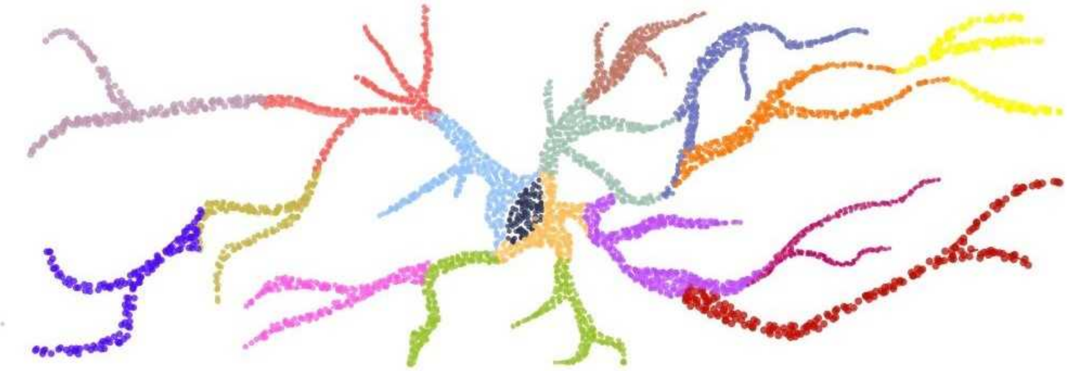


Profiling ageing-associated gene expression alterations in human astrocytes using single cell transcriptomics

Rita Martins Tereso Borges da Silva

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Susana de Almeida Mendes Vinga Martins
Dr. Nuno Luís Barbosa Morais

Examination Committee

Chairperson: Prof. Cláudia Alexandra Martins Lobato da Silva
Supervisor: Dr. Nuno Luís Barbosa Morais
Members of the Committee: Dr. Luísa Vaqueiro Lopes
Prof. Ana Luísa Nobre Fred

November 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was performed at the Disease Transcriptomics (NMorais) Lab of Instituto de Medicina Molecular João Lobo Antunes, Faculty of Medicine of the University of Lisbon (Lisboa, Portugal), during the period February-November 2021, under the supervision of Dr. Nuno Morais. The thesis was co-supervised at Instituto Superior Técnico by Professor Susana Vinga.

This work was supported by the GenomePT project [(POCI-01-0145-FEDER-022184), supported by COMPETE 2020 – Operational Programme for Competitiveness and Internationalization (POCI), Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Programme (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF)] and the European Molecular Biology Organization [EMBO Installation Grant 3057 to N.B.-M.].

Acknowledgments

First of all, I would like to sincerely thank Dr. Nuno Morais, for giving me the opportunity to join the disease transcriptomics lab at iMM and for guiding me, as a mentor, in this entire journey that culminates with my master's degree in biomedical engineering. His enthusiasm for science, genuine interest and scientific rigor motivates me to want to be a better scientist, reinforcing my taste for academia and the field of transcriptomics. I would also like to thank Professor Susana Vinga, my IST supervisor, for all her bureaucratic assistance and availability, and for teaching bioinformatics / computational biology in such an exciting way and with such diverse themes and invited speakers.

I am also deeply grateful to all my colleagues at the NMorais lab at iMM: José Ferrão, Mariana Ferreira, Marta Bica, Miguel Casanova, Nuno Agostinho and Rita Belo. Thank you for your mentorship, availability, and good fellowship that, even in conditions where presence in person was not possible, made me feel very welcome. Thank you for showing me what life as an investigator is like, and for making me believe more in myself and my abilities.

I could not fail to thank my fellow classmates of Biomedical Engineering at IST, for motivating me to always do my best, for helping me and always being by my side during these 5 years. To my *Unstable Unicorns*, *Ricous* and *Enhocados*, I sincerely thank you for always being there for me, especially in these pandemic times. You are the best friends anyone could ask for. In particular, I want to express my deepest gratitude to Vicente Garção, for being a source of motivation and self-confidence, and for believing in me when I didn't believe in myself. I could not have done this without you.

Finally, I would also like to leave a huge thank you to all my family, especially my parents and my sister, for always believing in me and providing me with help and a safe space whenever my academic responsibilities started to feel too much.

Without all these people my academic path would not have been the same. A big thank you to everyone.

Resumo

Sabe-se que o cérebro humano envelhecido cai num fenótipo inflamatório, referido como “inflammageing”, com os astrócitos a serem das células mais afetadas. Em condições fisiológicas, estas células são responsáveis por manter a homeostase do Sistema Nervoso Central e, entre outras funções, são capazes de modular a atividade neuronal, providenciar fatores tróficos e nutrientes aos neurónios, e estabelecer e manter a barreira hematoencefálica. No entanto, com o envelhecimento, os astrócitos sofrem mudanças de expressão génica, perdendo parte das suas funções normais. É plausível admitir que mudanças no fenótipo dos astrócitos possam deixar o cérebro mais vulnerável a lesões e a doenças relacionadas com a idade (envelhecimento patológico). No entanto, as tecnologias de análise transcricional atualmente aplicadas a tecido cerebral humano (e.g. RNA-seq) falham em distinguir estados astrocíticos mais subtis.

Dito isto, o principal objetivo deste trabalho é caracterizar as assinaturas de expressão génica de astrócitos humanos em envelhecimento fisiológico utilizando dados públicos de sequenciação de transcritomas de células individuais (scRNA-seq). Os resultados deste trabalho demonstram um claro aumento da heterogeneidade de astrócitos com a idade, incluindo um grupo de astrócitos que aparenta estar enriquecido em características de envelhecimento, tais como neuroinflamação, excitotoxicidade, perda de funções de suporte neuronal e de homeostase sináptica, e cujo enriquecimento com a idade foi validado em dados independentes. Este trabalho contribui cientificamente com a descoberta de alvos moleculares para validação do fenótipo *in vitro* e *in vivo*, bem como potenciais compostos terapêuticos capazes de reverter o fenótipo daqueles astrócitos associados a envelhecimento patológico.

Palavras-chave: Astrócitos, Envelhecimento, Doenças Neurodegenerativas, Single-cell RNA sequencing

Abstract

It is known that the ageing brain falls into an inflammatory phenotype, referred to as “inflammageing”, with astrocytes being within the most affected cells. In physiological conditions, these cells are responsible for maintaining the central nervous system homeostasis and, among others, can modulate neuronal activity, provide trophic factors and nutrients to neurons, and establish and maintain the blood brain barrier. However, with ageing, astrocytes undergo gene expression changes, losing part of their normal functions. It is plausible that changes in astrocyte’s phenotype can make the brain more vulnerable to injury and age-related diseases (pathological ageing). However, transcriptomic profiling technologies currently applied to human brain tissue (such as RNA-seq) fail to discriminate more subtle astrocyte activity states.

As such, the main objective of this work is to characterize the gene expression signatures of human astrocytes in physiological ageing using publicly available single-cell transcriptomic (scRNA-seq) data. Such results demonstrate a clear increase on astrocytic heterogeneity with age, including a group of astrocytes enriched in ageing hallmarks, such as neuroinflammation, excitotoxicity, loss of neuronal support and synaptic homeostasis functions, and whose enrichment with age has been further validated in independent datasets. This work’s main scientific contribution was the discovery of molecular targets for phenotype validation *in vitro* and *in vivo*, as well as candidate therapeutic compounds for the reversal of astrocytic phenotypes associated with pathological ageing.

Keywords: Astrocytes, Ageing, Neurodegenerative Diseases, Single-cell RNA sequencing

Contents

List of Tables

List of Figures

List of Acronyms

AD	Alzheimer’s Disease
ALS	Amyotrophic Lateral Sclerosis
BBB	Blood-Brain Barrier
CNS	Central Nervous System
DEA	Differential Expression Analysis
ER	Endoplasmic Reticulum
GSEA	Gene Set Enrichment Analysis
PCA	Principal Component Analysis
PC	Principal Component
PD	Parkinson’s Disease
RNA-seq	RNA sequencing
RNA	Ribonucleic Acid
ROS	Reactive Oxygen Species
scRNA-seq	single-cell RNA sequencing
scRNA-seq	single-nucleus RNA sequencing
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMI	Unique Molecular Identifier
UPR	Unfolded protein response

Chapter 1

Introduction

1.1 Motivation

Ageing consists of a multitude of genetic, biological and environmental factors that compromise the function of several cells in our body [1]. It is the strongest risk factor for numerous neurodegenerative disorders, such as Alzheimer's or Parkinson's diseases [2–4]. As the average life expectancy and median age of the world population increase, there is an urge to uncover the age-associated mechanisms of pathology, in order to adopt therapies that pair this increase in longevity with an increase in health span and quality of life [5].

It is known that the ageing human brain presents an inflammatory phenotype, referred to as "inflammageing" [6], with astrocytes being within the most affected cells. In healthy conditions, these cells are responsible for maintaining the central nervous system (CNS) homeostasis and, among others, can modulate neuronal activity, provide trophic factors and nutrients to neurons, and establish and maintain the blood brain barrier [1, 7, 8]. Given the critical role of astrocytes in the proper functioning of the CNS, it is plausible that age-associated changes in their phenotype can leave the brain more vulnerable to injury and age-related diseases. It has been previously demonstrated that, with ageing, astrocytes undergo gene expression changes, losing part of their functions and acquiring a pro-inflammatory phenotype [9], which may lead to an exacerbation of the low-grade inflammation state concomitant with ageing [1, 9].

In the past few years, there have been some exploratory assays focusing on ageing in human astrocytes. With transcriptomic analysis techniques, such as RNA sequencing (RNA-seq), several ageing-associated changes in gene expression and signalling pathways have already been identified. However, this technology lacks the resolution to elucidate on the real heterogeneity of brain tissue, given that it pools all cell types together and measures their average activity, not distinguishing cell groups with different functions. In 2009 the first test was carried out with a new technique, single cell RNA-seq (scRNA-seq), which permitted profiling the transcriptome of each cell individually [10]. In recent years, the application of this technique to human tissues, namely the brain, has become quite common. It appears as a way to solve the lack of cellular resolution of bulk RNA-seq, allowing to look at the distribution of gene expression levels across a population of individual cells, and hereby to acquire insights into novel candidate genes, signalling pathways and biological processes characteristic of ageing tissues.

Furthermore, there are numerous therapies that momentarily improve the symptoms of patients with several neurodegenerative diseases but whose maximum effectiveness is observed in animal models or cell cultures (where they were idealized), suggesting that these systems are not perfect surrogates for modelling age-related illnesses [11].

With the increase in average life expectancy, it is urgent to more deeply understand the role of astrocytes in **normal ageing** of the human brain with single cell resolution, in order to better grasp how the dynamics of the ageing human brain leads to predisposition to neurodegenerative diseases.

1.2 Objectives and Methodology

The main objective of this work was to characterise gene expression alterations in human astrocytes in physiological ageing. Such study is lacking and is needed to discover the transcriptional individuality of human aged astrocytes that can contribute to the predisposition of the aged human brain for neurodegenerative diseases.

For this, publicly available scRNA-seq data were used to obtain gene expression signatures of human ageing astrocytes. These signatures were used **(1)** to estimate the relative abundance of the distinct functional types of astrocytes in post-mortem brain tissues based on their transcriptomes (bulk RNA-seq); **(2)** to identify relevant associated pathways and cellular processes for future *in vitro* or *in vivo* validation studies; and **(3)** as molecular targets for the *in silico* identification of candidate drugs capable of phenotype reversal. Several analysis and visualisation packages, implemented in statistical software `R`, were used in obtaining those signatures.

1.3 Thesis Outline

This document is divided into five main parts. Chapter 1 is the present chapter, where the motivation and objectives of this work are summarised. Chapter 2, or Background, consists of the basic concepts for understanding the problem under study, through a careful and systematic literature review. This chapter explores the basics of physiological and pathological ageing, central nervous system cells, and ageing astrocytes as a potential aetiology of the increased predisposition of the elderly to neurodegenerative diseases. Chapter 3, or Materials and Methods, provides an overview of the public datasets used in this work, as well as the main scRNA-seq data processing and visualisation packages. By personal choice, this chapter also includes the results of scRNA-seq pre-processing, so that Chapter 4, or Results and Discussion, focuses only on the downstream data exploration and on addressing the proposed biological problem, including *in silico* validation. Finally, Chapter 5, or Concluding Remarks, consists of a synthesis of the results of this work, as well as considerations on its main limitations and future work to be developed in this topic.

Chapter 2

Background

2.1 The Ageing Human Brain

The Central Nervous System (CNS) is possibly the most complex system in the human body, being divided into two main parts: the spinal cord and the brain. The brain can be further divided into six main parts, namely the medulla oblongata, pons, cerebellum, midbrain, diencephalon (comprising the thalamus and hypothalamus), and cerebrum (comprising the cortex, basal ganglia, hippocampus and amygdaloid nuclei) [12].

The brain is responsible for numerous functions, controlling all the organs of the human body by processing, integrating, and coordinating multiple information coming from and to them [12]. However, given its importance in the control and homeostasis of the entire human body, any malfunction can have disastrous consequences, with age being a crucial factor of predisposition to illness and injury [13].

2.1.1 The Cell Types of the Human Brain

The concept of cell type is an interesting matter, since each cell is unique and behaves differently from the others, depending on the resolution at which we choose to analyse it [14]. However, many cells share similar activities in tissues, so it is common to group cells according to their morphology and perceived function. In addition, a cell can have several classifications: take as an example the so-called "canonical" cell type, such as muscle or nerve cells, defined according to the overall function that the tissue in which they are found presents. Though, within muscle cells, it is possible to further increase the resolution of such classification, finding skeletal, smooth, or cardiac muscle cells, among others [15]. It all depends on the resolution we choose to look at the cells and what we hope to achieve with this classification – a critical point intended to be explored in this work.

The broadest cell type classification in the brain assumes two main groups (figure 2.1): **neurons** and **glial cells** [16].

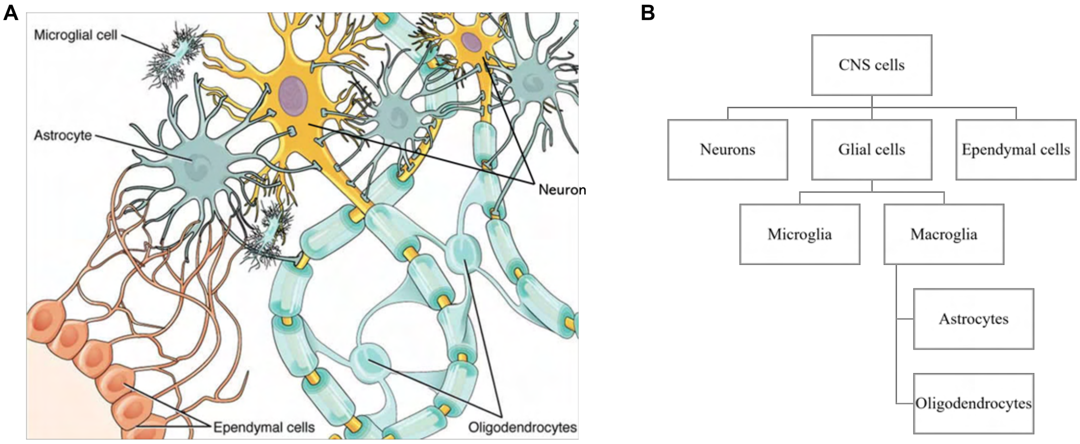


FIGURE 2.1: Cells in the human brain
(A) Cells that make up the human brain, including neurons, glial cells (astrocytes, microglia and oligodendrocytes) and ependymal cells. Adapted from NeuroscienceNews: New Cause of Schizophrenia Uncovered (2017) [17]. **(B)** Schematic representation of the CNS cell type classification, according to Kandel and Shadlen (2021) [16].

Neurons

Neurons are cells capable of being excited and transmitting information by conducting electrical stimuli. Morphologically, these cells are constituted by a cell body, dendrites, axons, and axon terminals. Furthermore, neurons are extremely heterogeneous, and can be classified, among others, according to their morphology (unipolar, bipolar, multipolar) or function (sensory, motor) [16]. It is estimated that the human brain has around 86 billion neurons, this density being possible due to our great efficiency in food ingestion and energy consumption [18].

Glial Cells

Glial cells are non-neuronal cells that do not produce electrical impulses and are not capable of being electrically excited, but are essential to proper functioning of neurons, being responsible for, among other functions, protecting and nourishing them [16, 19]. Glia comes from the Greek for ”glue” and was initially thought to be responsible for holding neurons in place and act as supportive cells [19]. Currently, these cells are known to be undoubtedly more complex than that: they surround the cell bodies, axons, and dendrites of neurons, maintaining the homeostasis of the entire system [16, 19].

In the human brain, as in all vertebrate brains, we can further divide glial cells into two large groups: macroglia and microglia. **Microglia** are immune cells responsible for presenting antigens and acquiring a phagocytic phenotype (i.e. becoming specialized macrophages) in response to injury and infection. Within macroglia we can also define oligodendrocytes and astrocytes. **Oligodendrocytes** are cells that cover the axons of neurons with their own cell membrane, forming the myelin that is responsible for increasing the speed of neuronal transmission of the electrical message, restricting the action of voltage-sensitive ion channels only in specific regions (Ranvier’s nodes). Finally, **astrocytes** are highly heterogeneous cells responsible for maintaining synapse homeostasis and forming the blood brain barrier, among others, but whose function is not yet fully understood [16].

We may consider another type of cells in the central nervous system that do not fall in the previous classification: **ependymal (endothelial) cells** cover the ventricular system of the brain, and they are responsible for the creation and secretion of cerebrospinal fluid [16].

The relative abundance of cell types varies between brain regions. It is estimated that oligodendrocytes are the most abundant glial cell type (45-75% of total human brain glial cells), followed by astrocytes (19-40%), and microglia (10% or less) [20]. For several decades, more specifically from the 1960s until 2009, it was thought that the ratio between neurons and glial cells was around 1:10 (100 billion neurons to one trillion glial cells) [21]. However, with the emergence of a novel way of quantifying the abundance of cell types in the human brain (i.e., *isotropic fractionator* [22]), it has now been established that this ratio is highly dependent on the brain area and age, and generally less than 1:1, meaning that there are more neurons than glial cells [18, 21].

2.1.2 Physiological Ageing

What is Ageing?

Although there is no unique definition for ageing, several studies have discussed this topic [23]. The broad scientific consensus points to ageing as a set of genetic, biological, and environmental factors, which act together and lead to physiological and cognitive changes, compromising cells in their functions [1]. Therefore, ageing can be seen as a deterioration of the physiological integrity of an organism caused by the passage of time [24].

The mechanisms and causes of ageing have always aroused curiosity in humans. However, the “new era” of exploring ageing had its origins 40 years ago, as the cellular and molecular mechanisms of life and disease began to be explored [13].

Ageing affects every part of the body differently, with some showing obvious changes, such as skin atrophy [8], and others showing more subtle alterations. Tissues composed primarily of postmitotic cells, such as the brain, are especially prone to the nefarious effects of ageing [25].

The Hallmarks of Ageing

The effect of ageing in the brain is noticeable to us mainly through cognitive decline, which can manifest itself in different stages of severity [26]. There is ample evidence that ageing begins by affecting episodic memory, involving consciously remembering events and experiences, and executive functions, a set of capacities involved in planning, mental flexibility, inhibiting inappropriate actions, attending to relevant and ignoring irrelevant sensory information [27]. However, irrespectively of reaching such a perceptible consequence, the ageing brain, such as the remaining human organs and tissues, shows some consistent changes amongst the elderly population, the so called “hallmarks of ageing” [13].

According to López-Otín and colleagues (2013), there are **nine fundamental hallmarks of ageing** (Figure 2.2), with these being divided into three main categories: primary, antagonistic, and integrative [13, 25]. Furthermore, they all follow the following criteria: they manifest themselves during physiological ageing, their aggravation accelerates ageing, and their amelioration retards the physiological ageing process, being thus very interesting candidates for ageing reversing therapeutics [13].

The **primary hallmarks** are the first to occur and negatively affect the human body. These include genomic instability, telomere attrition, epigenetic alterations, and loss of proteostasis, which is the dynamic balance of the formation and maturation of functional proteins. Genomic instability is at the heart of the main theories of ageing [13]. It is described as the dysregulation of mechanisms that control and correct DNA mutations or other changes during cell division, which may lead to non-functional cells.

Antagonist hallmarks arise in response to primary hallmarks and tend to counteract their effects through compensatory mechanisms, including mitochondrial dysfunction, cellular senescence, and deregulation of nutrient sensing. Although these actions start by being beneficial, as they compensate for the damage caused by the primary hallmarks, their prolonged action may also be detrimental. Mitochondrial dysfunction is particularly harmful, since although the production of reactive oxygen species may be beneficial, as it is responsible for cell signalling and survival, its prolonged effect is associated with toxicity and cell death [13].

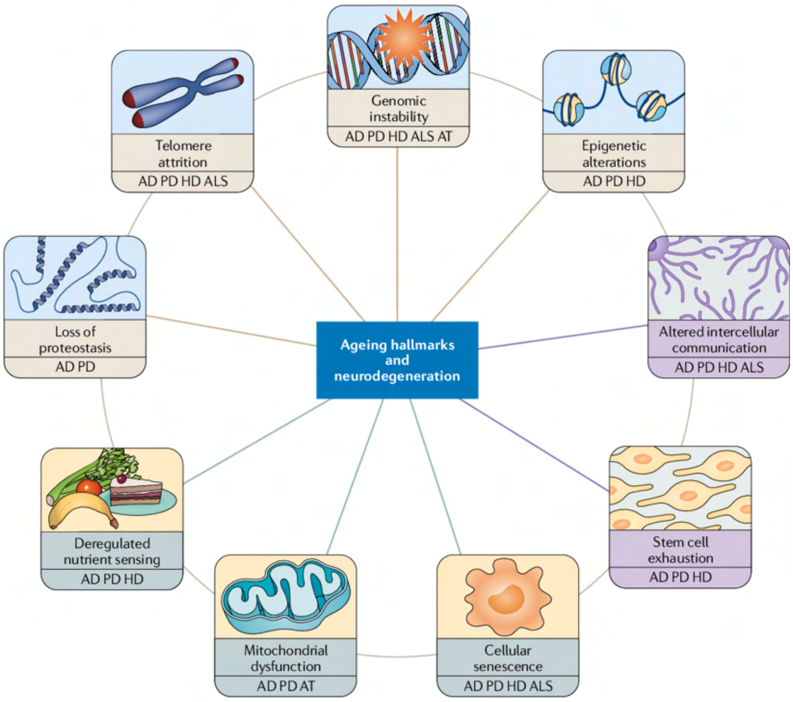


FIGURE 2.2: The nine hallmarks of ageing
The nine hallmarks of ageing and the neurodegenerative diseases associated with them (AD – Alzheimer's disease; PD – Parkinson's disease; HD – Huntington's disease; ALS – amyotrophic lateral sclerosis; AT – ataxia telangiectasia). Primary hallmarks are identified in beige, antagonist hallmarks in grey and integrative hallmarks in purple. Retrieved from Hou *et al.*, 2019 [25].

Finally, **integrative hallmarks** arise as a consequence of the joint damage of the remaining hallmarks and when the cellular homeostasis mechanisms become unable to maintain their proper functioning. These hallmarks include stem cell exhaustion and altered intercellular communication (such as calcium signalling), and are thought to be responsible for the effective functional decline associated with age [13].

Taking all of the above into consideration, it becomes clear that the proposed nine hallmarks of ageing have a hierarchical relationship between them [13]. Thus, elucidating on the exact causal relationships between them through their individual study for each organ and cell is critical for a better understanding of the mechanisms of ageing and age-related diseases.

