d **(G)** 15-year-old, **(H)** 23-year-old**, (I)** 31-year-old, and **(J)** 46-year-old. T1. T1, technical replicate 1, T2, technical replicate 2 and cryo, cryopreserved sample. The X-axis shows the insert size (BP), and the Y-axis shows the fluorescence.

Supplementary table 5: *MACS2* and *Genrich* peak calling results. The total number of peaks for each sample called using the *MACS2* or *Genrich* peak-caller.

Sample	Total no. of peaks (MACS2)	Total no. of peaks (Genrich)
19-month-old fresh T1	12 346	27 482
19-month-old fresh T2	19 549	30 266
19-month-old cryo	21 204	42 174
23-month-old fresh T1	6 776	26 294
23-month-old T2 old fresh	9 764	19 166
23-month-old cryo	6 212	34 471
5-year-old fresh T1	17 664	20 648
5-year-old fresh T2	51 855	45 588
6-year-old fresh T1	41 238	76 883
6-year-old fresh T2	44 292	65 779
9-year-old fresh T1	21 631	45 184
9-year-old fresh T2	18 745	45 141
14-year-old frozen	25 558	53 113
15-year-old fresh	29 483	62 104
15-year-old cryo	44 478	79 012
23-year-old fresh	17 257	70 572
31-year-old fresh T1	29 483	86 116
31-year-old fresh T2	44 478	76 268
46-year-old fresh T1	20 700	68 594
46-year-old fresh T1	4 688	31 658



Supplementary figure 9: ***Peak annotations after performing peak calling using MACS2 and Genrich peak callers (A-B).*** **(A)** shows the total peaks called using *MACS2* and **(B)** shows the total peaks called using *Genrich* for the indicated samples.



Supplementary figure 10: Feature plots generated in Seurat showing the expression of important genes identified in the ATAC-seq analysis. (A) shows the expression of *PACS2*, **(B)** *OPALIN*, **(C)** *CNTN2* and **(D)** *TUBB4A*. Nuclei are colored purple if the markers are highly expressed and grey if there is no expression.