Conclusively, one that is worth mentioning and exploring further is the **oxidative damage theory of ageing**, as well as its repercussions and influence on other detrimental mechanisms. It postulates that "age-associated reductions in physiologic functions are caused by a slow steady accumulation of oxidative damage to macromolecules which increases with age, and which is associated with life expectancy of organisms" [28]. A corollary of this theory also postulates that "the rate of ageing should be retarded by attenuation of oxidative damage". However, oxidative stress is only one of many types of "stress", which will be further detailed (figure 2.3). Cellular stress can be of different sorts, such as caused by radiation, chemotherapy, oncogene activation or hypoxia, as well as oxidative and endoplasmic reticulum stress [29]. The endoplasmic reticulum is responsible for, among other functions, preventing protein aggregates, by ensuring correct transcription and translation, as well as post maturation and folding of proteins. This is maintained through several mechanisms, such as the unfolded protein response (UPR). Endoplasmic stress can be caused by a multitude of factors, including several ageing hallmarks (nutrient depletion, disturbances in calcium signalling), and can result in the disruption of the UPR, disturbing tissue homeostasis. Moreover, the ER stress induces an inflammatory response that in chronic conditions can result in senescence, apoptosis, or triggers a compensatory immunosuppression mechanism, with the release of mediators of immunosuppression secreted by immune cells, such as TGF-β, IL-10 and ROS [29, 30]. There are several mechanisms to deal with cellular stress, namely cellular stress responses, stress-induced cell death, and senescence [29]. Cellular stress responses include the aforementioned UPR, heat shock response and DNA damage response. When these mechanisms fail to recover cell homeostasis, stress-induced cell death can happen in order to prevent further damage, including programmed cell death (apoptosis or autophagy) and necrosis [29–31].

The Hallmarks of the Ageing Brain

The effect of physiological ageing in the brain is noticeable mainly through cognitive decline. To reach this visible consequence, the brain, like other organs, manifests the aforementioned hallmarks of ageing. More specifically, the aged brain suffers from loss of dendritic spines [1], mitochondrial dysfunction, dysregulated energy metabolism, compromised DNA repair, stem cell exhaustion [2], loss of stem cells in the hippocampus [1], aberrant neural network activity [4] and inflammation [1, 2, 4, 27], that will be discussed next.

Dendritic spines are specialized protrusions of the neuronal dendritic surface, and are fundamental for excitatory transmission and synaptic plasticity in the brain [32]. The loss of dendritic spines, as well as the change in proportions of the various types of spines that occur naturally with age, can greatly impact normal cognitive functions [1, 33]. Furthermore, mitochondria are responsible for a plethora of functions, such as production of ATP, calcium signalling, lipid biosynthesis and cellular apoptosis, and when their function is compromised as a result of natural ageing, they can increase the production of

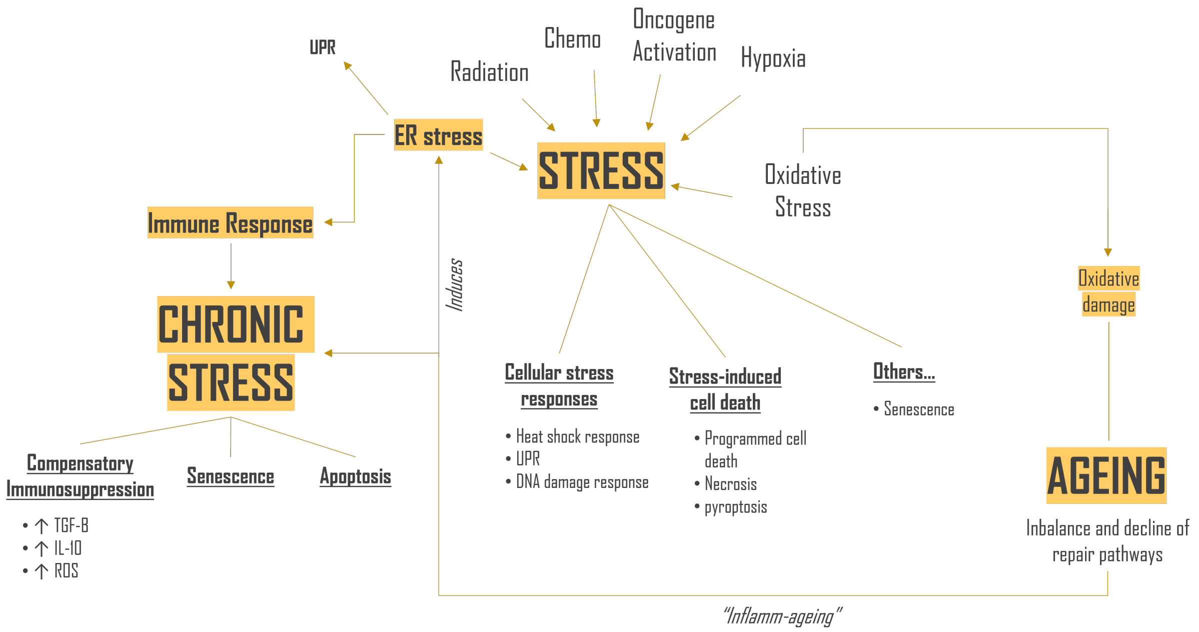


FIGURE 2.3: Stress and the oxidative damage theory of ageing
Illustrative diagram of stress relationships, including endoplasmic reticulum stress and oxidative stress, with natural physiological ageing [29–31].

ROS, which can be damaging [25]. Additionally, neuronal precursor stem cells are lost, particularly in the hippocampus [1], being this region extremely important for the consolidation of memory and the establishment of spatial memory [2, 34]. Several hallmarks of brain ageing can also make neuronal circuits more prone to excitotoxicity, which is damage caused by hyper-excitability, including oxidative stress caused by mitochondrial dysfunction [2, 4]. Finally, perhaps one of the most studied phenotypes in ageing is inflammation [1, 2, 4, 27]. Interestingly, under the scope of different causal and temporal mechanisms sometimes this is even being referred to as “inflammageing” [35].

Molecular Profile of the Ageing Brain

The pace of ageing is controlled by genetic pathways and biological processes conserved in evolution [13]. Although the physiological changes in gene expression that occur in the ageing brain depend on several factors, such as the brain region, the sex of the subject and even the inter-subject biological variability, there are some consistent and interesting changes. In the human brain, most genes have a trend for decreased expression with age, mainly associated with protein processing and energy generation [36]. However, there are some whose expression is increased in aged brains, these genes being associated with immune activation and inflammation. Glial cells shift their gene expression towards an inflammatory phenotype, and this shift is not observed in neurons [1, 36].

Functional and Morphological Changes in Aged Brain Cells

Functions and relationships between the various cells also change in the aged CNS, although it is not yet clear why.

Briefly, all glial cells show changes that can compromise their neuroprotective role (figure 2.4). In particular, ageing microglia demonstrate a predisposition to the inflammatory phenotype, astrocytes appear to lose their synapse-maintaining ability (as will be further explored in the next sections), and oligodendrocytes modify their axon myelination capabilities, which may impact the speed of electrical message transmission. Since all cells are essential for the proper functioning of the CNS, and since they are all in constant interaction, any disruption in their functioning can compromise neuronal support and increase neuronal vulnerability to aggression, which could explain the ageing brain's predisposition to cognitive decline and neurodegenerative diseases [37].

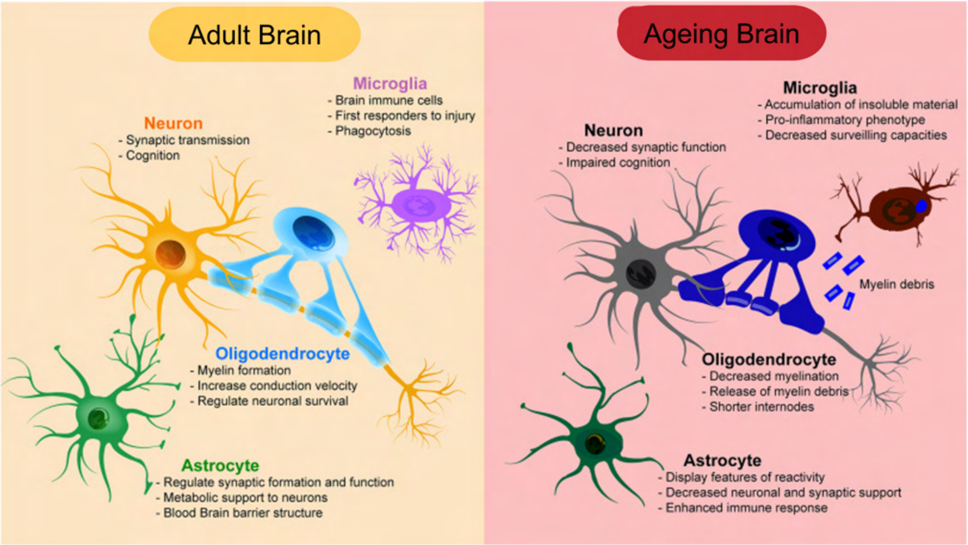

Adult Brain
Microglia
- Brain immune cells
- First responders to injury
- Phagocytosis
Neuron
- Synaptic transmission
- Cognition
Oligodendrocyte
- Myelin formation
- Increase conduction velocity
- Regulate neuronal survival
Astrocyte
- Regulate synaptic formation and function
- Metabolic support to neurons
- Blood Brain barrier structure
Ageing Brain
Microglia
- Accumulation of insoluble material
- Pro-inflammatory phenotype
- Decreased surveilling capacities
Neuron
- Decreased synaptic function
- Impaired cognition
Myelin debris
Oligodendrocyte
- Decreased myelination
- Release of myelin debris
- Shorter internodes
Astrocyte
- Display features of reactivity
- Decreased neuronal and synaptic support
- Enhanced immune response

FIGURE 2.4: Ageing brain cells
Schematic illustration of the functional changes that brain cells undergo with the physiological ageing process, and that are believed to impair their neuroprotective roles and increase the predisposition of the CNS to neurodegenerative diseases. Adapted from Salas *et al.* (2020) [37].

It is also known that the brain loses volume and weight with ageing, with a volume loss of about 5% per decade after the age of 40 [38]. However, what causes this to happen is not fully understood. Specifically, the loss of volume associated with gray matter is thought to be associated with the decrease in dendrite branches [38]. These volume losses are also highly dependent on the brain area: the prefrontal cortex appears to be the most affected, as well as the hippocampus and cerebellum [38]. More studies are needed to clarify these questions.

In conclusion, ageing consists of a multitude of factors that cause disruptions in the CNS homeostasis. Given that these effects are responsible for cognitive decline in the elders [26], it is urgent to discover what are the specific mechanisms responsible for the appearance of the primary hallmarks of ageing.

2.1.3 Pathological Ageing, Neuroinflammation and Neurodegenerative Diseases

Taking all of the above into consideration, the brain is still quite resilient to physiological ageing. In fact, there are no physiological, cellular, and molecular changes that alone compromise the elderly to the point of being completely dependent on others and cognitively inept. For instance, it is not uncommon to observe two 70-year-olds in the opposite side of the spectrum: one fully capable of their cognitive faculties, and one already with evidence of Alzheimer's Disease or other neurodegenerative disease. Such may be underlain by neurobiological differences between subjects with the same chronological age [39], and the subtle changes that occur and may predispose the tissues to age-related diseases (cancer, cardiovascular and neurodegenerative disorders) is called pathological ageing. There is a need to uncover the functional and molecular differences between functionally impaired and unimpaired elders to tackle disease prevention.

What are Neurodegenerative Diseases?

Neurodegenerative diseases are some of the most nefarious disorders affecting the human body, as they compromise neurons and thus impact several systems. These diseases are characterized by a chronic, progressive loss of the structure and functions of the nervous system, resulting, in the case of the brain, in deep cognitive and functional decline [40]. However, the aetiology of neurodegenerative diseases remains poorly understood.

The most common neurodegenerative diseases are Alzheimer's Disease (AD), Parkinson's Disease (PD), and Amyotrophic Lateral Sclerosis (ALS) [25]. Alzheimer's disease is the most common cause of dementia, associated with 60-70% of all cases worldwide [41]. Dementia is defined as the deterioration of cognitive function beyond what would be expected with physiological ageing [41]. According to the World Health Organization (WHO), dementia affects, among others, memory, reasoning, and orientation, leaving patients very dependent as it progresses [41]. AD is also a progressive disease, since the patients' symptoms worsen as time passes. Unfortunately, AD has no cure and, on average, the life expectancy of a patient with AD is estimated to be around 8 years after diagnosis [41]. Some of the brain areas most affected by AD are the hippocampus and the entorhinal cortex, both involved in memory processes [42, 43]. PD is also a progressive neurodegenerative disease that primarily affects motor functions. The first symptoms of this disease are tremors, commonly in the hands, and then patients start to progressively feel stiffness and slowness of their movements [44]. Contrary to AD, the area most affected by PD is the basal ganglia, more specifically the substantia nigra [45]. Finally, ALS is a neurodegenerative neuromuscular disease that causes the loss of motor neurons, which mostly affects the primary motor cortex, brainstem and spinal cord [46]. The main symptoms of these patients include muscle stiffness and twitches, as well as difficulty in speaking and/or swallowing. A small percentage of patients with ALS may also show signs of dementia, being a disease with a quite wide spectrum of symptoms [47]. Like AD, PD and ALS have no cure [44, 46].

However, there are some treatments capable of delaying their symptoms. For example, there is currently one Food and Drug Administration (FDA) approved treatment for AD, although under the so called "accelerated approval pathway" [1], that addresses the underlying molecular biology of the disease, removing plaques of beta amyloid protein [48], thought to be associated to cognitive decline.

[1] The FDA can approve a drug for a life-threatening disease that may have a therapeutic benefit over existing treatments, even when there is uncertainty about the drug's clinical benefit [48].

What is the Prevalence of Neurodegenerative Diseases?

The prevalence of neurodegenerative diseases is alarmingly increasing in the population as it ages [25]. It is estimated that one in 10 individuals over 65 years of age has AD and this number increases to 50% in individuals over 95 years old [25]. Unfortunately, the numbers associated with PD and AD are equally frightening, as it can be observed in figure 2.5 (data from the United States of America (USA) [25]). There is plenty of evidence that structural changes in the brain occur years before the cognitive and functional decline associated with neurodegenerative diseases [49]. Finding the molecular basis of these diseases relies on exploring their mechanisms before their onset. However, the impossibility of predicting the onset of disease, as well as the difficulties regarding collection of brain tissue to study its cellular mechanisms in humans, make that task very complicated [25, 49].

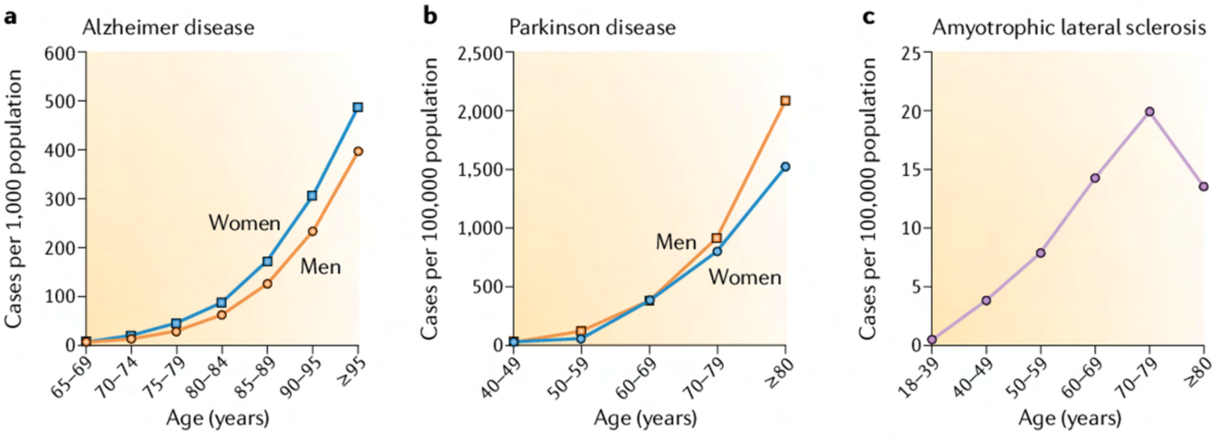


FIGURE 2.5: Prevalence of neurodegenerative diseases
Prevalence of **a)** Alzheimer's Disease per 1000 citizens in the USA; **b)** Parkinson's Disease per 100000 citizens globally; **c)** Amyotrophic Lateral Sclerosis per 100000 citizens in the USA. Retrieved from Hou *et al.*, 2019 [25].

What are the Causes of Neurodegenerative Diseases?

Although the timeline and causality of events associated with neurodegenerative diseases are not known [24], and despite these diseases probably having different causes, there are already several clues that associate them with the main hallmarks of ageing [25] (figure 2.2). Some authors suggest that the onset of AD includes, among others, inflammation (**inflammageing**), DNA damage and mitochondrial dysfunction [25]. In older adults, inflammation has been related to cognitive decline and structural changes in the brain, and chronic inflammation has also been associated to many of the most nefarious neurodegenerative diseases [50]. Moreover, the antagonist process of inflammation – the compensatory immunosuppression – has also been in the centre of many studies, given that even this process can be damaging by itself, evoking harmful effects in the brain tissue, and possibly promoting the risk of tissue degeneration and age-related diseases [31, 51].

There are several theories that try to explain the exact relationship between ageing and neurodegenerative diseases. Specifically, one of them admits a continuum between ageing and the appearance of neurodegenerative diseases, such that all ageing will eventually lead to neurodegeneration [25]. Furthermore, inflammation has also been suggested as the primary cause of pathological ageing [27]. Some studies established the relationships between ageing and cognitive decline, and between inflammation and cognitive dysfunction, thus it will not be unreasonable to hypothesize that inflammation may exacerbate age-associated cognitive decline and pathological ageing [27]. However, it is increasingly agreed among the scientific community that the key to a better understanding of neurodegenerative diseases, and their possible therapeutic reversal, lies in a deeper study of the physiological mechanisms of ageing.

Lately, there is an interesting relationship between two of the most harmful classes of diseases of the 21st century, neurodegenerative diseases and cancer. It is known that the physiological integrity concomitant with ageing is not only the primary risk factor for neurodegenerative diseases, but also for cancer [13], among other diseases. Interestingly, an inverse correlation between the incidence of cancer and the incidence of neurodegenerative diseases has already been demonstrated [52]. While the prevalence of cancer reaches its peak around the age of 60, it tends to decrease with age while the prevalence of neurodegenerative diseases starts to increase [52]. Additionally, it is thought that many genes are differentially expressed in both neurodegenerative diseases and astrocytoma (tumour of the central nervous system originated from an uncontrolled proliferation of astrocytes [53]), which illustrates the need for further understanding the mechanisms behind both conditions. In particular, the gene encoding for the TAU protein is expressed in both conditions. Although in AD the accumulation of TAU protein causes the formation of neurofibrillary tangles, thought to be toxic, in astrocytoma the overexpression of this protein can be also beneficial, acting as a break for the formation of blood vessels, being thus associated with better prognosis [54].

In summary, the ageing brain incurs in several molecular and functional changes, felt mainly through cognitive decline. However, unknown subtle changes in the environment, molecular signalling, or cell behaviour are thought to trigger pathological ageing, manifested as numerous neurodegenerative diseases, such as AD and PD, and resulting in the impairment of everyday tasks [26]. It remains then to know what mechanisms might be behind this change between physiological and pathological ageing.

2.1.4 Socioeconomic impact of the Ageing Brain in the Modern Era

Ageing is natural, normal, and an irrevocable event in all our lives, affecting every cell and organ in our body - the brain is no exception [1]. It is already common knowledge that, with ageing, most of individuals have a mild cognitive decline which leads to greater dependence in common daily tasks [26]. Moreover, many other ageing-associated diseases occur, with very high costs for the elderly and for the social system that supports them [25]. Ageing is the main risk factor not only for “physical” diseases or symptoms, such as typical muscle pain, but also for numerous neurodegenerative diseases, such as AD and PD [2–4, 13].

Both the average life expectancy and the world population’s median age have been increasing for the past 70 years [55]. According to the World Health Organization, several factors may be associated with these indicators, such as advances in medical care and changes in the leading causes of death (for example, from infections to chronic diseases) [56]. The median age of the world population has grown from 23.1 years in 1950 to 28.5 years in 2010, with prospects of an increase to 32.0 years in 2025, although there is still a huge gap between developed and undeveloped countries, as can be seen in figure 2.6 [55].

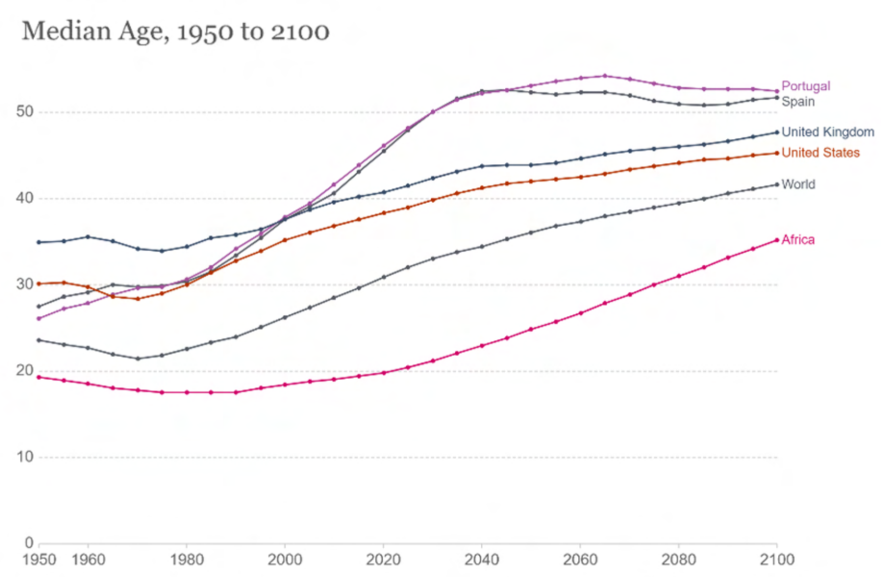


FIGURE 2.6: Median age of the world's population
Median Age by country and of the world population from 1950 to 2100, adapted from Our World in Data (https://ourworldindata.org/age-structure), by Max Roser. Data Source: United Nations Population Division (Median Age) – 2017 [55].

According to the same source, there is a trend also towards an increase in the median age of the world population in the next 30 years, both in developed and non-developed countries, indicating the growing ageing of the population worldwide [55]. However, despite these – apparently – positive indicators, the quality of life of the elderly does not follow this growing trend: we live longer, but we may not have proportional healthspan extension. Amongst other factors, the higher prevalence of neurodegenerative diseases in the elder population contributes for this.

In figure 2.7 are represented the disability-adjusted life years (DALYs) per sex and neurological disorders [57], that is, ”the sum of the years of life lost to due to premature mortality (YLLs) and the years lived with a disability (YLDs) due to prevalent cases of the disease or health condition in a population”, according to WHO [58]. It can be noted that, as people age, the number of DALYs associated with neurodegenerative diseases, in light blue and dark green, increases immensely. However, it is not yet clear what happens and what drives the change from normal to pathological ageing, only that age is a factor that increases this predisposition.

There is therefore an urge to find out more about the mechanisms that are associated with ageing, in order to ensure that the increase in life expectancy is accompanied by an increase in the associated quality of life [27]. In fact, considering the socioeconomic impact that ageing and neurodegenerative diseases have on the world population, there are increasing efforts by the scientific community to understand these conditions [24], reflected by the number of associated publications per year (figure 2.8).

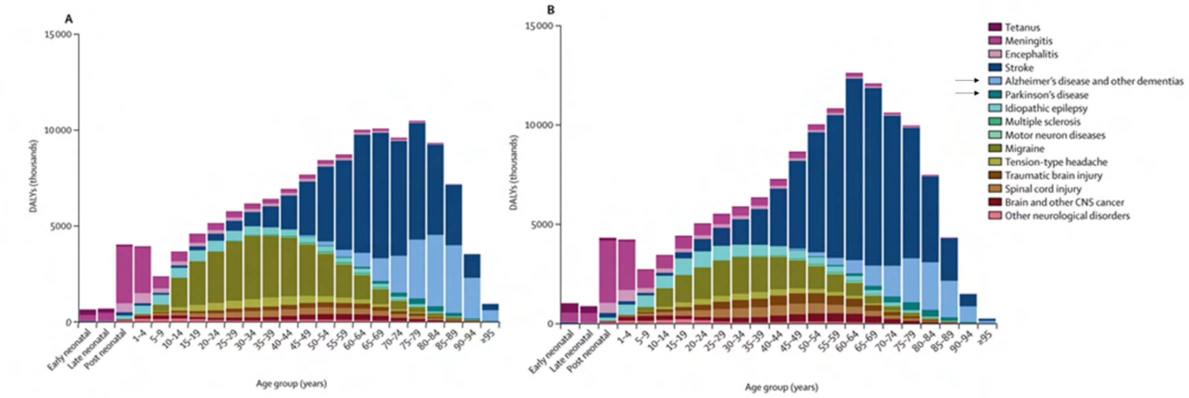


FIGURE 2.7: Global DALYs for neurological disorders (identified by black arrows) by sex and age Global DALYs for neurological disorders by sex and age, 2016 **(A)** Females. **(B)** Males. DALY=disability-adjusted life-year. Adapted from Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016 [57].

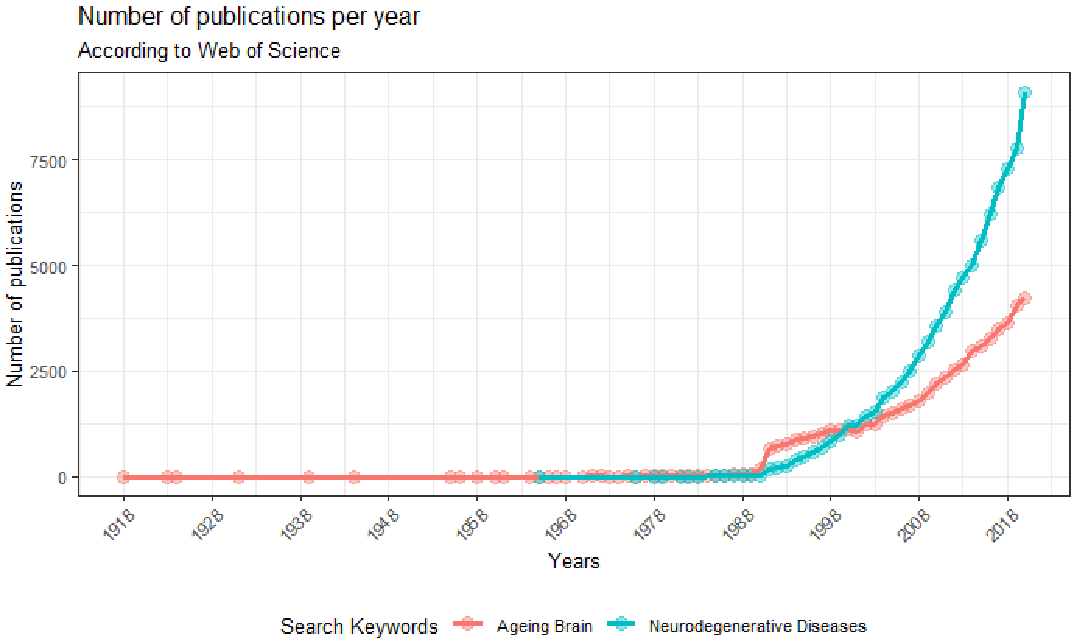


FIGURE 2.8: Web of science publications on the ageing brain and neurodegenerative diseases Number of publications per year associated with keywords ”ageing brain” (red) and ”neurodegenerative diseases” (blue), from 1918 to 2019, according to Web Of Science. Plotted using RStudio software.

2.2 Astrocytes

Astrocytes are a fascinating group of cells, indispensable for the proper functioning of the CNS. Given the high number of functions that these cells perform, any disruption of their homeostasis can be expected to result in a major impact on brain functions [7]. Astrocytes are shaped like stars, hence their name (from the ancient Greek *astron* – star – and *kútos* – cell), and are present in the brain and spinal cord. Interestingly, a single astrocyte interacts with up to two million synapses [59]. These cells are very important from the point of view of neuroprotection and maintenance of the CNS, which is why they deserve a prominent position in this Master's thesis [7].

2.2.1 Heterogeneity

For many years, astrocytes were thought to play a primarily structural role. However, with the increase in scientific knowledge, it is now known that these cells are essential for the proper functioning of neurons and CNS homeostasis [19].

Structurally, astrocytes are classified into two main groups. **Fibrous astrocytes** (figure 2.9 (A)) have long, unbranched processes, being mainly located in white matter. **Protoplasmic astrocytes** (figure 2.9 (B)) have highly branched processes and are present in gray matter [60]. The expression level of glial fibrillary acidic protein (GFAP), an intermediate filament protein, has been widely used as an astrocytic marker, although it is known that other glial cells and even neurons can express this marker [61]. Moreover, fibrous astrocytes have higher expression of GFAP than protoplasmic astrocytes [60].

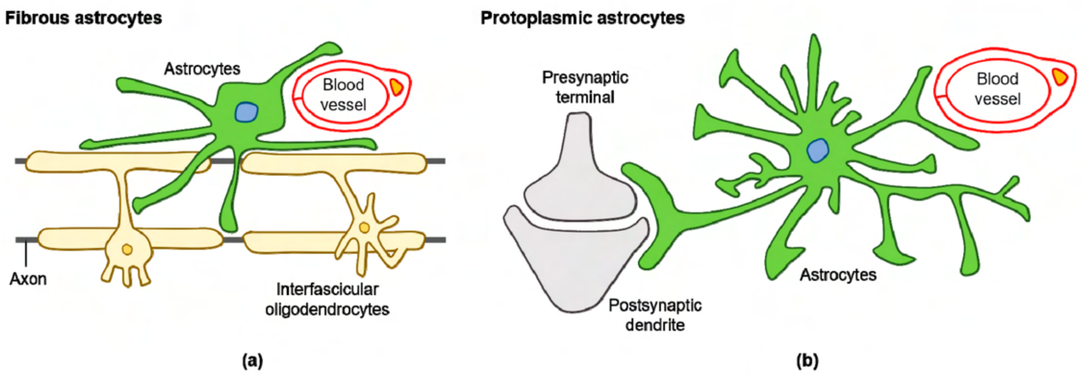


FIGURE 2.9: Types of astrocytes
Schematic illustration of the different structural classification of astrocytes: **(a)** fibrous astrocytes, with slender and longer processes, interacting with other brain cells; **(b)** protoplasmic astrocytes, with highly branched processes, being part of a tripartite synapse [62].

Among astrocytes' main functions (figure 2.10), we can highlight the modulation of neuronal activity, synapse homeostasis (including synaptic activity and plasticity, as well as neurotransmitter clearance), provision of trophic factors and nutrients to neurons and establishment and maintenance of the blood brain barrier [1, 7, 8, 63]. New studies also include in the functions of astrocytes the ability to generate brain rhythms and neuronal network patterns [7].

In terms of synaptic functions, astrocytes are responsible, together with pre- and post-synaptic nerve endings, for forming tripartite synapses (figure 2.9 (b)). This is a very typical structural and functional configuration of chemical synapses and allows astrocytes to perform their synapse homeostasis functions [64].

Different astrocytes can present different functions, with this not being an on-off state (just as a motor neuron cannot become a sensory neuron) and some brain regions presenting specific distributions of astrocytic functions. Such heterogeneity and possible variability of astrocytic proportions may contribute to the differences in the effects of neurodegenerative diseases between brain areas, with some regions suffering more than others [7].

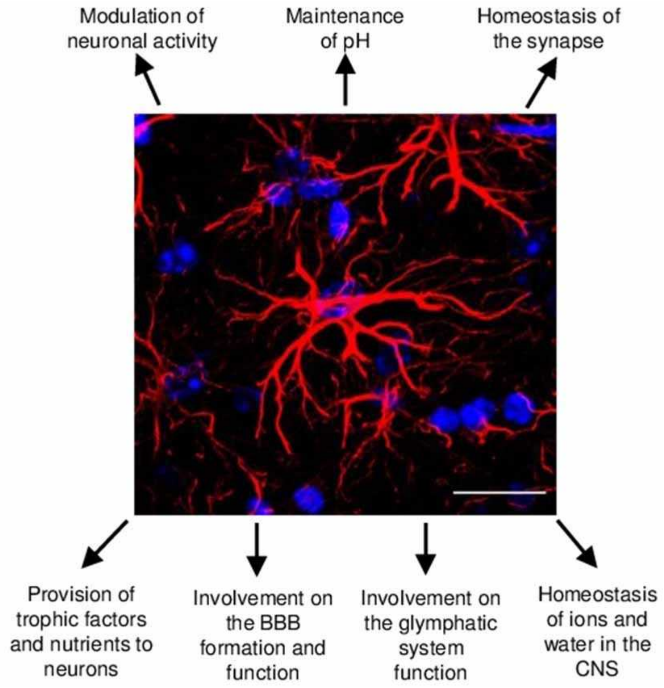

Modulation of neuronal activity
Maintenance of pH
Homeostasis of the synapse
Provision of trophic factors and nutrients to neurons
Involvement on the BBB formation and function
Involvement on the glymphatic system function
Homeostasis of ions and water in the CNS

FIGURE 2.10: Astrocyte functions
Schematic representation of the heterogeneity of functions of astrocytes in the CNS [7].

Out of curiosity, the ratio of astrocytes to neurons increases with organism complexity [63]. There is no consensus yet on the relationship between that ratio and cognitive capacity: some authors argue that the relative increase in astrocytes is only associated with enhanced metabolic support required by the higher energy output associated with larger neurons and brains; conversely, other authors argue that larger neurons do not need more energy, and therefore the increase in the astrocyte-neuron ratio in humans may be associated with greater cognitive capacity [63]. Such apparently contradictory information increases the mystery and curiosity about these cells, which will certainly motivate the search for more knowledge on them.

In conclusion, astrocytes have a wide spectrum of possible functions, being therefore normal for any disruption of the CNS homeostasis to cause a major impact on their physiological functions [7].

2.2.2 Reactivity

Astrocytes can be activated by various stimuli and diseases, further increasing their complexity and range of functions [7]. **Reactive astrocytes** are characterized by morphological, molecular, and functional remodelling, in response to injury, disease, or infection of the CNS [7, 65]. This means that, following some stimuli, astrocytes can reversibly shift their molecular expression, while enlarging and losing some of their functions [7] (figure 2.11). A major difference between reactivity and heterogeneity of phenotypes is that reactivity is reversible and therefore should not be confused with astrocyte functional heterogeneity [65]. However, it is not yet clear whether this reactivity is beneficial or detrimental to the homeostasis of the CNS.

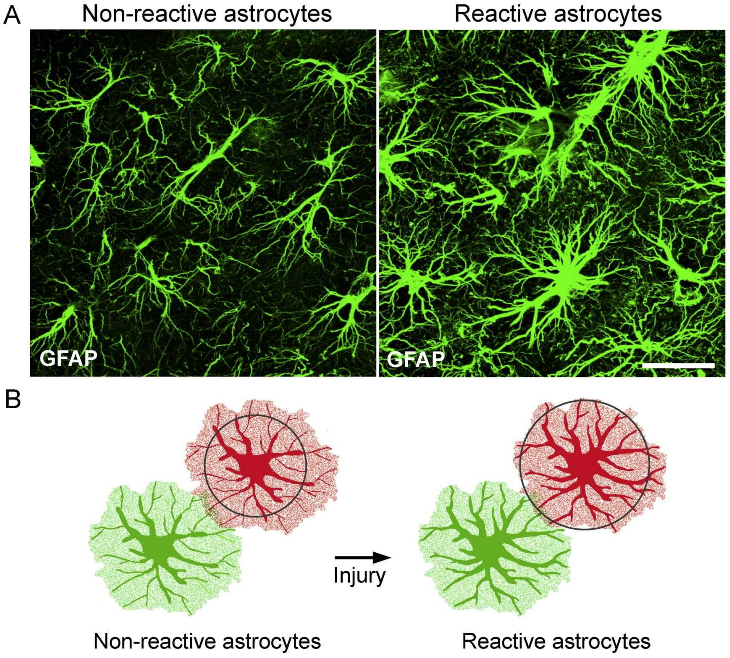


FIGURE 2.11: Reactive astrocytes
Reactive astrocytes undergo hypertrophy of cellular processes. **(A)** Staining of astrocytes with GFAP, a marker of astrocyte reactivity; **(B)** schematic illustration of morphological changes that reactive astrocytes undergo after their activation. [66]

One classification for astrocyte reactivity was proposed by Liddelow and his colleagues in 2017 and advocates two states of reactivity: **A1** and **A2** [67]. According to this classification, reactive astrocytes can develop into pro- (A1) or anti-inflammatory (A2). A1 astrocytes have been associated with neuroinflammation and tend to upregulate the expression of classical complement cascade genes shown to be destructive to synapses. This means that they lose their ability to promote neuronal survival. It is also known that microglia are fundamental and sufficient to trigger the reactivity of A1 astrocytes. On the other hand, A2 astrocytes are mainly associated with ischaemia and upregulate many neurotrophic factors, promoting CNS recovery and repair. Although being firstly described in mice, the same study validates their presence in human samples [67].

There are several putative reactivity marker genes. The most used marker to describe reactivity in general (called "pan" reactivity, that is, not restricted to the differences between A1 and A2 astrocytes) is GFAP (figure 2.11) [65]. This gene encodes for the Glial Fibrillary Acidic Protein, an intermediate filament protein that is a constituent of astrocytes. However, GFAP is not expressed by all reactive astrocytes (although those who express it are necessarily reactive) [65]. Other genes used to tag pan-reactive astrocytes are *VIM* (which also encodes for an intermediate filament protein but is expressed by endothelial cells and immature astrocytes), *CHI3L1* (whose function is still unknown) and *S100B* (a gene encoding Ca2+ binding protein that is expressed after injury) [65]. On the other hand, the *STAT3* gene is a marker of A2 reactive astrocytes and, while it encodes for a transcription factor, it can also be expressed by neurons [65]. Finally, the complement factor protein encoded by the *C3* gene is a marker of A1 reactive astrocytes, although it is also expressed by other glial cells [65].

However, the same group admits that this classification may be too simplistic and binary, and that reactivity may be a continuum between these two states, or even *n* distinct activation states [68]. More studies are needed to clarify this notion of reactivity. However, this simplistic binary classification is already beginning to give several clues on the true complexity of astrocytes and their potential harmful role in ageing contexts.

2.2.3 Astrocytes Communication

As stated above, it was thought that astrocytes were just “the glue of the brain”, having thus a minor structural role. However, it is currently known that this is not the case.

To perform their function, astrocytes need to communicate between them and with other cells. Astrocytes have ionotropic receptors in their membranes, made of ion channels that open or close in response to a ligand (such as neurotransmitters), and metabotropic receptors, which use signal transduction mechanisms (such as G Proteins) to modulate cellular activity [69].

These receptors may be activated by several neurotransmitters (such as noradrenaline, glutamate, GABA, among others) released by neurons in the tripartite synapse and have the ability to respond according to the intensity of synaptic activity, modulating their activity and maintaining homeostasis of the CNS. Furthermore, in response to these neurotransmitters, astrocytes undergo fluctuations in the intracellular levels of Ca2+ ions, depending on the intensity of the neuronal response, which leads to the release of glio-transmitters (neurotransmitters, but of glial origin), such as ATP and glutamate (figure 2.12) [70]. Despite this, the astrocytic response to calcium transients is not fast enough, with astrocytes mainly performing a modulating activity, rather than a message transmission one like neurons [71]. However, it is still unclear how astrocytes release these transmitters. With all these mechanisms, astrocytes are able to maintain the proper functioning of synapses, avoiding excitotoxicity through the renewal of neurotransmitters, and plasticity [64].

On the other hand, it is also known that astrocytes demonstrate spontaneous oscillation of Ca2+ levels, not dependent on neuronal activity. These “calcium waves” are distinct from the electrical transients that characterize neurons, as they have high amplitude, last for longer periods and are regular but spaced in time [64]. It is not yet known why this happens, but it has been shown that it occurs in some astrocyte subpopulations, and that it may be associated with modulation of neuronal activity, as well as communication between astrocytes. Furthermore, it is thought that these ionic transients may also be sufficient to excite adjacent neurons [64].

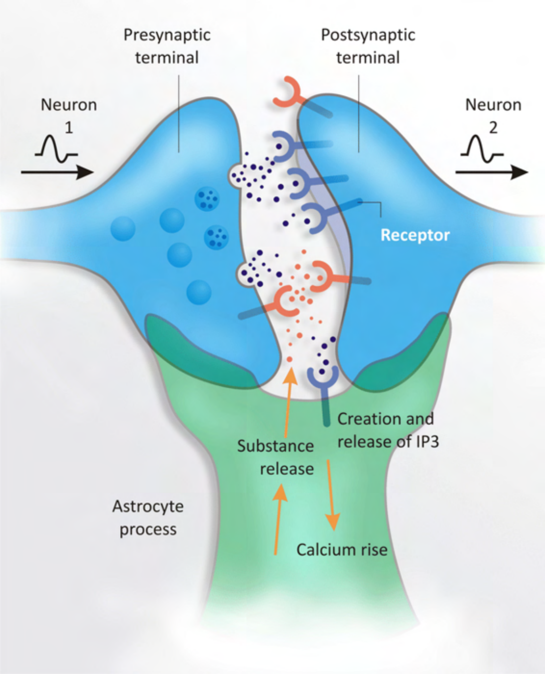


FIGURE 2.12: Tripartite Synapse
Schematic illustration of the tripartite synapse. Pre-synaptic neurons (neuron 1) release neurotransmitters to the synaptic cleft. These neurotransmitters create an action potential in post synaptic neurons (neuron 2) and trigger a Ca2+ transient inside the astrocyte (green). These calcium waves will trigger the release of glial transmitters that can modulate the synapse, depending on the neuronal activity. Astrocytes are also capable of having calcium transients independent from neuronal activity. IP3, or inositol 1,4,5-trisphosphate, is a second messenger that mediates the release of intracellular calcium. [72]

Although astrocytes cannot propagate action potentials, they are somewhat excitable, in the sense that they can be activated and communicate with other cells through glial transmitters and calcium waves [64]. This discovery has revolutionized the way we look at electrical transmission in the brain and the fundamental role of astrocytes in the CNS.

2.2.4 Astrocytes and Ageing

Functional Profile of Ageing Astrocytes

With age, our brain develops an inflammatory phenotype, often called inflammageing. With inflammation there is a recruitment of microglia, which in turn secretes inflammatory factors that trigger the reactivity of astrocytes. These can develop, simplistically, into pro-inflammatory or anti-inflammatory states [67]. However, the pro-inflammatory phenotype (A1 astrocytes) is the most prevalent with ageing. This causes transcriptional and functional changes, making astrocytes unable to perform their functions of promoting neuronal survival [9].

It is thought that some functional changes that ageing astrocytes undergo may be increasing the pro-inflammatory phenotype of the brain [1]. Specifically, aged astrocytes are thought to activate the complement system, which is part of the innate immune system and is responsible for regulating inflammation through the release of complement factors C3 and C4B [1]. Since astrocytes participate in the tripartite synapses, one hypothesis is that the complement system's action on these synapses decreases the strength of the connection between neurons and astrocytes, and that it potentiates memory loss in older people [1]. Furthermore, this may also be associated with the loss of the capacity to maintain synaptic homeostasis, with excitotoxicity being an important hallmark of brains affected by ageing and/or neurodegenerative diseases [73]. Excitotoxicity is mainly the result of prolonged or exacerbated activation of glutamate receptors, caused by the inability of astrocytes to control the levels of glutamate in the synaptic cleft, resulting in loss of neuronal function and cell death [73]. Furthermore, it is known that aged astrocytes have an increase in ROS release, which is related to the oxidative stress theory of ageing [1, 28]. It is also known that aged astrocytes lose part of their ability to maintain the proper functioning of the Blood-Brain Barrier (BBB) [1]. Finally, as they are an extremely heterogeneous cell group, any impairment on their function will irrevocably impact the function of other glial cells in addition to neurons, creating feedback mechanisms that result in dysfunction of the entire CNS [1].

This inflammatory phenotype has been shown to be detrimental to the proper functioning of astrocytes. However, recent studies are also starting to elucidate on the effect of natural compensatory immunosuppression in astrocytes. A review by Salminen (2020) states that the number of cells that fall into this phenotype tends to increase with age, and this compensatory immunosuppression phenotype has harmful effects on the tissues on which it acts, such as brain tissue, and may also be associated with the exacerbation of neurodegeneration and age-related diseases [31].

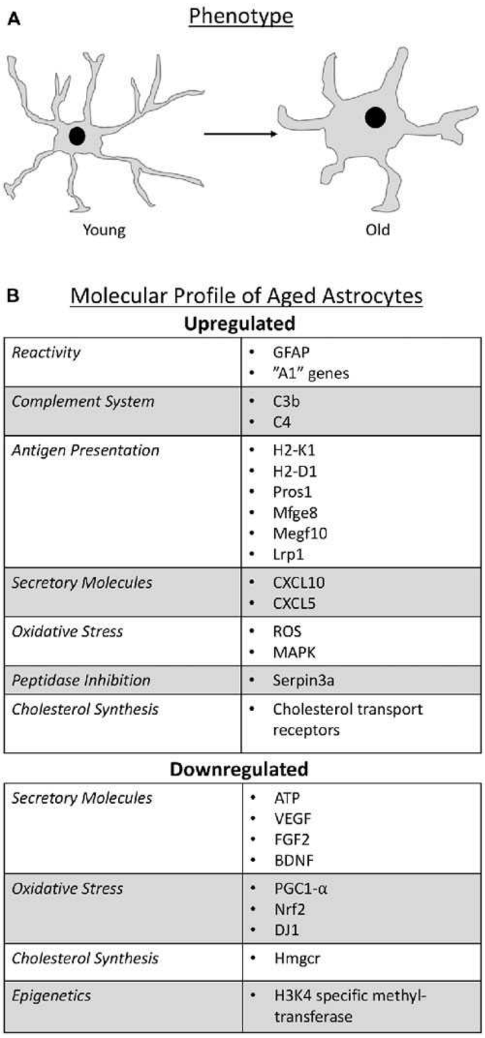
Finally, considering the oxidative stress theory of ageing with astrocytes, we can still find evidence that astrocytes are quite sensitive to both oxidative and endoplasmic reticulum stress, compromising their neuroprotective and homeostasis functions and adapting worse to these conditions as they get older [74]. However, more studies need to be done in order to clarify this issue, as other studies suggest that stress-reactive astroglia is not necessarily neurotoxic and that intense oxidative stress does not result in its exacerbation by glia or neurons [75].

Structural Profile of Ageing Astrocytes

In structural terms, despite being highly dependent on the region in which they are found, it is known that, with age, astrocytes start to have shorter and stubby processes, as opposed to the fine and branched processes of normal astrocytes (figure 2.13 (A)) [1]. There is no evidence that their number changes significantly with age [37].

Transcriptomic Profile of Ageing Astrocytes

In general, consistent with astrocytes becoming more reactive with age, mainly by acquiring a pro-inflammatory phenotype (A1 reactive astrocytes), we find genes such as *GFAP*, *S100b* and other ”A1” genes whose expression is increased in aged astrocytes (figure 2.13 (B)). Furthermore, we find up-regulated genes associated with the complement system, such as *C3* and *C4B*. On the other hand, we find down-regulated genes associated with secretory molecules, such as *ATP* and *VEGF*, and genes that regulate oxidative stress, such as *NRF1* and *DJ1*. However, all these transcriptomic studies were carried out in mice and/or humans, using RNA-seq of pools of cells [1]. This means that all these changes will reflect an “average” transcriptomic profile of the astrocytes, not having the sensitivity to identify more subtle changes in the transcriptome of individual cells [76].



Molecular Profile of Aged Astrocytes

Upregulated

Reactivity	• GFAP • "A1" genes
Complement System	• C3b • C4
Antigen Presentation	• H2-K1 • H2-D1 • Pros1 • Mfge8 • Megf10 • Lrp1
Secretory Molecules	• CXCL10 • CXCL5
Oxidative Stress	• ROS • MAPK
Peptidase Inhibition	• Serpin3a
Cholesterol Synthesis	• Cholesterol transport receptors

Downregulated

Secretory Molecules	• ATP • VEGF • FGF2 • BDNF
Oxidative Stress	• PGC1-α • Nrf2 • DJ1
Cholesterol Synthesis	• Hmgcr
Epigenetics	• H3K4 specific methyl-transferase

FIGURE 2.13: Molecular and morphological changes of ageing astrocytes
Summary of **(A)** the morphological changes and **(B)** molecular changes that astrocytes undergo with ageing. With ageing, astrocytic processes become shorter and stubbier, and most of their neuroprotective functions become dysregulated [1].

Ageing Astrocytes and Neurodegenerative Diseases

Aged astrocytes have also been associated with several neurodegenerative diseases. Specifically, they have been related to AD, that shares many of the hallmarks of ageing brain and ageing astrocytes, such as oxidative stress, mitochondrial dysfunction, and inflammation [25]. PD and ALS have also recently been associated with ageing astrocytes and their consequent loss of function [1, 7, 8, 77]. Since this cell group is very affected by age and given its complexity, it is plausible that there are subtle transcriptional changes associated with ageing astrocytes that remain unnoticed and make the brain more vulnerable to age-related diseases.

In short, and citing Soreq and colleagues (2017), "the intimate relationship between ageing and neurodegeneration raises the possibility of shared transcriptional and post-transcriptional gene regulation programs" [78] - a better understanding of neurodegenerative diseases involves a better understanding of the processes in physiological ageing.

2.3 Single Cell RNA sequencing Data

High-throughput RNA sequencing (RNA-seq) allows the study of the molecular mechanisms of health and disease by sampling the transcriptomes of tissue samples, enabling the quantification of the relative levels of different RNAs therein [79]. This technology comprises reverse transcription from RNA to complementary DNA (cDNA), DNA fragmentation, fragment amplification and detection of the resulting base pair sequences, the so-called reads. After aligning reads against an annotated genome sequence, they are counted for each gene in each sample and summarised in a matrix of **read counts**, with genes as rows and samples as columns.

2.3.1 What is Single Cell RNA sequencing?

Single-cell RNA sequencing (scRNA-seq) is the current gold standard for profiling the transcriptomes of individual cells and thereby inferring their phenotypes. Being a high-throughput technology, it can profile thousands of cells per experiment, allowing at the same time for the study of a single cell transcriptome in an unbiased manner, not targeting specific genes like microarrays [80]. It is widely used for discovering new cell states in heterogeneous samples, such as the tumour micro-environment [81].

To reach single-cell resolution, scRNA-seq protocols require, among others, a step for cell isolation and transcript amplification. Two of the main categories of these protocols include well-based protocols and droplet-based protocols [82, 83]. **Well-based protocols** rely on methods, such as fluorescence-activated cell sorting (FACS) or microfluidic chips, for physical separation of cells in separate wells. Although allowing for flexible experimental set-ups, as the cells can stay in the wells for a certain amount of time, these protocols require manual pipetting for each individual well, in order to perform reverse transcription, being also very expensive and potentially introducing more noise in the samples [83].

Droplet-based protocols (figure 2.14) are based on the mechanical isolation of each cell using a droplet of oil. Each droplet contains a small bead (each coloured dot in figure 2.14), coated by many repeated complementary DNA (cDNA) sequences with five main parts: a linker region, a primer region to allow for further molecular amplification of each transcript, an unique barcode, a second series of barcodes called unique molecular identifiers (UMI), and finally the poly-d(T) region that allows the capture of mRNA [84]. Although each bead has only one unique barcode (allowing for the identification of each cell), it has numerous distinct UMIs, allowing for the unique identification of each transcript in each cell. These molecules can then be pooled after reverse transcription [83], amplified with PCR and sequenced with high-throughput state-of-the-art-technologies, such as Illumina, without losing the single-cell resolution of the transcriptome.

One of the most used droplet-based protocol is Chromium from 10X Genomics, that ensures one of the highest capture efficiencies amongst scRNA-seq technologies while being relatively affordable [86]. In this work, I have used scRNA-seq data prepared with this protocol.

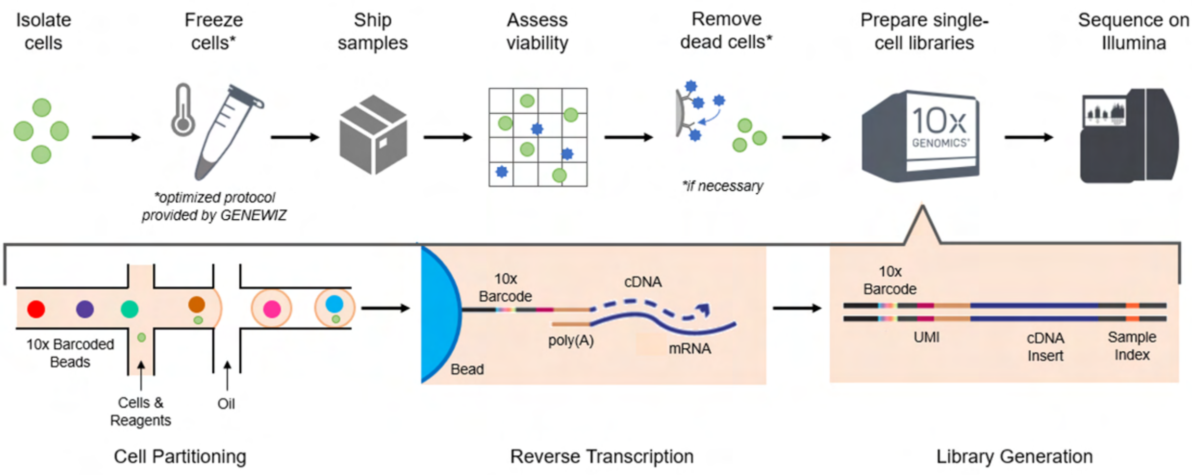


FIGURE 2.14: Droplet-based single-cell RNA sequencing platform
Droplet-based protocol for scRNA-seq with Chromium 10X Genomics® . Adapted from "Single-Cell RNA Sequencing Frequently Asked Questions" [85]

2.3.2 Single Cell versus Bulk RNA Sequencing

The first single-cell transcriptomics study was published in 2009 and focused on a mice blastomere [10]. Since then, scRNA-seq has been increasingly applied, given its enormous advantages in unravelling the complexity and heterogeneity of cell groups. Bulk RNA-sequencing experiments allow to measure gene expression levels as averages across thousands of cells. However, if there is high heterogeneity within the group of cells to be sequenced, transcriptome individualities are lost [76]. With single-cell RNA-seq, we can study each cell individually, obtaining the distribution of gene expression levels across a population of individual cells (figure 2.15). Together with clustering algorithms, we can see, among others, the differences in expression between cell types, heterogeneity within cell types, study differentiation trajectories and differences between different cell type-specific responses [87].

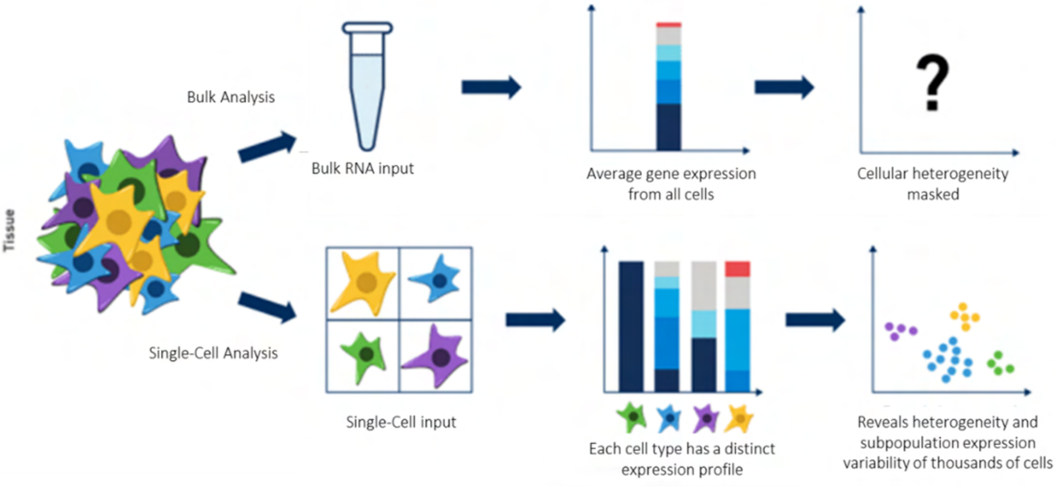

Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked
Tissue
Single-Cell Analysis
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells

FIGURE 2.15: Bulk RNA-seq versus scRNA-seq
Unlike bulk RNA-seq, which provides an average transcriptomic profile of a sample with numerous cells, scRNA-seq has the ability to find distinct cell groups within the same sample. Adapted from "Single-Cell RNA-Seq: An Introductory Overview and Tools for Getting Started" [87].

2.3.3 Advantages and Disadvantages of Single Cell RNA Sequencing

Given the technical limitations of the scRNA-seq technologies (low amounts and inefficient capture of mRNA molecules, leading to sampling bias, in individual cells), there is a high number of lowly expressed genes in each cell with no reads [88]. These dropout events lead to a zero-inflation of the count matrix highly characteristic of scRNA-seq data, motivating the adaptation of protocols for the analyses of bulk transcriptomes. However, the possibility of unravelling the true heterogeneity of a tissue compensates for those drawbacks, and there are already plenty of bioinformatics tools that aim at dealing with them [89]. Furthermore, we are now witnessing the emergence of new protocols that even couple this enormous transcriptional resolution with spatial information.

2.3.4 Single-Cell versus Single-Nucleus RNA Sequencing

Single-nucleus RNA sequencing (snRNA-seq) is an important variation of single-cell RNA sequencing. The single-nucleus protocol was developed based on the scRNA-seq protocol to extend its applicability to tissues that cannot be easily dissociated into a single-cell suspension [90], such as the human brain (given that neurons are highly connected and very long, being difficult to dissociate entirely [91]), or frozen tissues (given that nuclei are better preserved than the whole cell [92]). At the same time, snRNA-seq minimizes the alteration of gene expression that may be introduced by artificial interactions between cells in suspension [93].

This technology is based on four steps: tissue processing, nuclei isolation, nuclei sorting and sequencing, being the first two steps those that differ the most from scRNA-seq [94]. For this, nuclei dissociation protocols are used, where the cells are suspended and lysed, for the nuclei to be separated from the cytoplasm using centrifugation.

2.3.5 Advantages and Disadvantages of Single-Nucleus RNA Sequencing

The main advantage of snRNA-seq is that it combines the advantages of scRNA-seq with the possibility of applying such technique to brain or frozen tissues, without losing the cells identity. However, it has lower RNA input amounts, which can also increase noise [90]. Furthermore, due to technical limitations, some extra-nuclear contents may be encapsulated along with the nuclei and may also increase the amount of cell debris and background RNA (i.e., RNA from the cytoplasm, mitochondria, etc.) [95].

Chapter 3

Materials and Methods

3.1 Data Availability

All frozen human brain tissue snRNA-seq datasets used in the present work are publicly available through the National Center for Biotechnology Information (NCBI) data repository Gene Expression Omnibus (GEO) [96]. Processed snRNA-seq data from these datasets (read count tables) were downloaded from the GEO data portal[1]. Moreover, the independent brain RNA-seq validation datasets used were retrieved from the Genotype-Tissue Expression (GTEx) project, a publicly available resource to study tissue-specific gene expression and regulation, with an associated tissue bank with relevant clinic metadata. Processed GTEx v8 RNA-seq data (read count tables) were downloaded from the project's data portal[2]. Donor metadata were obtained from dbGaP - database of Genotypes and Phenotypes (Accession phs000424.v8.p2).

TABLE 3.1: Summarised description of the human datasets used in this work, including the ageing astrocyte datasets (scRNA-seq) and the validation brain datasets (RNA-seq).

Dataset ID	Designation	Technology	Number of individuals	Brain Area	Title of the study	Ages
GSE153807	Young	snRNA-seq	4	Temporal Cortex (TC)	"Single nucleus RNA-Seq is not suitable for detection of microglial activation genes in humans"	7, 20, 24, 50
GSE141552	Youngoldish		4	Pre frontal cortex (PFC)	"Single cell transcriptome profiling of the human alcohol-dependent brain samples"	44, 56, 58, 69
GSE159812	Oldish		4	Medial pre frontal cortex (MPFC)	"Dysregulation of brain and choroid plexus cell types in severe COVID-19"	58, 77, 79, 82
GSE160936	Old		6	Entorhinal Cortex (EC)	"Diverse human astrocyte and microglial transcriptional responses to Alzheimer's pathology"	73, 74, 77, 80, 81, 91
phs000424.v8.p2	GTEx Cortex	RNA-seq	245	Cortex	-	20-70
	GTEx Hippocampus		191	Hippocampus		
	GTEx Cerebellum		234	Cerebellum		

[1] Gene Expression Omnibus - `https://www.ncbi.nlm.nih.gov/geo/`

[2] Genotype-Tissue Expression project - `https://www.gtexportal.org/`

3.1.1 Ageing Astrocytes Datasets

In order to find human transcriptomic data from single-nucleus RNA sequencing of healthy individuals of various ages, i.e, whose cause of death is not related to neurodegenerative diseases, control samples from several publicly available studies were used (table 3.1). For this, an extensive search was made in snRNA-seq data available on the GEO portal, using the keywords *scRNA-seq, epilepsy, Memory, Alzheimer, Alcohol, COVID-19*[3] *and Huntington*, in order to select **control individuals** for these studies, who did not have pathologies associated with the brain that could greatly impact the conclusions of this work (table A.1). The 60-year-old threshold was chosen to separate young and aged individuals, as it is thought that the first signs of neurodegenerative diseases may start to appear up to 30 years before their onset (80-90 years old for AD, for example [49]). To balance the number of young and old individuals, four datasets were chosen (GSE153807, GSE141552, GSE159812, GSE160936), with overlapping ages so that age ranges are not totally confounded with the studies and thereby facilitate the removal of batch effects. With these criteria, the joint dataset consisted of eight samples from young donors and ten samples from old donors. The names given to the datasets in this work are "young", "oldish", "youngoldish" and "old", and reflect the dominant ages in each of them, with the numbering of each sample within each dataset (e.g., young1, young2, etc.) being random (table A.1). To avoid introducing a bias towards a chosen brain area, different brain areas were used. Furthermore, the young dataset is the only one not comprising post-mortem samples but samples from living individuals with epilepsy, as it is naturally very difficult to find post-mortem samples from young individuals.

Although some count matrices were made available already with a few quality filters applied, I decided to revise the quality control and apply my own filtering criteria to data pre-processing.

All datasets were supposed to include a mixture of all CNS brain cells and the identification of cell types was performed in the downstream analysis, through known cell type-specific markers. However, it is suspected that the old dataset only had astrocytic and microglial cells. Finally, although all the data used in this analysis are snRNA-seq, to simplify the language, and given that scRNA-seq is the foundation of snRNA, I will henceforth not make a distinction between the two and refer everything as scRNA-seq.

3.1.2 Validation Datasets

To validate the results of this work in independent samples, the GTEx project RNA-seq dataset [97] was used, comprising donors with a wide range of ages (table 3.1). The full dataset was available as a matrix of gene counts, to which we performed sample filtering to obtain only the brain regions of interest, namely the Cortex, Hippocampus and Cerebellum. As the detailed metadata of this project are confidential, institutional authorised access to them was needed to study the change in the proportion of cells with age.

In order to avoid the bias associated to brain pathologies, GTEx samples associated with dementia, PD, cerebral vascular accident, and unknown cause of death were removed, which reduced the number

[3] Given the great amount of data recently made available.

of samples for analysis in roughly 3-4 % (from *n*=255 to *n*=245 in the cortex, from *n*=197 to *n*=191 in the hippocampus and from *n*=241 to *n*=234 in the cerebellum).

GTEx cortex samples were chosen to ensure the comparability between GTEx validation samples and scRNA-seq cortex samples. However, hippocampal and cerebellum datasets were also considered, as these areas are very affected by neurodegenerative diseases.

3.2 R Statistical Software

Most of the work of this thesis was performed using the `R` software environment for statistical computing and graphics [98]. This programming language is widely used by statisticians and computational biologists, as it is an intuitive and efficient language for the analysis of big data. `R` was used to import and pre-process data, as well as to render plots that illustrate the main results of this work. R is an open-source programming language, being constantly improved in terms of resources for data scientists, and with a helpful large online community. R-Studio [99] is an integrated development environment (IDE) for `R`, which allows its usage in a graphical, user-friendly way, including resources such as debugging, plotting and help in its graphical interface. In this work, R Studio Web (with `R` version 4.1.0) was used to run the analysis in the laboratory's server, given that single-cell analysis can be computationally demanding, easily reaching 50GB, and sometimes almost 200 GB of RAM. The main `R` packages used in this work and their respective versions are summarised in table 3.2. However, it should be stressed that the number of packages used in this work was virtually higher through package dependencies.

TABLE 3.2: Summary of the main `R` packages used in this work, all free and open-source.

Package	Version	Description
`ggplot2`	3.3.9000	Used for declaratively creating graphics, based on The Grammar of Graphics. It works based on mapping variables to aesthetics (colour, fill, shape).
`SingleCellExperiment`	1.14.1	Used for storing data from single-cell experiments in an intuitive, efficient, and organized way, containing functions for filtering, visualisation, and dimensionality reduction.
`Seurat`	4.0.4	Used for filtering, analysis, and exploration of scRNA-seq data. It also offers its own algorithms for batch effect correction (data integration) and clustering.
`limma`	3.46.0	Uses linear models for analysing designed experiments and the assessment of differential expression between variables.
`slingshot`	2.0.0	Used for trajectory inference in single-cell data, to uncover global structure and infer smooth lineages or relations between cell clusters.
`fgsea`	1.18.0	Uses preranked gene lists to perform gene set enrichment analysis (GSEA), a computational method that determines whether a set of genes shows statistically significant and concordant differences between two phenotypes.
`cTRAP`	1.10.0	Uses differential gene expression results to compare with those from known cellular perturbations (gene knockdown, small molecules), derived from the Connectivity Map, and allows for, among others, identification of candidate drugs mimicking/reversing those perturbations.

Amongst the enormous number of packages offered by `R`, `ggplot2` [100] is one worth mentioning. This package was developed to make data plotting easy and intuitive, allowing the plotting of different types of graphs always based on the same reasoning. Most, if not all, graphics included in this work were made using this package as a basis.

The two main packages used for data handling, pre-processing and results visualization in this work are `SingleCellExperiment` (SCE) [101] and `Seurat` [102]. These packages are quite similar in terms of data storage, as they allow the association of gene counts and metadata related to each cell in a single object, in an efficient way and requiring the minimum possible computational resources. They are also similar in terms of visualisation, as they are based on `ggplot2`. Although `Seurat` is more powerful and includes more visualisation and data processing tools (including its own batch effect correction algorithm), it is less intuitive than `SingleCellExperiment` in terms of data access. Since this thesis was my first contact with scRNA-seq data, I chose to use a mix of the two packages: using `SingleCellExperiment` in exploratory data analysis and first pre-processing and filtering of data, and `Seurat` for more complex tasks, such as batch effect correction, clustering and some types of data visualisation.

Other `R` packages were used in this work, namely `limma` (for differential expression analysis), `slingshot` (for pseudotime inference analysis), `fgsea` (for gene set enrichment analysis) and `cTRAP` (for assessing drug re-purposing potential). These packages have several methods and algorithms associated that are worth mentioning in more detail in the following sections.

3.3 Dimensionality Reduction

Matrices of read counts obtained from scRNA-seq experiments have high dimensionality, as there are thousands of genes that can be detected (figure 3.1). Thus, dimensionality reduction techniques are fundamental for data visualisation. Amongst the most used in scRNA-seq data are Principal Component Analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE).

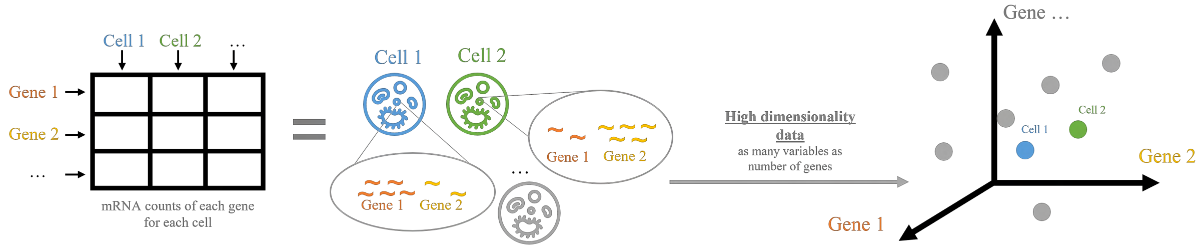


FIGURE 3.1: High dimensionality data representation
Schematic representation of single cell data dimensionality and the need for dimensionality reduction techniques. Each gene in a scRNA-seq count matrix corresponds to a variable.

3.3.1 Principal Component Analysis (PCA)

PCA is one of the oldest dimensionality reduction techniques [103] and is based on finding the directions of greatest variance in the data, which, therefore, contain most of the statistically relevant information. This problem is based on the discovery of pairs of eigenvectors / eigenvalues, that is, orthonormal vectors and their associated length, ordered by decreasing percentage of variance explained. Thus, data with high dimensionality can be projected in this new coordinate system, so that the projection reflects as much variance in the data as possible (figure 3.2).

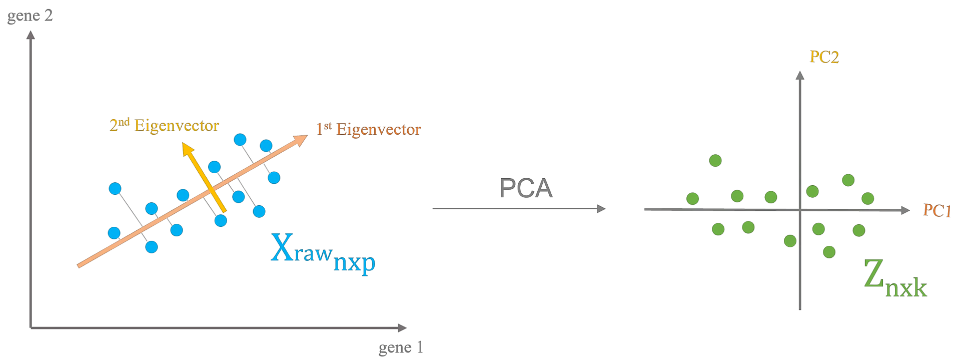


FIGURE 3.2: Principal Component Analysis (PCA) schematic representation
Schematic representation of performing a PCA on two dimension data, by projecting the data onto the main axes of variance. Each dot is a cell, and each variable is a gene.

Let X_{raw} be the high-dimensional data matrix, n the number of cells, p the number of genes (variables), k the number of new coordinates, μ the mean of the data and V the k eigenvectors with the highest eigenvalues. The centred data, that will ensure that the first principal component is proportional to the maximum variance of the data, may be given by:

$$X_{n \times p} = X_{raw_{n \times p}} - \mu \tag{3.1}$$

The new coordinates of the centred data, Z, will be given by the projection of the data in the space defined by the eigenvectors:

$$Z_{n \times k} = X_{n \times p} \times V_{p \times k} \tag{3.2}$$

In this work, I chose to use 50 main components and the `SCE` `runPCA` function to project the data in these 50 new dimensions, storing the results in an `SCE` object along with the data and metadata.

3.3.2 t-distributed stochastic neighbour embedding (t-SNE)

Unlike PCA, that is a deterministic and linear method, t-SNE is a probabilistic and non-linear method for visualising high dimensional data in n new dimensions [104]. The t-SNE n-dimensional plot is constructed such that, if two points/cells are close, they are most likely (given by a probability distribution) to be close in the real high dimensional space as well. This technique usually forms clusters of similar points but the absolute distance between two points, as well as their absolute position in space and cluster size, is arbitrary – only the relative distances between points hold some biological meaning. This means that this technique is highly prone to distortions in the data, can possibly find patterns in random noise if the parameters are chosen wrongly (unlike PCA that does not have *a priori* parameters), and is non-deterministic. However, the fact that it does not assume a normal distribution of the data, like PCA, makes it very useful for scRNA-seq data [105].

In this work, the `runTSNE` function of the `SCE` package was used to perform t-SNE and save the new coordinates for visualisation in the `SCE` object, together with the data and metadata. This function also takes a parameter called `perplexity`, which balances the attention between local and global similarities in data, forming clusters, being this set to 30 (the default).

3.4 Statistical Analysis

Statistical analysis allows to investigate trends, patterns, and relationships in quantitative data (such as count matrices), being the basis of scientific reasoning. And because every answer has a question, to draw valid conclusions, we must be very careful and precise when designing an experiment to address our research question. Not all scientific questions can be answered, but those that can must have an associated testable hypothesis (even if it then does lead to inconclusive results). Typically, these hypotheses translate into null (no relationship or no difference between groups) and alternative (there is relationship or difference between groups) hypotheses, which can be tested using representative data samples for the problem at hand, and appropriate statistical tests [106].

Statistical tests

There are a huge number of statistical tests, designed to test the null hypothesis. We can have tests applicable to cause-effect relationships (Pearson's correlation test), tests that explore the relationship between variables (t-test, ANOVA, etc.), and even tests that allow us to infer characteristics of a given population through a sample of it (linear regressions). These tests can be parametric, that is, they can be defined through a set of parameters such as the mean and standard deviation, thus making assumptions about the data (it follows a normal distribution, etc.), or non-parametric, that do not make assumptions about the data. Generally, for a parametric test (t-test, Pearson's correlation test) there is a non-parametric equivalent (Wilcoxon Rank-Sum test, Spearman's test), so the statistical test appropriate to the data and suitable to answer the scientific question should be carefully chosen.

Statistical significance

Regardless of their purpose, all statistic tests have an associated test statistic. This test statistic describes how far the data are from the null hypothesis (for example, that there is no difference between two groups, which can be translated into as the difference between the means of each group being null). Each test statistic will have an associated statistical significance, that is, a number that denotes how likely it is that the data would have occurred by random chance under the null hypothesis. The statistical significance is normally given by the p-value. P-value $<$ 0.05 is a common threshold for statistical significance, that is, if the null hypothesis is true, the data under question is likely to occur less than 5% of the time [107, 108]. We can reject the null hypothesis, but we can not deem it as true when it is not rejected.

Effect size

The effect size should always be considered when performing a statistical analysis. The effect size is the actual meaningful difference between groups being tested under the null hypothesis. Let's say that my null hypothesis is that the mean expression of gene k in cluster of cells A is equal to that in cluster of cells B; my null hypothesis is then $\mu_A - \mu_B = 0$. Let's say that I statistically test if my null hypothesis can be rejected, with a resulting p-value of 0.01 for a log fold change ($log_2FC = log_2(A/B)$) of 0.007. This means that the null hypothesis can be confidently rejected because there is only a 1% chance of obtaining an at least as extreme result under the null hypothesis. However, the average difference in expression between the groups (effect size) is quite small and therefore likely not biologically meaningful. We can then conclude that the tested is not relevant, although it is statistically significant [107, 109].

3.5 Pre-processing of scRNA-seq data

Although it can be considered as part of the results, taking around two months to be completed, I chose to include the pre-processing of single-cell data as a part of the Methods section. Albeit not directly answering any biological question, it has influenced all results and their interpretation.

“Getting intimate” with the data is essential in research projects; there is no set of packages, functions and algorithms that works universally well for all data. However, a pipeline of general steps for single cell data pre-processing can be defined, in order to guide researchers in the initial contact with this analysis. Such a pipeline is exemplified in figure 3.3 and will be explored in the following sections.

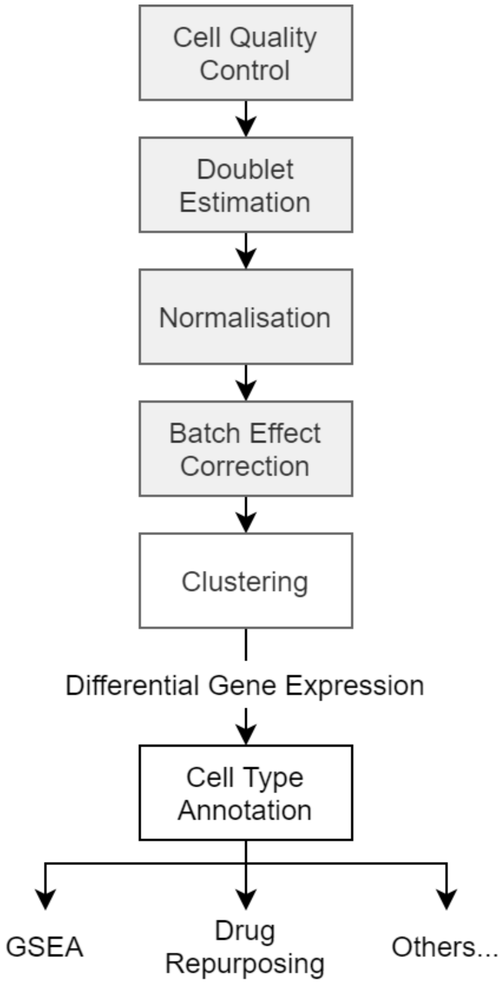


FIGURE 3.3: Common workflow for scRNA-seq data processing
Schematic of the processing pipeline used in this work. The pre-processing steps are identified in grey, and were applied to the raw count matrix retrieved from public databases of scRNA-seq data, comprising cell quality control (QC), normalisation, batch effect correction and clustering. Based on the main steps presented in the resources of the online course “Analysis of single cell RNA-seq data” (2019) [110].

3.5.1 Filtering

The first step in the analysis of single cell data is filtering based on cell quality and gene expression, as there may be several technical aspects that make the sequenced cells of poor quality. If this is not done, poor quality cells that do not reflect the molecular phenotype of any human brain cell will be included in the analysis, which could lead to wrong conclusions. In addition, filtering non-informative cells and genes is very important to optimize the computational burden of analysis.

There are several ways to identify a poor-quality cell [111, 112]:

- **Library size:** The library size of a cell can be described as its total number of sequencing reads. Cells with small library size may correspond to wells/droplets that captured ruptured cells/nuclei, or just background noise. This could also happen due to inefficiency in capturing and amplifying cDNA.

- **Detected genes:** Cells with few identified genes will have low library complexity, suggesting that the real population of transcripts was not correctly sampled. A characteristic of single cell data is the existence of a "heavy left tail" at the extreme of the distribution of identified gene counts (figure 3.4 (C)).

- **Percentage of mitochondrial genes:** High percentages of mitochondrial genes amongst those detected are indicative of low quality cells, as the remaining mRNA may have been lost due to cell lysis or RNA degradation. In single nucleus, this high percentage could also be indicative of cells where the cytoplasm was not efficiently removed.

It is also important to filter genes whose expression is considered undetectable, as they will not add relevant new information to the analysis.

The thresholds chosen for filtering are highly dependent on the data themselves, and therefore there are no recommended values. In general, thresholds should be chosen such that, in addition to acting on the heavy left tail of the UMI counts, balance the library size around the median of counts, as well as reducing the computational cost to an affordable burden. Considering these points, the following thresholds were chosen for this analysis, illustrated in figure 3.4:

- **Library size:** Remove cells with less than 400 read counts.

- **Unique genes detected:** Remove cells with less than 300 unique genes detected.

- **Percentage of mitochondrial genes:** Remove cells with more than 15% of the reads mapped to mitochondrial (MT) genes.

- **Gene counts:** Keep genes with a minimum expression of 5 read counts in at least 10 cells.

3.5.2 Doublets

Another thing to consider when analysing single cell data is the presence of doublets. Doublets occur when an oil droplet in the cell sorting protocol encapsulates two or more cells. Despite being a technical artefact, the estimation and removal of doublets can be done computationally, for example through machine learning techniques.

The identification and removal of doublets in this work was done using the `scDblFinder` [113] `R` package. According to a benchmark study, `DoubletFinder` [114] is the best tool for finding doublets in

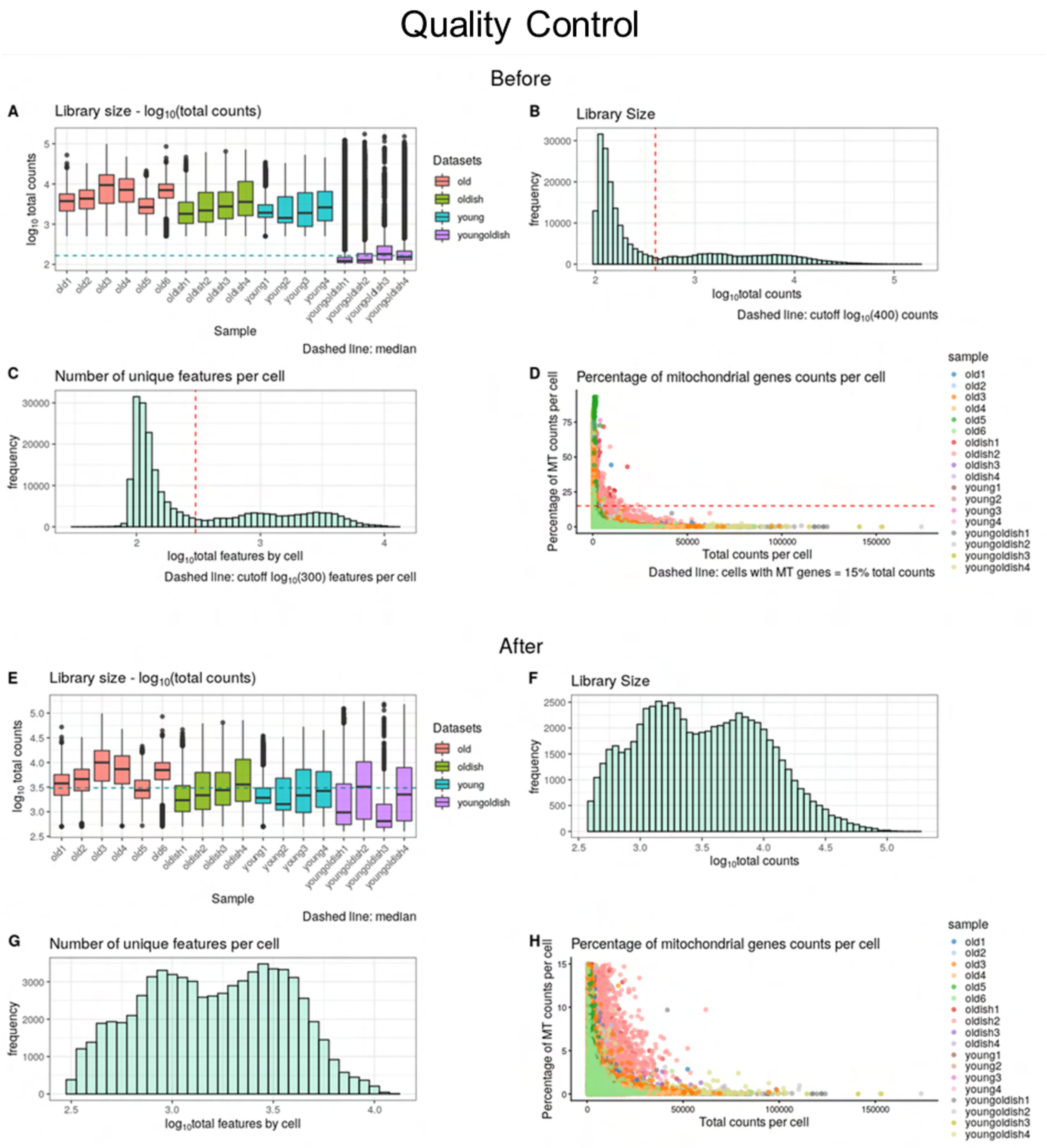


FIGURE 3.4: Quality control of scRNA-seq data - Filtering
The filtering thresholds were chosen considering the ”heavy left tails” in the counts of **(B)** reads and **(C)** unique genes identified, as well as **(D)** a reasonable percentage of mitochondrial genes. It an be also noticed that **(A)** the median of the library size was not around the median of the individual samples, and that **(E)** after filtering it is more balanced around them. The filtering results are shown in figures **(F)**, **(G)** and **(H)** for library size, unique genes per cell and percentage of mitochondrial gene counts per cell, respectively.

terms of accuracy, but not in computational efficiency [115]. Its main principle is that the algorithm first generates artificial doublets by combining gene expression profiles of two randomly selected droplets; subsequently, it defines a doublet score for each original droplet as the level of similarity the droplet has to those artificial doublets (via k-nearest neighbours); finally, it detects doublets as the original droplets whose doublet scores exceed the computed threshold. `scDblFinder`, published after the release of the benchmarking study, uses the same principles as `DoubletFinder`, but with a few alterations that make it more efficient and accurate. This function allowed for doublet identification in each sample,using the 1000 most variable genes to make the estimation, hence the importance of prior gene filtering.

Since this technique depends a lot on the number of cells used (the more cells, the more accurate the estimate), the estimation of doublets was done in two ways: with the unfiltered data and subsequent identification of doublets in the filtered results, and with the filtered data (figure 3.5). The estimate with the unfiltered data identified more doublets, totalling 10199 among the 68028 cells, in contrast to the estimate using only the filtered data, which identified 4269 doublets. Nevertheless, since the rationale for data filtering was that many of the cells were of poor-quality and did not reflect any real biological individuality, the estimate with the filtered data was chosen, removing 4269 doublets from the data.

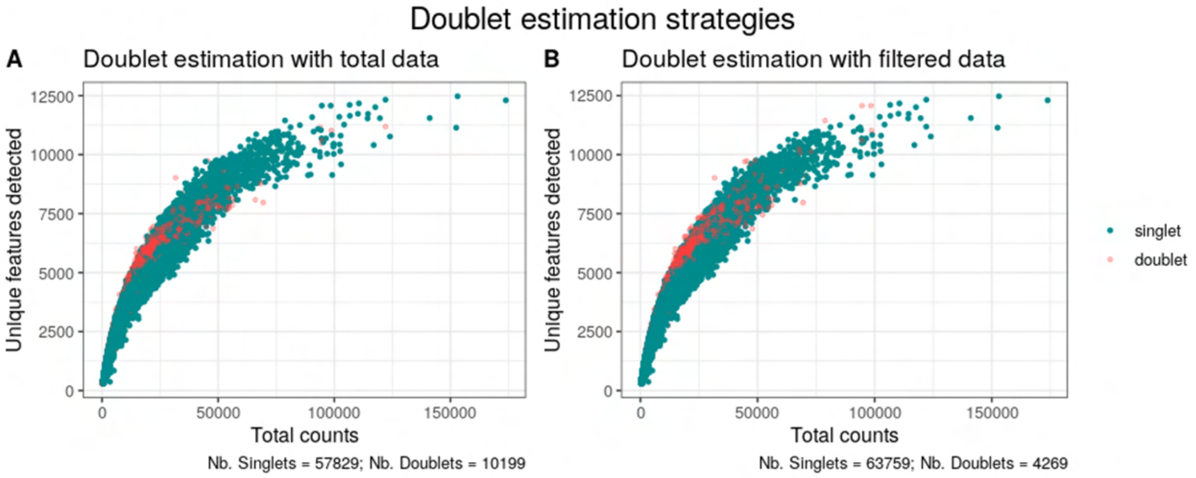


FIGURE 3.5: Doublet estimation strategies
Doublet identification strategies, using the scDblFinder R package. **(A)** estimation of doublets using unfiltered data, with consequent identification in the filtered data; **(B)** estimation of doublets using filtered data.

3.5.3 Normalisation

Normalisation is indispensable for analysing scRNA-seq data. When preparing the library for each cell, there are biological and technical factors that influence the associated library size. Since gene expression should not be confounded with cell sequencing depth, normalisation is important to eliminate technical variability while maintaining biological variability [116].

In this work, the `computeSumFactors` function (`scran` package [117] version 1.20.1) was used, which implements a deconvolution strategy for normalisation. Briefly, all cells are sorted by increasing library size, and a moving window is applied to this rank of cells, so that each window has cells with similar library sizes; then, the counts for those cells are summed together; furthermore, the count sums for this pool of cells are normalised against the average of the counts across all cells (average reference pseudo-cell); finally, sliding this window and performing these steps iteratively will construct a linear system that can be solved by least-squares methods to obtain cell-specific size factors, that can be applied to each cell (that is, gene expression of each cell is divided by its size factor). The default window sizes were used, being these around 20 cells for low library sizes, and with window size of around 100 for high library sizes. The results before and after normalisation are represented in figure 3.6.

Distribution of read counts usually have a log normal shape, and so the read counts were log-transformed with log_2. As the distribution of read counts for single-cell data is zero-inflated, a pseudocount of 1 was added in the transformation. The effect of normalisation on the data can be seen in figure 3.6, with the corrected library sizes becoming much more comparable between samples and datasets, mitigating sequencing depth-related biases.

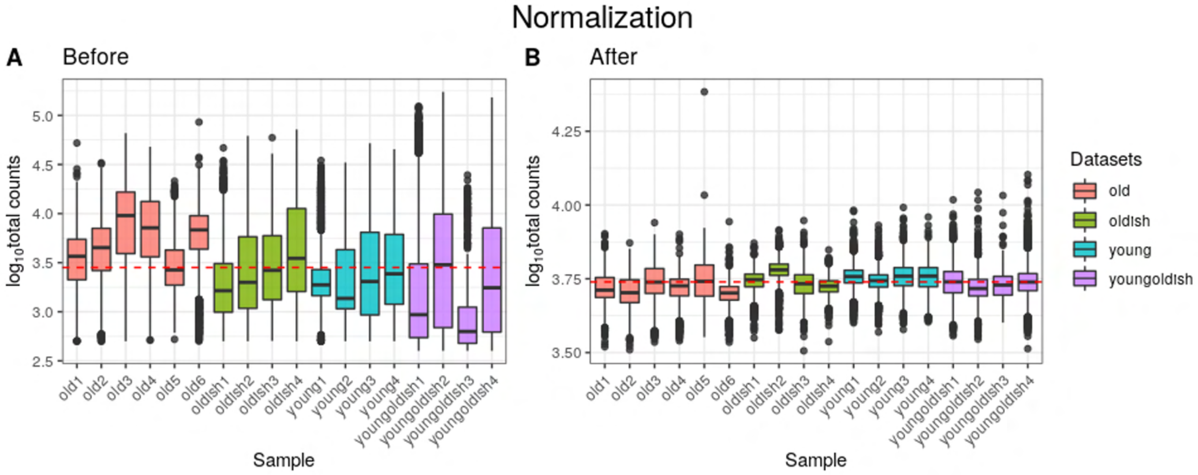


FIGURE 3.6: Normalisation of single-cell RNA-seq data
Boxplots summarising library sizes across cells for each sample in each dataset, **(A)** before and **(B)** after normalisation. The median counts across the entire data are represented by the red dashed line, and it can be clearly seen that, with normalisation, the library sizes are much more aligned with the median library size, mitigating biases associated with the sequencing depth.

3.5.4 Batch Effect Correction

Batch effects (differences in personnel, experimental conditions, etc.) cause variability in the data that is not associated with biological factors. This can lead to wrong conclusions if the variables of interest are correlated with the conditions of the experiment (e.g., clustering by dataset). Batch effects are of particular concern in scRNA-seq experiments (high resolution and noisy, due to low amounts of mRNA to work with) and, therefore, numerous computational tools have been developed to correct them.

Several batch effect correction methods were explored in this work, including `ComBat` (`sva` package [118] version 3.40.0) and limma's [119] function `removeBatchEffect`. However, `Seurat`'s batch effect correction method [120] showed the best results in terms of sample homogenization across all clusters. This algorithm was bench-marked as one of the best tools for batch effect correction, according to Tran *et al.* (2020) [121].

It identifies correspondences between cells of the same type in different experiments, called anchors, that can be used to harmonize datasets into a single reference, even when there are heavy technical differences. It has some similarities with reference assembly and mapping for genomic DNA sequences [120]. In summary, this algorithm performs, for each pair of datasets, canonical correlation analysis (CCA, similar to PCA), followed by L2 regularization, and mapping of both datasets in this new low dimensional space. Then, in this common space, pairs of mutual nearest neighbours (MNNs) are found between the two datasets, being considered as belonging to similar biological states – these are anchors between the datasets. Then, the anchors are given a score, based on the shared overlap of mutual neighbourhoods for the two cells in a pair - if this score is high, it means that many similar cells in one dataset are predicted to correspond to the same group of similar cells in a second dataset (figure 3.7). Finally, a non-linear transformation of the data based on these anchors and corresponding scores is performed, so that they can be jointly analysed. For this to work, correspondences of cell types between datasets are expected.

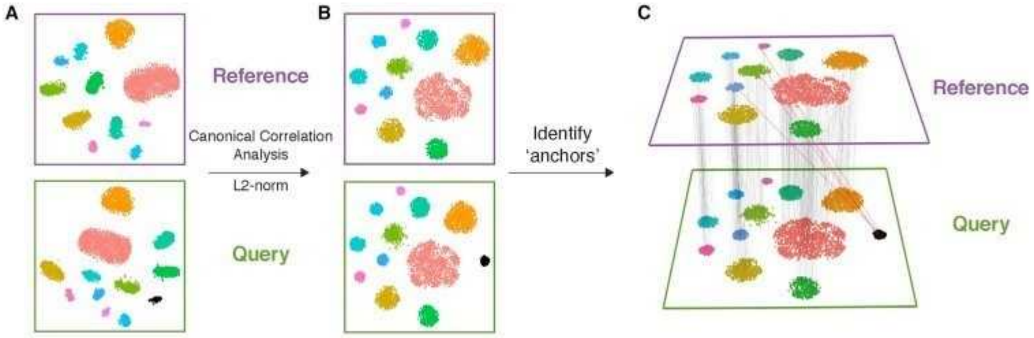


FIGURE 3.7: Seurat's batch effect correction algorithm
Schematic illustration of Seurat's integration algorithm for batch effect correction on single cell data. **(A)** For each pair of datasets, one is considered the reference, and the other the query. **(B)** CCA + L2 regularization is performed to project each dataset into a shared low dimensional space, where **(C)** anchors (pairs between cells) between datasets are computed. Non-linear transformations are applied, using these anchors, in order to "join" the datasets based on their biological similarities [120].

The functions used in this work were `FindIntegrationAnchors`, which identifies anchors between datasets in a list, and `IntegrateData`, which performs the non-linear integration of datasets. Using the four datasets (young, oldish, youngoldish, and old) as the input of the function, the algorithm was unable to integrate the data. Because of this, each sample was considered a dataset and the integration was performed with the 18 samples. The parameters used were the defaults of each function, including the use of 2000 highly variable genes as the number of features used for finding anchors, in the low-dimensional data with the first 30 main components. The batch effect with 4000 highly variable genes was also tested, however the result was roughly the same, which did not justify the higher computational cost and volume of the resulting data. The results of the batch effect correction are shown in figure 3.8. Before correction, cells were grouped by dataset,demonstrating a non-biological aggregation of data. After correction for batch effect, we can notice that the datasets and samples are now completely mixed, so the formed clusters should now be indicative of cell populations or interesting biological states that could be later explored.

Note that correction for batch effect may introduce technical variability and will not be equal to uncorrected normalized data. This correction should only be used for visualization and clustering purposes, as was done in this work, and not to perform differential gene expression, GSEA or similar [122].

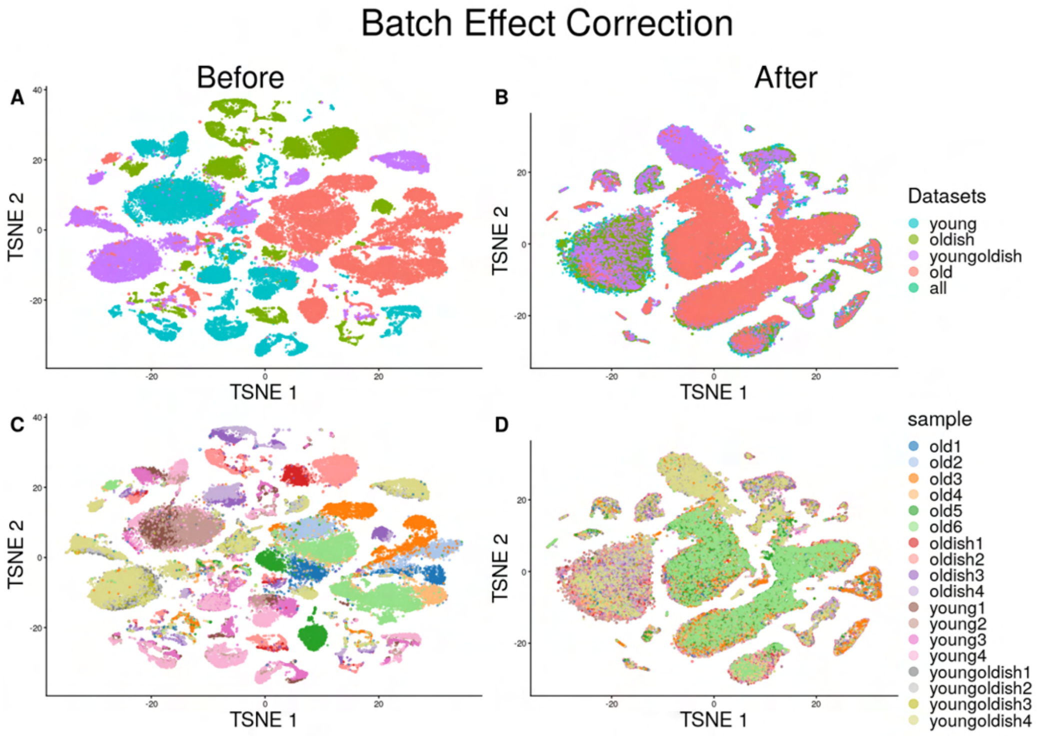


FIGURE 3.8: Batch effect correction
Single cell data of this work before **(A,C)** and after **(B,D)** batch effect correction. Before batch effect correction, cells were grouping by **(A)** dataset and **(C)** sample. After batch effect correction, the **(B)** datasets and **(D)** samples seem now more harmonized and integrated, clustering by possible cell types and cell states.

3.6 Clustering

One of the most important tasks in analysing single cell data analysis is the definition of clusters of cells (e.g., after t-SNE representation) and the following assignment to cell types, based on the expression of specific markers of each cell type. The definition of a cluster and the consequent assignment to a cell type is a very complicated task, so as not to incur in under-clustering (i.e., when cells of different types are assigned to the same cluster, masking the underlying biological structure of the data) or over-clustering (i.e., when if multiple clusters represent the same cell type, introducing non-relevant divisions in the data).

Concomitant with the growing interest in the area, numerous clustering algorithms have been developed in recent years. The one used in this work is modularity optimization [123], or Louvain's method. This is a heuristic method applicable to large networks and the clustering algorithm used by `Seurat`, having demonstrated its effectiveness in benchmarking studies [124]. This method consists of the so-called modularity optimization phase and the community aggregation phase. Modularity is a parameter between -1 and 1 that translates the density of links in each community of points, compared to links between communities. For each node, its neighbours are considered, and the modularity gain that would occur if a node were to belong to a community is calculated. This process is iteratively done for all nodes, until there is no gain in modularity (a local maximum). The community aggregation phase creates a new network, whose nodes are the communities found in the first step. These two steps are thus repeated iteratively until a stopping criterion, called resolution, is met.

In order to avoid under-clustering, an “educated clustering” strategy was allied to this algorithm (i.e., through a qualitative analysis of cell type markers [125]), allowing to have an initial idea of the cell types associated with each cluster (figure 3.9). In figure 3.10 (A) to (C) are depicted the resulting clusters when using different resolutions (0.01, 0.35 and 0.4). For this data, a resolution of 0.01 results in under-clustering, given that the qualitative evaluation of the clusters identifies microglia and oligodendrocytes appearing together in cluster 0; with a resolution of 0.35, cluster 4 could be merging a cluster not qualitatively identified as any cell type with a cluster that could theoretically be neuronal. Still in the same line of reasoning, with a resolution of 0.4 we can divide cluster 4 (resolution 0.35) into two new clusters (12 and 5), separating a putative neuronal cluster from a cluster whose cell type cannot be determined.

The clustering algorithm was applied using the `Seurat` functions `FindNeighbors` and `FindClusters`. A resolution of 0.4 was chosen, given that with higher resolutions the division of clusters gets noisier, by the displacement of residual cell groups from one cluster to another. This decision was aided by the construction of a cluster tree, using the `clustree` function from the package `clustree` (figure 3.10 (D)).

3.7 Differential Gene Expression

To infer differences in gene expression between groups of cells, we can linearly model gene expression. The differential gene expression analysis (DEA) was performed in this work in multiple ways. First, by modelling gene expression and comparing one cluster against the average of the remaining clusters:

$$GE_x = Cluster_i \times \beta_i \tag{3.3}$$

Where GE_x is a vector of expression of gene x across cells, and $Cluster_i$ is a logical matrix with an entry of 1 if the cell belongs to cluster i, and 0 otherwise. Given that this matrix has as many columns as clusters and as many rows as cells, and each cell will be in only one cluster, the resulting matrix (design matrix) will be sparse. β_i will be the average expression of gene x in cluster i. A contrast matrix (i.e., a matrix representative of linear combination of the unknown coefficients β_i) was then used to get the differences between a β_i coefficient and the average of the remaining coefficients.

The second way was by comparing two clusters' gene expression. The formulation was equivalent to equation 3.3, but with further use of a contrast matrix comparing specific pairs of coefficients.

The third way of using the linear models on gene expression in this work was to compare the gene expression profile of each cluster against a baseline one:

$$GE_x = \beta_0 + Cluster_i \times \beta_i \tag{3.4}$$

Where GE_x is a vector of expression of gene x across cells, β_0 is the expression of the baseline cluster, and $Cluster_i$ is a logical matrix with an entry of 1 if the cell belongs to cluster $i \setminus baseline$, and 0

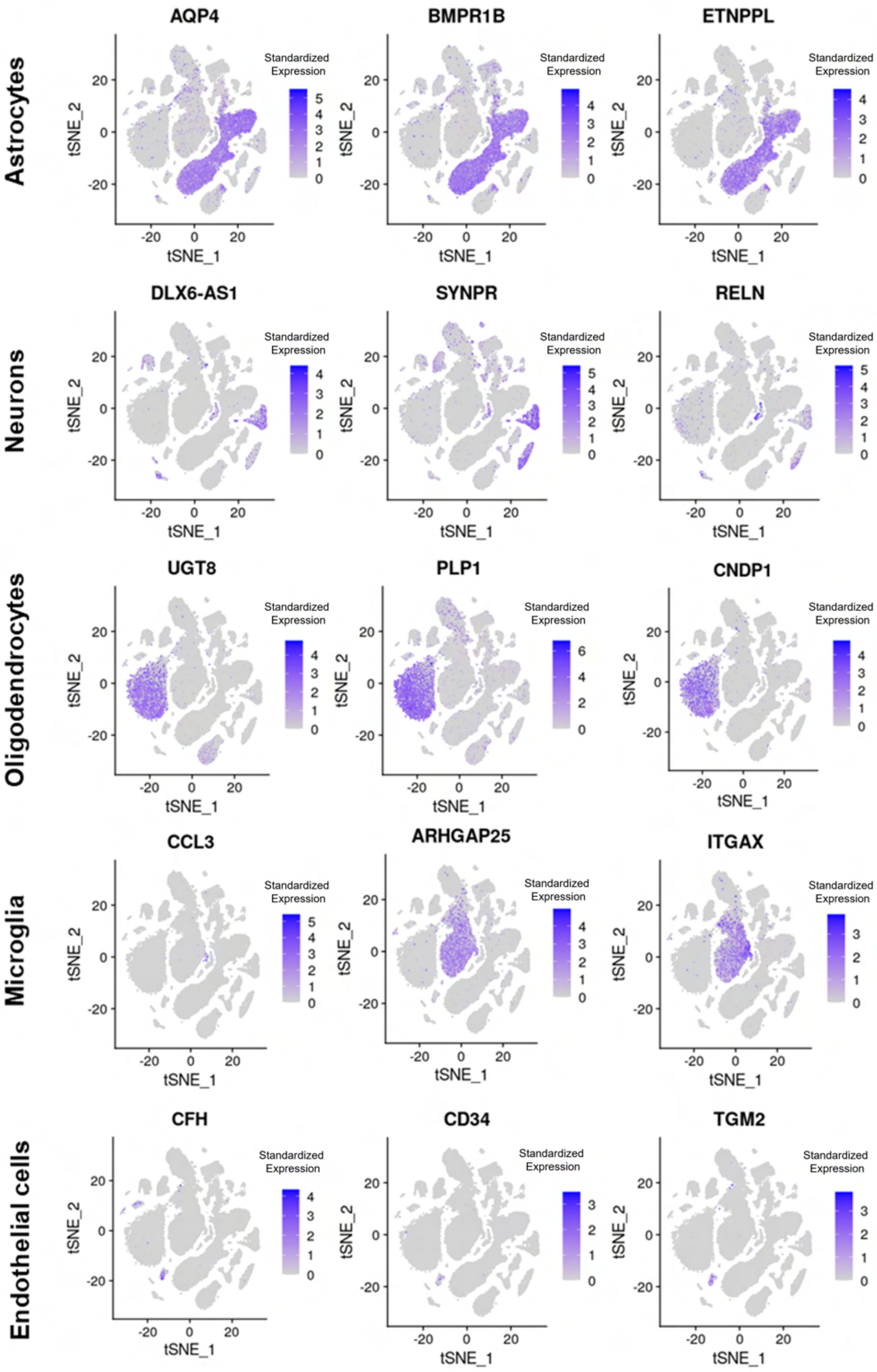


FIGURE 3.9: Marker genes of CNS cell types
t-SNE plots of the expression (log-normalised) of three marker genes of each of the main CNS cell types [125].

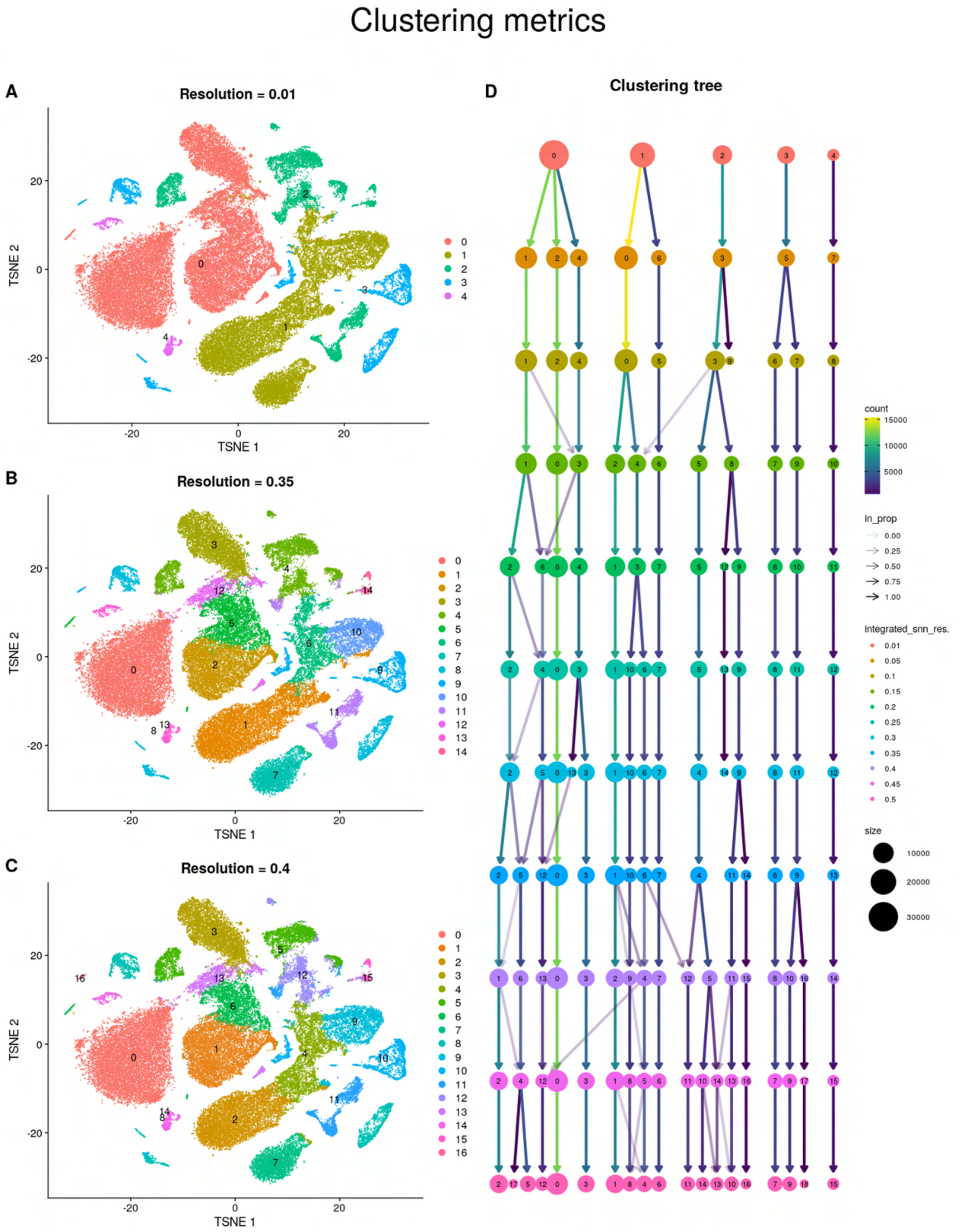


FIGURE 3.10: Clustering Metrics
Clustering results obtained by applying the Louvain method to our scRNA-seq data, with resolutions of **(A)** 0.01, **(B)** 0.35 and **(C)** 0.4. A resolution of 0.4 was chosen. **(D)** Cluster tree obtained by applying a range of resolutions between 0.01 and 0.5, with a step of 0.05, illustrating the relationship between clusters obtained with different resolutions.

otherwise. Under this formulation, the resulting β_i coefficients will be the log_2FC expression of gene x between each cluster and the baseline cluster.

DEA was performed using the `limma` [119] and `EdgeR` packages [126] and following the limma-voom pipeline [127], unless stated otherwise. The limma-voom pipeline was applied to the non-normalized filtered scRNA-seq data (using `edgeR` for normalization) and fit the data to a linear model, using then the moderated t-test (parametric) and empirical Bayes shrinkage of standard errors to assess the statistical significance of the differential expression results. `limma` is a *pseudobulk* method (i.e., aggregates cells within a biological replicate) for differential expression analysis, meaning that it can deal better with biological replicates than other state-of-the-art methods built exclusively for scRNA-sequencing, producing fewer false positives [128]. For each coefficient in the linear model, the magnitude of differences in gene expression was measured in log_2FC, and their significance given by an adjusted p-value lower than 0.05 for multiple comparisons (Benjamini-Hochberg correction (BH correction)), needed to control the false discovery rate arising from testing more than one hypothesis at a time.

3.8 Cell Type Annotation

After having the "ideal" number of clusters, it was possible to assign each one to a cell type. For this, the markers of each cluster, that is, differentially expressed genes of each cluster in comparison with the other clusters, were found.

There is still no consensus on the ideal statistical test to find the markers (i.e., the differentially expressed genes) for each cluster. However, if those are robust, their finding should be independent of the test used. Due to its relevance, the cell type annotation task was done applying a non-parametric test, using `Seurat`, as this approach does not make assumptions on the data. Yet, given the computational burden of it (around 5 hours to get the differentially expressed genes), `limma`'s workflow was followed in the remaining analyses.

The `FindAllMarkers` function from `Seurat` finds the markers for each cell cluster against the remaining clusters, using the non-parametric Wilcoxon Rank-Sum Test. A significance level of adjusted p-value $<$ 0.05 (BH correction) and a magnitude of the difference in expression between clusters of log_2FC above 0.25 were considered, for both the detailed and general analysis. All thresholds for significance and log_2FC were chosen with the help of volcano plots[4] and the usage of established marker genes (positive controls, figure B.1).

The cell type annotation task was divided into two main steps: the **detailed analysis**, where each of the clusters identified in figure 3.10 (C) was associated with a cell type; and the **general analysis**, where if one cell type was associated with more than one cluster, these smaller clusters were grouped into larger ones. These steps will be explained in detail in the next sections.

[4] Plots of magnitude of effect against the statistical significance, in logarithmic scale, of each gene after DEA, that resembles a volcano

3.8.1 Detailed Analysis

To associate each cluster in figure 3.10 (C) to a cell type, we found the percentage of differentially expressed genes between that cluster and the others that are known markers of a cell type, associating a cluster *C* to a cell type *A* if the following criteria were met:

1. More than 35% of cluster *C* marker genes being also markers for a cell type *A*.

2. Less than 8% of cluster *C* marker genes being also markers for each cell type $\neq$ *A*.

3. The percentage of cluster *C*'s markers not associated to any cell type being less than 60%

These criteria were chosen empirically (table B.1). Particularly for the first threshold, if this value was too strict, some cell types known to be present in the brain would be missing, such as endothelial cells (figure 3.11 (A)) and microglia, and given that, by looking at panels (B) and (D) from figure 3.8, the samples are homogeneously distributed by all clusters, it is plausible that the cluster formation will reflect the main cell types of the CNS. The second threshold was chosen to avoid an overly permissive classification that would lead to an almost random distribution of cell-type markers (figure 3.11 (B)). Finally, the third criterion, framed by the first two, was chosen to control for random technical clusters. If these three criteria are met, the clusters will be assigned with one of the cell types; otherwise, they will be assigned to "unknown cell types".

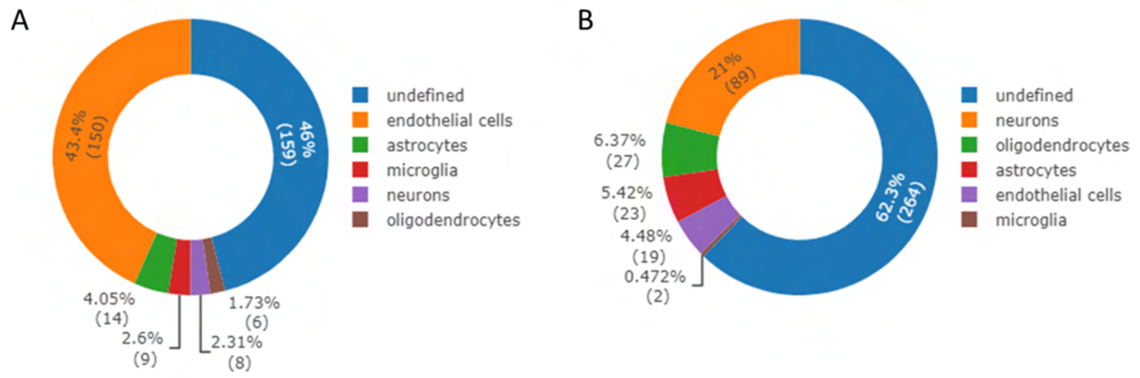


FIGURE 3.11: Distribution of know cell-type specificities of marker genes for clusters 14 and 7 in figure 3.10 (C)
The percentage of marker genes of clusters (A) 14 and (B) 7 from figure 3.10 (C) that are known to be specific markers of each neural cell type is shown.

Finally, for each of the assignments, the top 10 marker genes of each cluster (significant genes with the highest log_2FC) were always checked. This was done in order to confirm that there was agreement between the top genes for each cluster and the classification obtained in the detailed analysis step, increasing confidence in the results. All clusters associated with a cell type had five or more of the top 10 marker genes associated with that cell type.

3.8.2 General Analysis

After associating each cluster with a cell type (figure 3.12 (A)), 6 large clusters were formed, by joining together clusters of the same cell type (and one large cluster for undefined cells), as shown in figure 3.12 (B). After this, the markers of each of these clusters were again obtained and the conditions described above were verified (table B.2). With this, I managed to associate each cluster to a CNS cell type, through an “educated” approach of clustering and cell type annotation, which allows to discriminate the cells of interest and start asking the biological questions I am interested in.

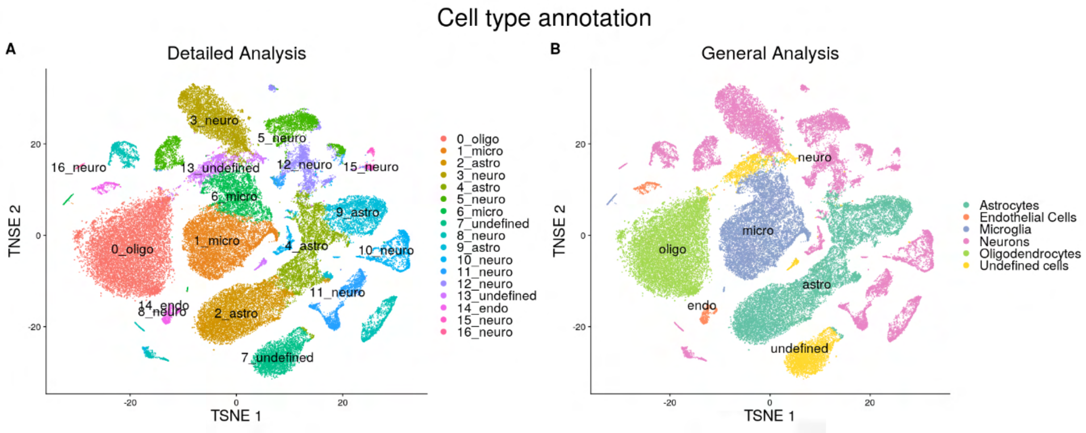


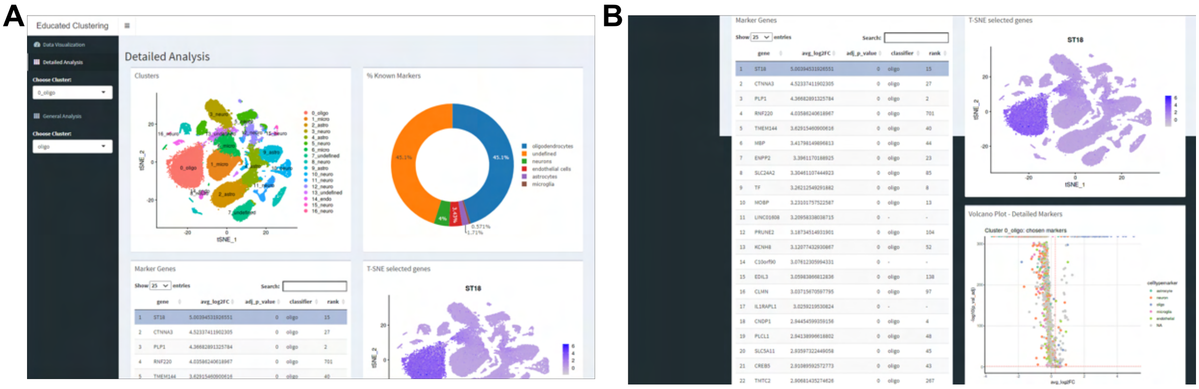
FIGURE 3.12: Cell type annotation of clusters
Annotations of cell types based on **(A)** the detailed analysis and **(B)** the general analysis. In the detailed analysis, each of the clusters obtained by the clustering algorithm with a resolution of 0.04 was associated with a cell type, based on specific markers of each cell type. In the general analysis, these clusters were grouped into 6 large clusters, which allows the isolation of the specific cell type of interest – astrocytes – for subsequent analyses.

The number of cells in each of the 6 clusters can be found in table 3.3, where we see a higher number of neurons, followed by astrocytes, oligodendrocytes, microglia, undefined cells, and finally endothelial cells. We cannot conclude much from these numbers, since we know that, for instance, the ”old” dataset only had astrocytes and microglia, therefore not following the expected 1:1 ratio between neurons and glial cells. Furthermore, cell proportions suggested by scRNA-seq should be handled with care, as they may be technically biased (e.g., cell quality dependent on cell type). Nevertheless, neurons are still the most abundant cell type in the combination of all datasets.

TABLE 3.3: Summary of the number of cells per cluster, obtained at the end of the general analysis of the “educated clustering” workflow.

Cluster	Number of Cells
neuro	20303
astro	14798
oligo	12327
micro	11279
endo	913
undefined	5242
total	64862

For this educated clustering process, since there were many clusters to be analysed, and many variables to take into account, I chose to develop a small decision support dashboard. This dashboard was made in R, using the `shinydashboard` [129] package. It is a very basic dashboard, with just a summary of all the metrics I talked about (figure 3.13 (A)), with the possibility of seeing in detail the expression of each marker of a certain cell type, in any cluster, coloured in a tSNE, as well as the visualisation of the markers in a volcano plot (figure 3.13 (B)). However, in pedagogical terms, it was another tool that I learned to use and included in my “toolbox”.



Marker Genes

Show 25 entries Search:

	gene	avg_log2FC	adj_p_value	classifier	rank
1	ST18	5.00394531926551	0	oligo	15
2	CTNNA3	4.52337411902305	0	oligo	27
3	PLP1	4.36682891325784	0	oligo	2
4	RNF220	4.03586240618967	0	oligo	701
5	TMEM144	3.62915460900616	0	oligo	40

Marker Genes

Show 25 entries Search:

	gene	avg_log2FC	adj_p_value	classifier	rank
1	ST18	5.00394531926551	0	oligo	15
2	CTNNA3	4.52337411902305	0	oligo	27
3	PLP1	4.36682891325784	0	oligo	2
4	RNF220	4.03586240618967	0	oligo	701
5	TMEM144	3.62915460900616	0	oligo	40
6	MBP	3.41798149896813	0	oligo	44
7	ENPP2	3.3961170188925	0	oligo	23
8	SLC24A2	3.30461107444923	0	oligo	85
9	TF	3.26212549291882	0	oligo	8
10	MOBP	3.23101757522587	0	oligo	13
11	LINC01608	3.20958338038715	0	-	-
12	PRUNE2	3.18734514931901	0	oligo	104
13	KCNH8	3.12077432930867	0	oligo	52
14	C10orf90	3.07612305994331	0	-	-
15	EDIL3	3.05983866812836	0	oligo	138
16	CLMN	3.03715670597795	0	oligo	97
17	IL1RAPL1	3.0259219530824	0	-	-
18	CNDP1	2.94454599359156	0	oligo	4
19	PLCL1	2.94138996618802	0	oligo	48
20	SLC5A11	2.93597322449058	0	oligo	45
21	CREB5	2.91089592572773	0	oligo	43
22	TMTC2	2.90681435274626	0	oligo	267

FIGURE 3.13: Dashboard for cell type annotation
Dashboard made, through the R shinydashboard package, with the purpose of helping the annotation of cell types for each cluster. As an example, we show the information available for the a cluster of oligodendrocytes (0_oligo), where it is possible to verify **(A)** the percentage of specific gene markers of each cell type present in the markers of this cluster, as well as **(B)** visualise a t-SNE plot coloured by the expression of a particular oligodendrocyte marker, and the volcano plot of differential gene expression associated with the markers of this cluster.

3.8.3 Astrocyte Isolation

With the educated clustering analysis and workflow, it is possible to select, with statistical and biological confidence, using robust genetic markers, the set of astrocytes for the remaining analysis. One could have directly used the already processed gene expression data for the astrocytes obtained in the educated clustering step. However, I chose to perform all steps on raw astrocytic data, including filtering, normalisation, batch effect correction, and clustering, as to avoid possible bias caused by the non-astrocytic data. Two samples had to be removed from the analysis, youngoldish3 and oldish4 (table A.2), because they had a very low number of astrocytes (31 and 99, respectively), which made the remaining analysis unpractical. Two of the pre-processing steps got special attention, as described below.

The normalisation did not go as well as what was expected from what happened previously with all CNS cells, where the median library size of each sample was around the median library size across all samples (figure 3.14 (C)). This could be explained by the reduction in the number of cells, which hampered the precision of the normalisation factors estimation. I hypothesised that the normalisation step could get better results if I used the gene expression of astrocytic cells directly from the cell type annotation step, and thus not running the pre-processing pipeline all over again. However, I tried both approaches and the results were pretty much the same (figure 3.14 (B) and (C)). That said, I chose to continue with the pipeline analysis from the raw astrocytic data.

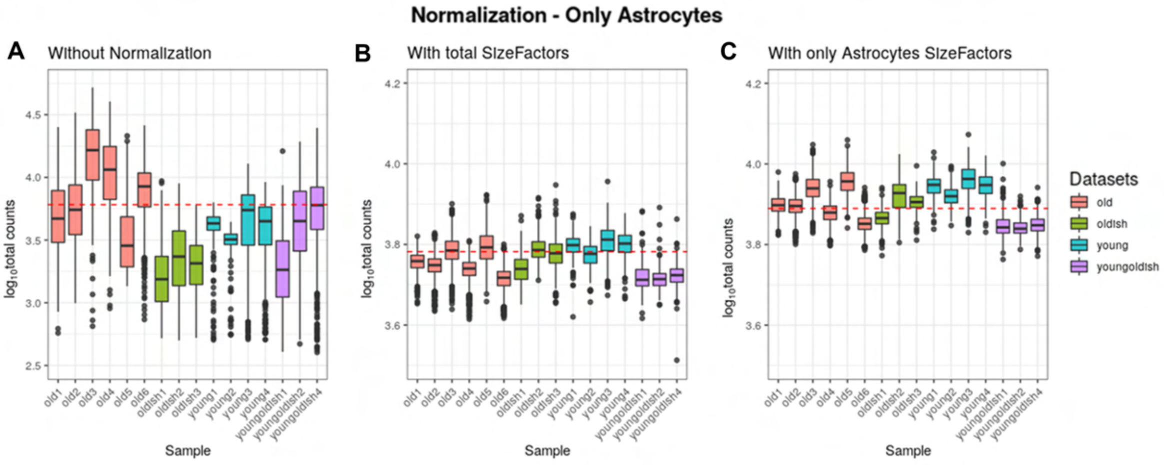


FIGURE 3.14: Normalisation of astrocytic data
Distribution of astrocytic library sizes across the various datasets, **(A)** before data normalisation; **(B)** after data normalisation using normalisation factors obtained from cells of all types; and **(C)** after data normalisation using astrocytic data normalisation factors, after redoing the pre-processing pipeline only in these cells. The red dashed line represents the median library size.

In the cell type annotation step (after clustering with a resolution of 0.3), since I only have one cell type under analysis, I chose to see the expression of a pair of markers specific for astrocytes, namely *AQP4* and *SLC1A2* [125]. In figure 3.15 (B) it can be seen that cluster 5 in figure 3.15 (A), coloured in pink, does not express any of these markers, which may indicate cells that were misidentified as astrocytes. That said, I chose to remove this cluster, and go back to performing the pre-processing pipeline again.

Finally, after applying the pre-processing pipeline only to these reliable astrocytic cells, the clusters shown in figure 3.15 (C) were obtained. However, cluster 1 had a large percentage of mitochondrial genes when compared to the other clusters (3.16 (A)), and the differentially expressed top genes were also mitochondrial genes. It was not clear why this was happening, but given that the expression of some mitochondrial genes had a bimodal shape (3.16 (B)), one hypothesis was that cells in poor condition could be grouped with cells that are actually biologically defined by higher expression of mitochondrial genes. To further divide those cell groups, looking for biologically interesting distinct astrocytic states, I decided to run clustering again on the cells of cluster 1. The chosen resolution was 0.31 and, after analysing the markers of each cluster, I concluded that one of the (four) clusters obtained with this subdivision did not have mitochondrial genes among its differentially expressed genes, and that the bimodal expression of mitochondrial genes was less pronounced. Thus, I associated the group of cells without mitochondrial marker genes in the new cluster 1, and the others in a new cluster 6.

The clusters shown in figure 3.17 (B) are those chosen for the subsequent analyses. Out of curiosity, I finish this chapter by mentioning that I have reduced my data by about 93% since the beginning of pre-processing, going from 209,187 CNS cells (nuclei) to about 13,694 astrocytes. These data will be the basis of the remaining analysis, given that they are already normalised, with the dimensionality reduction parameters (t-SNE and PCA) computed, and with interesting clusters to study.

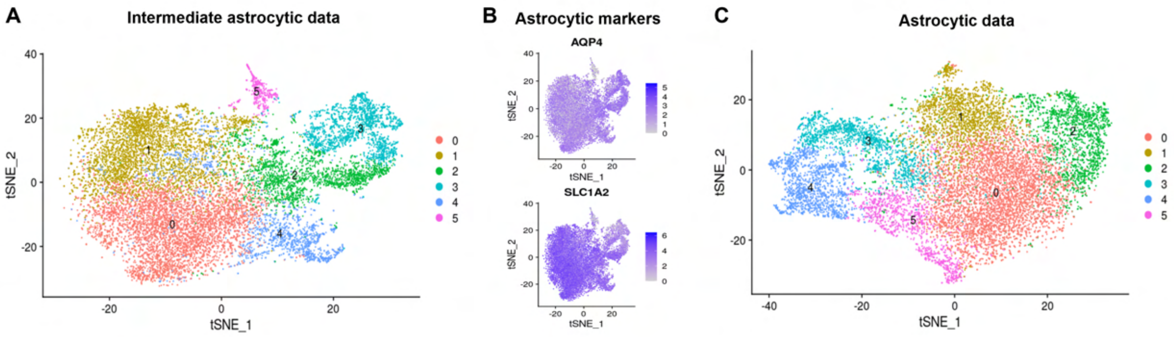


FIGURE 3.15: Final selection of astrocytic data
(A) Clustering of re-processed gene expression of cells previously classified as astrocytes based on the clustering of cells of all types; **(B)** t-SNE plots coloured by the expression of specific astrocyte markers, where it is verified that the cells of cluster 5 express less of them; **(C)** clustering of re-processed gene expression of astrocytes after the removal of cells in cluster 5 of **(A)**.

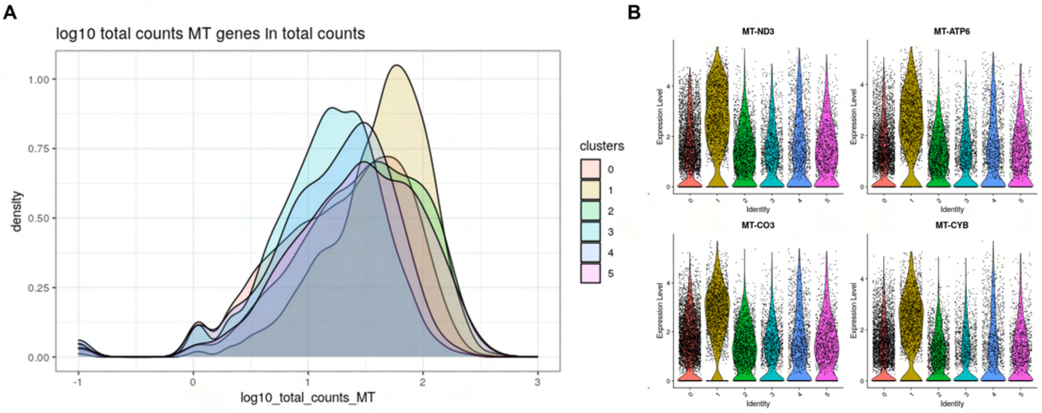


FIGURE 3.16: Mitochondrial genes in astrocytic clusters
(A) Smoothed histograms of distributions of log10 of read counts of mitochondrial genes in each cluster of astrocytes, with cluster 1 standing out. **(B)** Expression of some mitochondrial genes found to be cluster 1 markers, with evidence for their bi-modal form distribution across cluster 1 cells, which may indicate that this cluster houses two distinct cell groups.

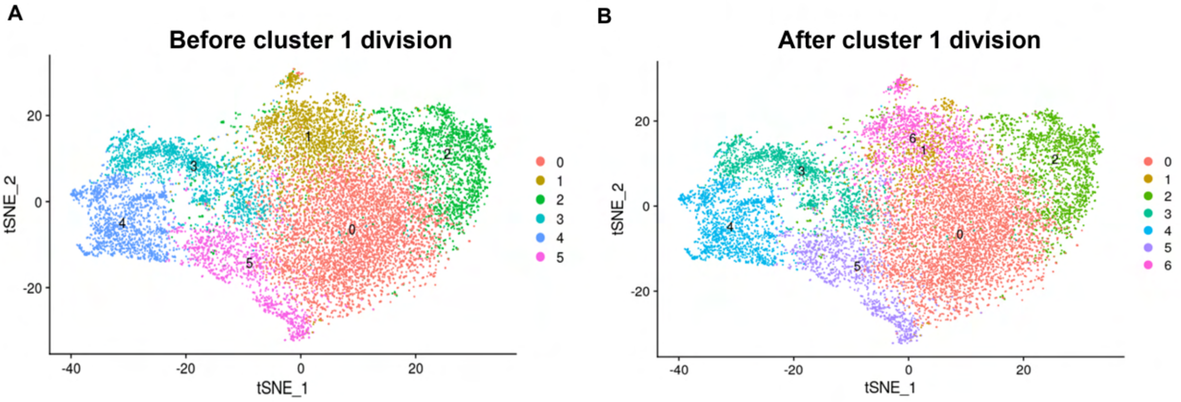


FIGURE 3.17: Sub-clustering of cluster 1 astrocytes
Clustering of astrocytes based on their gene expression data **(A)** before and **(B)** after the division of cluster 1 into two new clusters, 1 and 6. The new cluster 1 does not have mitochondrial marker genes, unlike cluster 6.

3.9 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a computational method used to infer whether a defined list of genes shows statistically significant and concordant differences between two biological states [130]. This tool looks at a list of genes ranked by a certain statistical metric, and if a gene belongs to the gene set of interest (pathway, biological process, etc.), it increases a running-sum statistic, and decreases it otherwise (figure 3.18). After going through the entire list of genes, an overall enrichment score (ES) will be given together with its statistical significance for each pathway / biological process included in the GSEA database [130], under the reasoning that if the gene set is enriched at either the top or bottom of that list (that is, under or over-expressed), it is thought to be related to phenotypic differences. GSEA uses a collection of publicly accessible annotated gene sets, divided into, among others, hallmark gene sets (representing specific well-defined biological states or processes), curated gene sets from online pathway databases (such as KEGG), and genes annotated by the same ontology term (GO).

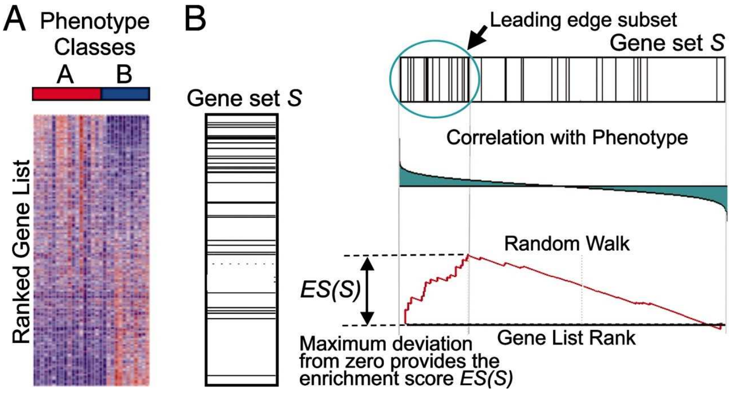


FIGURE 3.18: Gene Set Enrichment Analysis (GSEA) overview
Overview of the GSEA method, applied to two phenotype classes, A and B. **(A)** DEA is performed between two groups A and B, resulting in an ordered list of differentially expressed genes; **(B)** The location of genes from a gene set S within the sorted list of differentially expressed genes is obtained, on top of which GSEA will be performed, resulting in a maximum ES [130].

In this work, GSEA was performed using the `fgsea` package [131] to infer phenotypes or biological processes that underlie the biology of each astrocytic cluster. Upregulated or downregulated KEGG pathways, hallmarks and biological processes (GO) were inferred through the application of GSEA to each of the astrocytic clusters. The ranked lists of genes used were the differentially expressed genes in one cluster against the remaining, the differentially expressed genes of one cluster versus the baseline astrocytic cluster, and the loadings of each gene for each principal component. These genes were ordered by their t-statistic, that is, the ratio of the difference between the estimated value of a parameter and its hypothesized value to its standard error, combining effect size and statistical significance.

3.10 Pseudotime Inference

Pseudotime inference is a technique that associates each cell with a pseudotime, i.e., a measure of how much progress an individual cell has made through a certain process that can be translated by changes in gene expression [132]. It is normally applied to cell differentiation data, on top of a representation of dimensionality reduction, such as PCA or t-SNE. This technique is based on the idea that, for many systems, such as cell differentiation, there are not clear distinctions between cell states, but a smooth transition – a lineage -, where each cell is associated with a given pseudotime in that lineage (figure 3.19). There are several methods to calculate such pseudotimes, whose explanation is not in the scope of this master thesis.

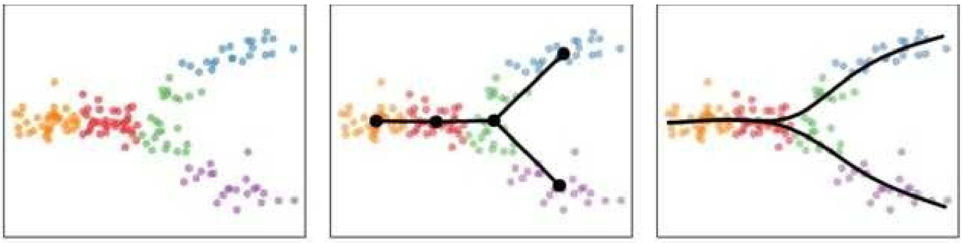


FIGURE 3.19: Slingshot for pseudotime inference
Representation of the application of a pseudotime inference algorithm, slingshot, on data represented in the PCA space. In this space, relationships between cell clusters are inferred through the assignment of pseudotimes, allowing a continuous lineage between cell states / cell clusters to be traced [133].

In methodological and logical terms, this tool was not developed with the aim of exploring the type of single-cell data used in this work. However, this was a key tool for clarifying the possible relationships between clusters, and thus ease the functional classification of each one. In this work, `R` package `slingshot` [133] was used to infer a progression between astrocytic clusters, starting from the baseline cluster with associated normal astrocytic functions. Although it is not expected to obtain differentiation trajectories with these data, it allowed to look in another way at the representative PCA of astrocytic data, and to associate each axis of variance with a biological process.

3.11 Drug Repurposing (cTRAP)

`cTRAP` [134] is a computational tool developed to compare differential gene expression results with the transcriptomic profiles from known cellular perturbations (such as drug administration), derived from the Connectivity Map (CMap) [135]. `cTRAP` can compare an ordered list of differentially expressed genes with known transcriptional alterations caused by gene knockdowns or compounds administration and find which perturbations are more correlated (positively or negatively) with the phenotype of interest (hereinafter referred to as *phenotype strategy*). Similarly, this tool can, from online databases of drug sensitivity, infer which drugs are the most likely to target cells expressing the distinctive/candidate genes of the phenotype of interest (hereinafter referred to as *top gene strategy*). These are two different but important concepts, that allow for an educated search for phenotype emulation in *vitro* / *in vivo*, and for drug repurposing. Given that most drugs will have therapeutical applications other than those they were originally described for and have been proven safe in the human body, the strategy of drug re-purposing has gained tremendous popularity, as it overcomes ethical issues and the expensive process of drug development and approval processes [136, 137]. In this context, drugs known to be able to cross the BBB in neuro-related works are particularly important.

3.12 Cell-type deconvolution (CIBERSORTx)

CIBERSORTx [138] is a machine learning technique that allows the performance of digital cytometry, that is, cell type deconvolution using gene signature matrices of each cell type and with application in bulk RNA-seq samples. Cell type deconvolution is a technique that allows estimating the proportions of different cell types in bulk samples, for further association with the sample's metadata. CIBERSORTx bases its algorithm on a linear support vector regression, firstly applied in its predecessor, CIBERSORT[139]. In short, to find these proportions (weights) for each cell type, the nu-support vector regression (ν-SVR) method, a variation of support vector machines (SVM), is first applied to gene expression data in order to perform feature (i.e. gene) selection, to minimise the possibility of over-fitting. The support vectors transform the total count matrix into a sparse solution that can be applied to a linear regression problem, depicted in figure 3.20, where m is the sparse mixture matrix of cell types (bulk RNA-seq) after feature selection, B is the single-cell signature matrix of each cell type, and f the unknown weights associated with each cell type, in each sample.

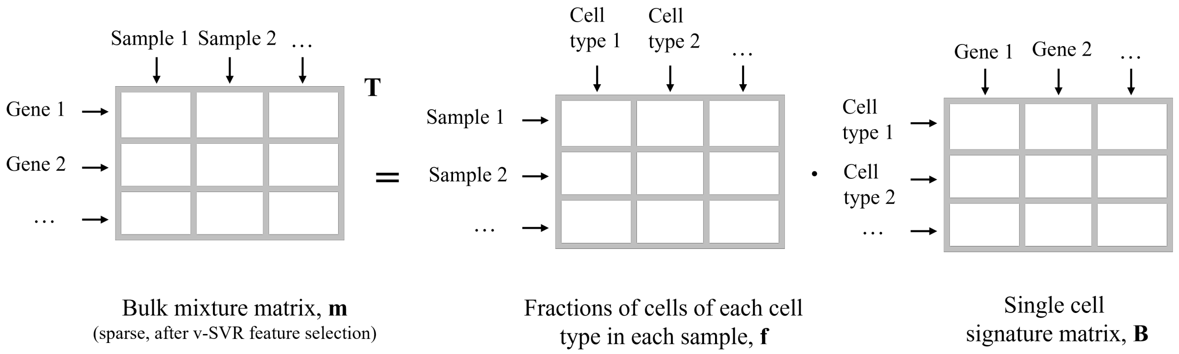

Sample 1
Sample 2
...
T
Gene 1
Gene 2
...
=
Cell type 1
Cell type 2
...
Sample 1
Sample 2
...
•
Gene 1
Gene 2
...
Cell type 1
Cell type 2
...
Bulk mixture matrix, **m**
(sparse, after v-SVR feature selection)
Fractions of cells of each cell type in each sample, **f**
Single cell signature matrix, **B**

FIGURE 3.20: CIBERSORTx system of equations for problem solving
System of equations associated with the solution of the cellular deconvolution problem, where m is the sparse mixture matrix of cell types (bulk RNA seq) after feature selection, B is the single cell signature matrix of each cell type, and f the unknown weights associated with each cell type, in each sample.

CIBERSORTx is implemented on a web platform [140]. By submitting a matrix of single cell counts, the platform builds a cell type signature matrix (B), and, by subsequent submission of the mixture matrices (bulk RNA seq), it can infer the proportions of each cell type in each mixture (f). The platform has a space limit of 1GB, so I had to filter the matrix of single cell raw counts (comprising all CNS cells) and remove undefined cell clusters, neuronal clusters whose definition was dubious, and perform random sub-sampling of neurons, oligodendrocytes, and microglia, to remove 4000, 2000 and 1000 cells, respectively. Genes expressing less than 5 counts in at least 100 cells were further removed. For the construction of the cell type signature matrix, an average minimum gene expression threshold of 0 log_2FC was chosen, as advised in the platform for single cell data from 10X Genomics (otherwise, the sparsity of the data could lead to an unreliable signature matrix). Bulk RNA-seq data were taken from the GTEx project, in this work referred to as "validation data".

Chapter 4

Results and Discussion

4.1 Problem Description

After pre-processing the scRNA-seq data and defining astrocytic clusters with potential biological relevance, the next step was the identification of clusters enriched in older samples and to unveil their biological functions. The goal of the remaining analysis was indeed to unravel gene expression alterations in ageing astrocytes that may be pathological and can contribute to the predisposition of the ageing brain to neurodegenerative diseases.

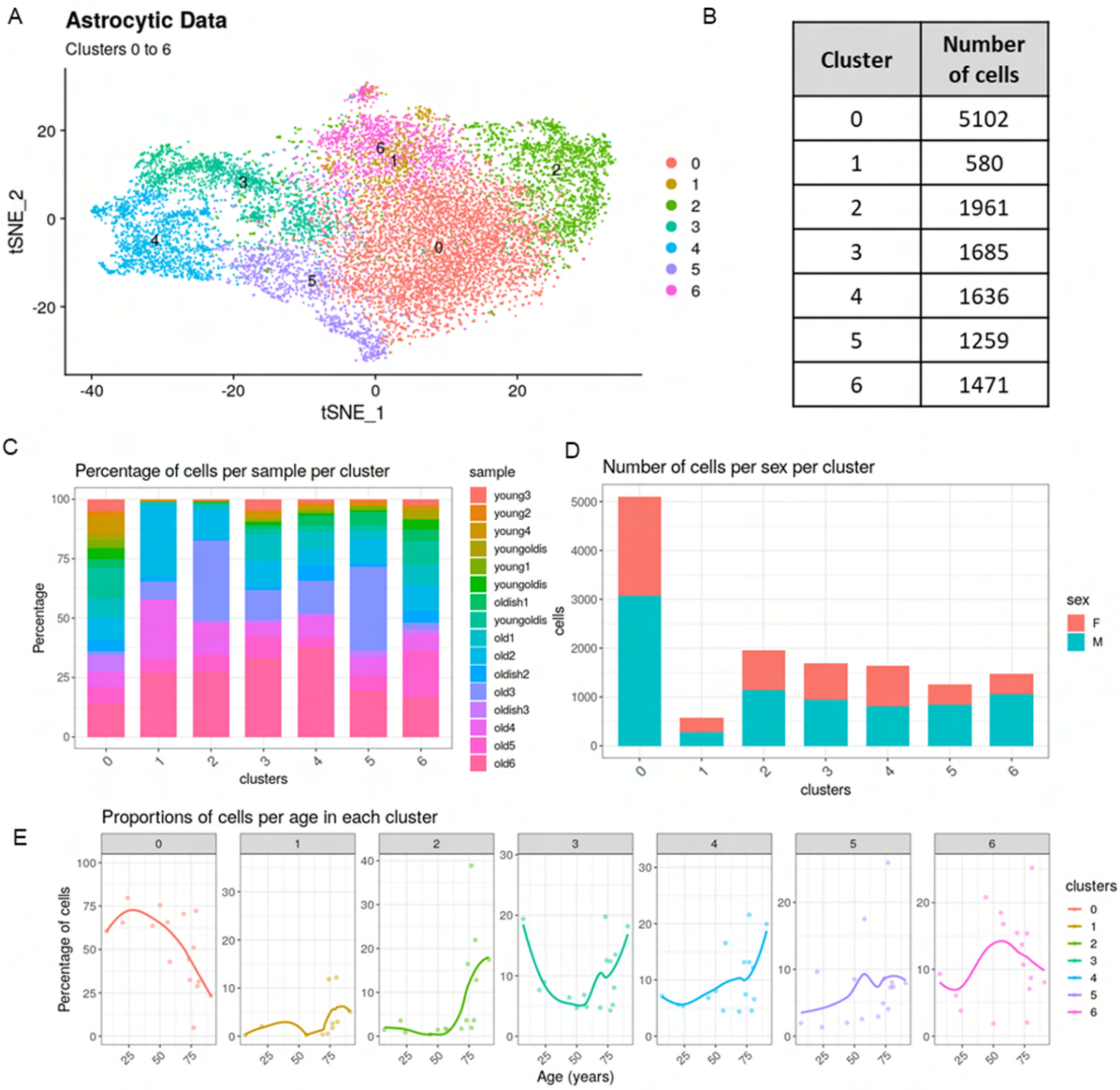
4.2 Exploratory analysis of astrocytic data

4.2.1 scRNA-seq data of human postmortem brain tissue reveal distinct clusters of astrocytes with unique characteristics

Astrocytes are a very heterogeneous group of cells in terms of function and molecular individuality. Each of the clusters in figure 4.1 (A) can be associated with a different type of astrocytes (type 0, type 1, etc.), whose distinctiveness is explored in the following sections.

Each cluster has cells from all individuals. Some astrocytic clusters (e.g., 2 and 4) are more represented in older samples, and some (e.g., 0) in younger samples (figure 4.1 (E)). However, any association between clusters and possible age-related biological functions should benefit from information about signalling pathways and cellular processes therein, to be explored in the next sections.

To define the transcriptomic profile of each of the clusters, I performed DEA of one cluster against the average of the others (table (C.1)). I obtained a list of genes, ordered by the magnitude of the difference (logFC), for each one of the astrocytic clusters. With this list of genes for each cluster, I performed gene set enrichment analysis (GSEA), in order to infer which pathways or biological processes may be up-regulated or down-regulated therein. However, as the clusters are of cells of the same cell type (astrocytes), a “blind” GSEA analysis was not very useful to unravel the subtle functional individualities of each cluster, as the GSEA hits were too vague to draw solid conclusions (e.g., down-regulation of ”Locomotion” in cluster 3 or ”Behavior” in cluster 4). Therefore, I chose to perform, as a first approach,



Cluster	Number of cells
0	5102
1	580
2	1961
3	1685
4	1636
5	1259
6	1471

FIGURE 4.1: Exploratory analysis of astrocytic clusters 0 to 6
(A) Representation in a t-SNE plot of the astrocyte clusters that make the basis of this work; **(B)** Distribution of the number of cells per cluster; **(C)** Number of cells per sample per cluster; **(D)** Number of cells per sex per cluster; **(E)** Proportions of cells of each cluster along age, with each dot representing a percentage of cells of each cluster in each sample.

GSEA with only known astrocytic biological pathways and processes, such as synapse maintenance and neuronal support (figure 4.2). Although I am aware that, by doing this, I am not allowing for the discovery of novel astrocytic functions, this is only the first step of the analyses. In the remaining sections, I will explore other astrocytic functions and relationships between clusters.

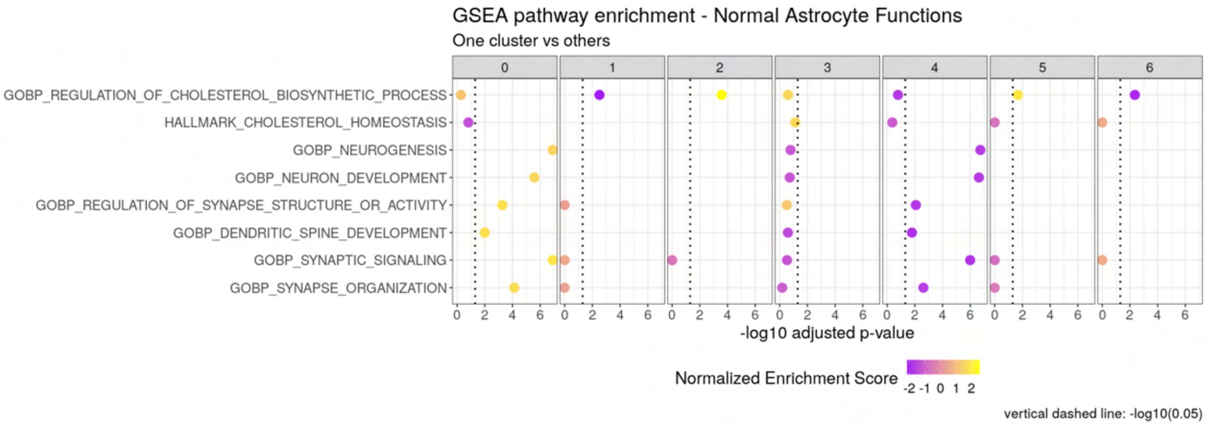


FIGURE 4.2: GSEA of normal astrocytic functions in defined clusters of astrocytes
Results of GSEA of differences between the clusters and the average of the others, with gene sets associated with known biological pathways and processes associated with normal astrocytic functions.

It was possible to define cluster 0 as a group of ”baseline” astrocytes, that is, whose functions are in accordance with what is expected from a healthy astrocyte (hereafter referred to as ”normal” functions). This cluster has an up-regulation of biological processes such as synapse organization, synaptic signalling and neuron development when compared with the other clusters (figure 4.2), suggesting that the others may have undergone some decline in those processes. Although cluster 0 is the most populated cluster (figure 4.1 (B)), it has a decrease in proportion in older samples (figure 4.1 (E)). This reinforces not only that these may be ”baseline” astrocytes, present in all samples, but also allows to hypothesise that older samples may be down-regulated in some of this neuronal and synaptic support functions, in accordance with what is already known regarding aged brains. Cluster 4 exhibits a particularly emphatic down-regulation of those normal astrocytic functions (figure 4.2).

4.2.2 Gene Expression Similarities Between Clusters

By performing DEA of one cluster against the others, the most differentially expressed genes in that cluster may also be differentially expressed between other clusters. This happens because under that formulation of DEA, the differentially expressed genes in one cluster are defined as different from the average astrocyte in the others, and not from each other cluster individually.

This overlap on the top differentially expressed genes can be observed between clusters 3 and 4, unravelling similar differences between these clusters and the others, suggesting some similarity between them, potentially attributable to related activation states. We can see a similar, albeit not as strong, overlap in strong expression of one cluster’s top differentially expressed genes in other clusters (e.g., between cluster 6 and 0), although not as strongly as in clusters 3 and 4.

Such analysis may give more clues on the relationship between astrocytic types and/or states.

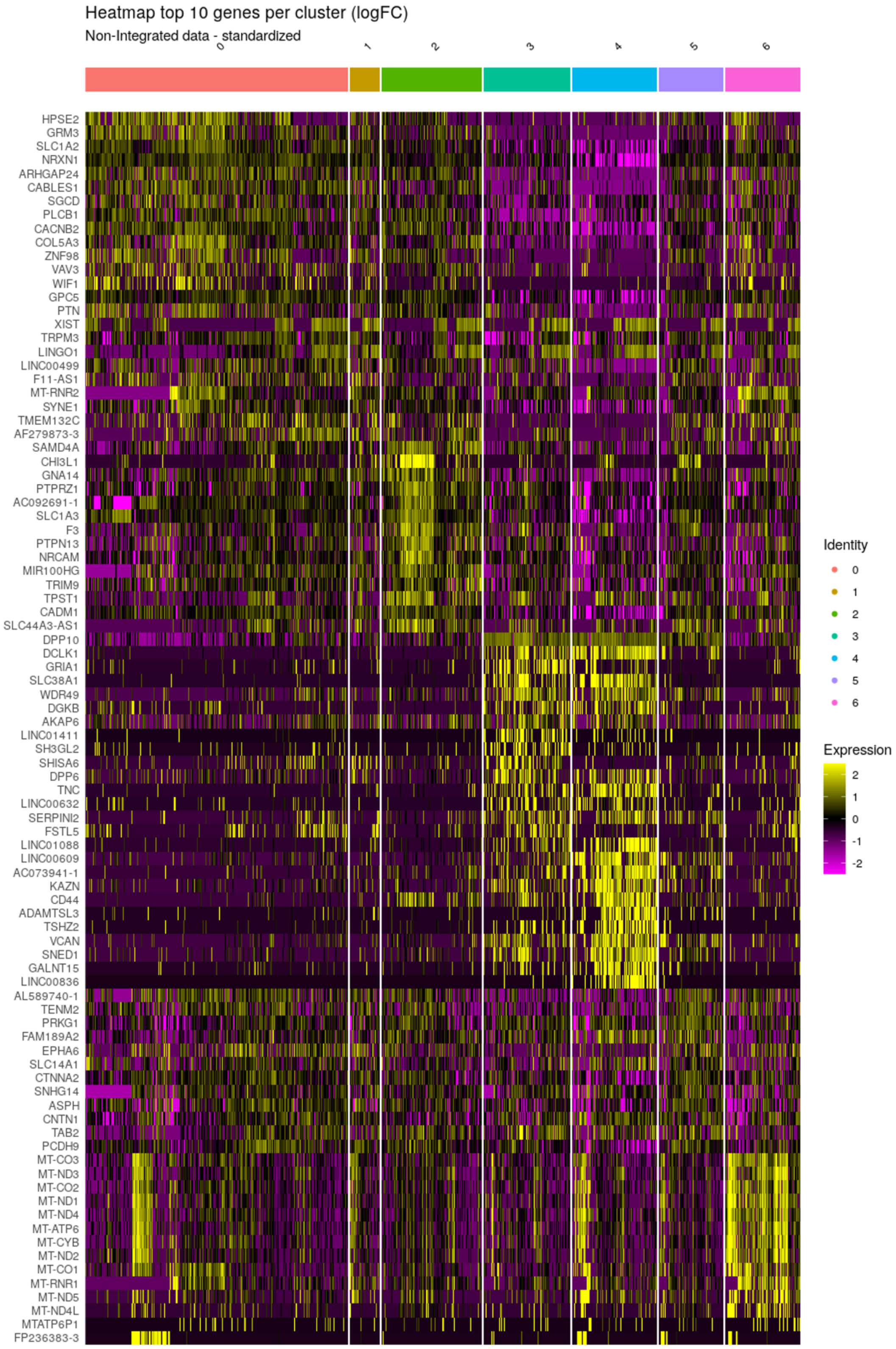


FIGURE 4.3: Heatmap of expression of the top 10 marker genes from each astrocytic cluster
Heatmap of the standardized expression of the top 10 marker genes of each cluster (based on the logFC in expression between that cluster and all others).

4.2.3 Reactivity does not explain the main differences between the analysed human astrocytic clusters

According to Clarke *et al.* (2018), ageing astrocytes acquire an A1 reactive, that is, pro-inflammatory phenotype. This could be due to the exacerbation of the inflammatory phenotype concomitant with age, with a mechanism similar to positive feedback. With this statement, it can be hypothesized that some of the clusters, perhaps clusters 2 and 4 as they appear to be enriched with age (figure 4.1 (E)), present a reactive phenotype. However, by studying the expression of specific reactivity markers [141], it can be noted that this does not appear to be the case (figure 4.4). Particularly in cluster 2, it can be noticed a down-regulation of the *SPARCL1* gene (also down-regulated in A1 astrocytes) but, at the same time, an up-regulation of the *STAT3* gene (up-regulated in A2 astrocytes) and a down-regulation of *FKBP5* (down-regulated in A2 astrocytes). In cluster 4, there is a similar reactivity dilution, with up-regulation of *C1QB* (up-regulated in A1 astrocytes) and *STAT3* (up-regulated in A2 astrocytes) and down-regulation of *SPARCL1* (down-regulated in A1 reactive astrocytes).

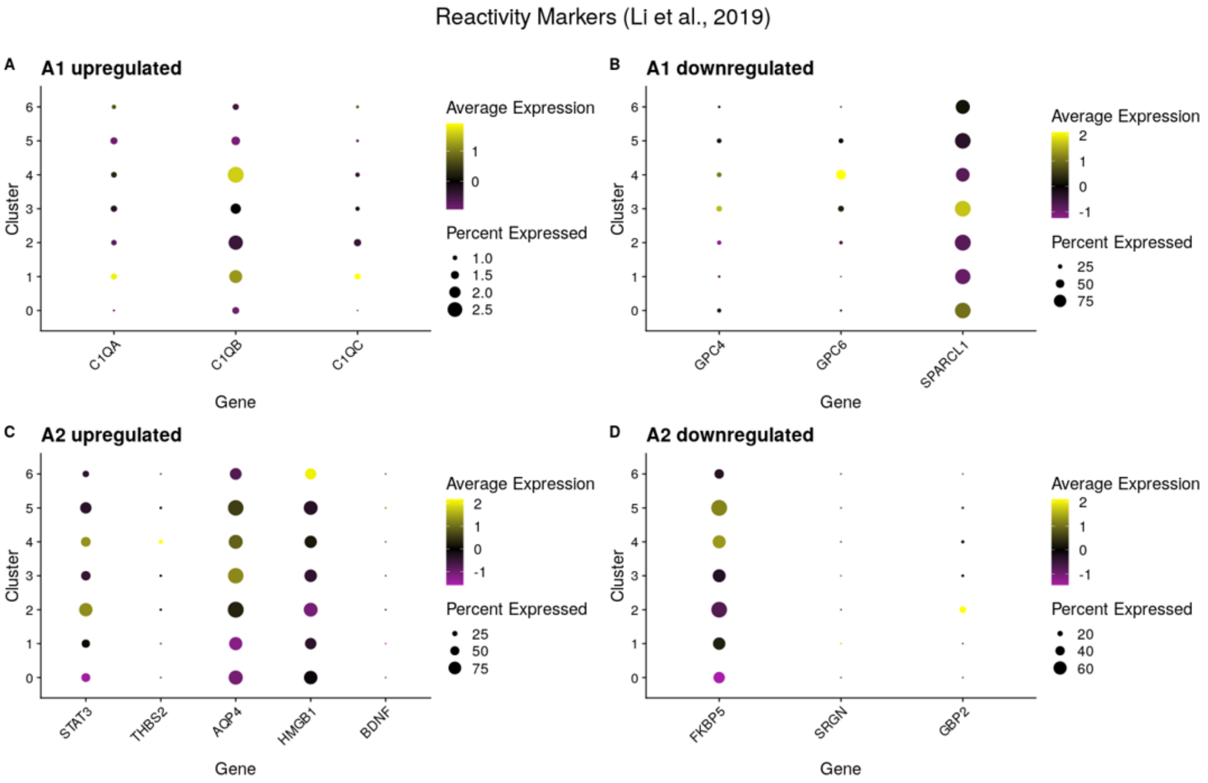


FIGURE 4.4: Expression of reactivity markers in clusters 0 to 6
Cross-cluster expression of markers for each type of reactivity [141], taking into account the binary classification suggested by Liddelow *et al.* (2017). **(A)** Up and **(B)** down-regulated markers of A1 astrocytes. **(C)** Up and **(D)** down-regulated markers of A2 astrocytes.

None of the clusters exhibits exclusive expression of markers of either A1 or A2 classification of reactivity, suggesting that for this dataset it is not reactivity that dominates the main differences in gene expression between clusters.

4.2.4 Pseudotime Inference as a glance on astrocytic relationships

Pseudotime inference was a key tool to unmask possible relationships between clusters. Although being developed for the inference of times associated with differentiation or temporal progression of cells, it generally allows the discovery of relationships associated with the greatest variance in the data, as it uses similarity between transcriptomes to infer proximity.

In practice, the application of pseudotime inference to astrocytic data enlightens a transitional relationship between clusters in PCA plots (figure 4.5), parallel to the major axes of variance. Using cluster 0 cells as the ”baseline” astrocytes and therefore as the origin of progression, the main axis of variance (PC1) seems to be associated with the 5 → 3 → 4 cluster progression. Similarly, the second main axis of variance (PC2) seems to be associated with progression to cluster 2. The third main variance component (PC3) does not reveal any meaningful trend, except perhaps the transition from cluster 1 to 6 (which, before being separated in the pre-processing step, belonged to the same cluster).

The relationships suggested by pseudotime inference using PCA may be a way of unmasking biological function and increase the molecular resolution of the identity of different astrocytic states, and will be explored in the remaining analysis.

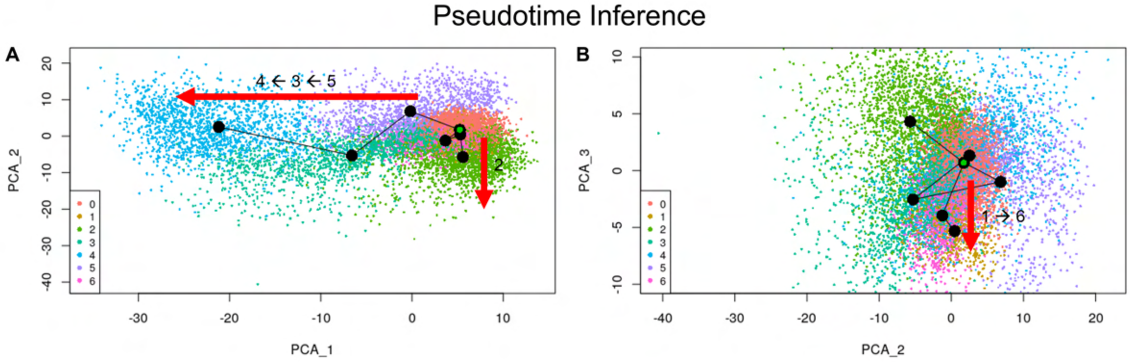


FIGURE 4.5: Pseudotime Inference in PCA space
Representation of the 7 clusters of astrocytes that are at the base of this work, in PCA plots (**(A)** first and second, and **(B)** second and third Principal Components), and with the results of pseudotime inference with potential trajectories between clusters, taking cluster 0 (coloured in salmon, and identified by a black dot with a green centre) as the ”origin”.

4.3 Gene Signature of Human Ageing Astrocytes

4.3.1 Stress as the presumed main source of variance in astrocytic gene expression

I have associated the main astrocytic gene expression variance axes with a progression between clusters (figure 4.6). Namely, the positive end of PC1 seems to be associated with clusters 0, 1, 2 and 6, with clusters 5, 3 and 4 along the progression of PC1 to more extreme negative values. Similarly, PC2 seems to be mainly associated with cluster 2 and 3 in the negative axis of variance, and clusters 4 and 5 in the positive axis. These progressions, also suggested by pseudotime inference analysis, can be speculated to be of various natures, such as transitions between types of astrocytes (e.g., type 5 astrocytes evolve to type 3 and then to type 4). The exact nature of these progressions will be explored in the next sections.

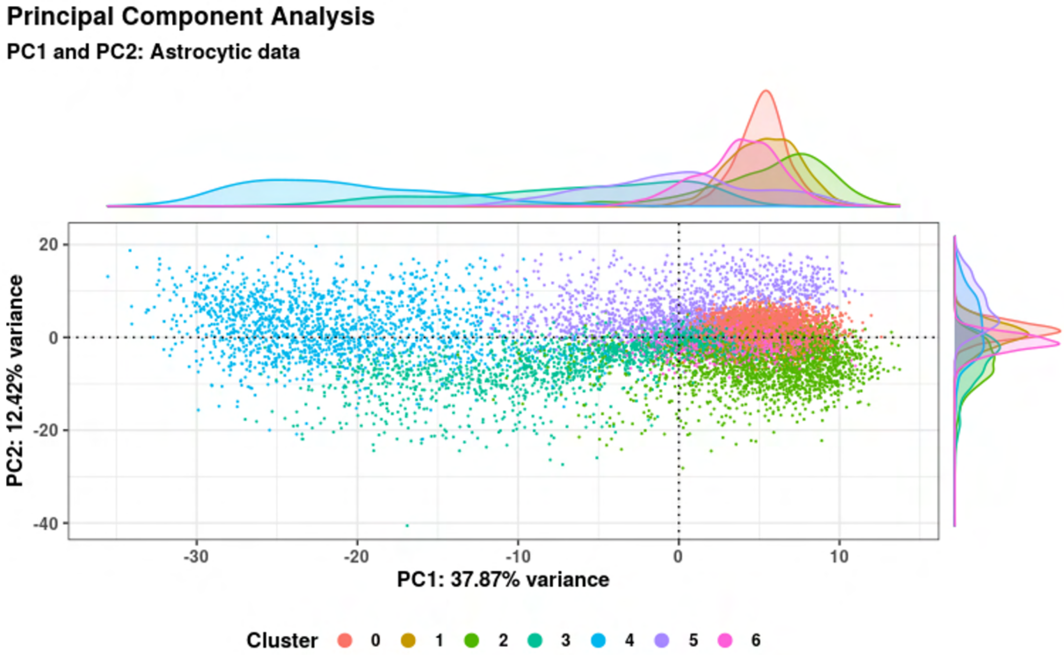


FIGURE 4.6: PCA of astrocytic gene expression
Representation of the 7 clusters of astrocytes that are at the base of this work, in a PCA plot of their gene expression, with adjacent curves associated to each clusters' density along each of the main axis of variance.

By performing GSEA on the differentially expressed genes between each cluster and the baseline cluster, some gene sets (hereinafter referred to as ”hits”) associated with endoplasmic reticulum (ER) stress were found enriched in clusters 2 and 5 (figure 4.7). Namely, these clusters present an up-regulation of biological processes of granule assembly stress and unfolded protein response, and of the unfolded protein response hallmark, already related to ageing (figure 2.3). This suggests that clusters 2 and 5 may be enriched in stress markers, which are known to be strongly associated with age.

However, there are some factors that differentiate clusters 2 and 5, and that allow a better definition of their biological identity. In addition to cluster 2 being enriched with age (figure 4.1 (E)), we also find reactive oxygen species pathways and TGF-β signalling down-regulated in cluster 5 when compared with cluster 2 (figure 4.8, panel ”5vs2”). This might suggest that cluster 2 has chronic stress markers associated with compensatory immunosuppression (figure 2.3). On the other hand, cluster 5 is not predominantly associated with young or old samples. Furthermore, this cluster is associated with the progression of clusters 5 $\rightarrow$ 3 $\rightarrow$ 4 in the main axis of variance. This may suggest that cluster 5 is an acute ER stress cluster (i.e., not associated to compensatory immunosuppression) that is somehow related to clusters 3 and 4.

Clusters 3 and 4 appear to be the most similar in gene expression (figure 4.3). In functional terms,

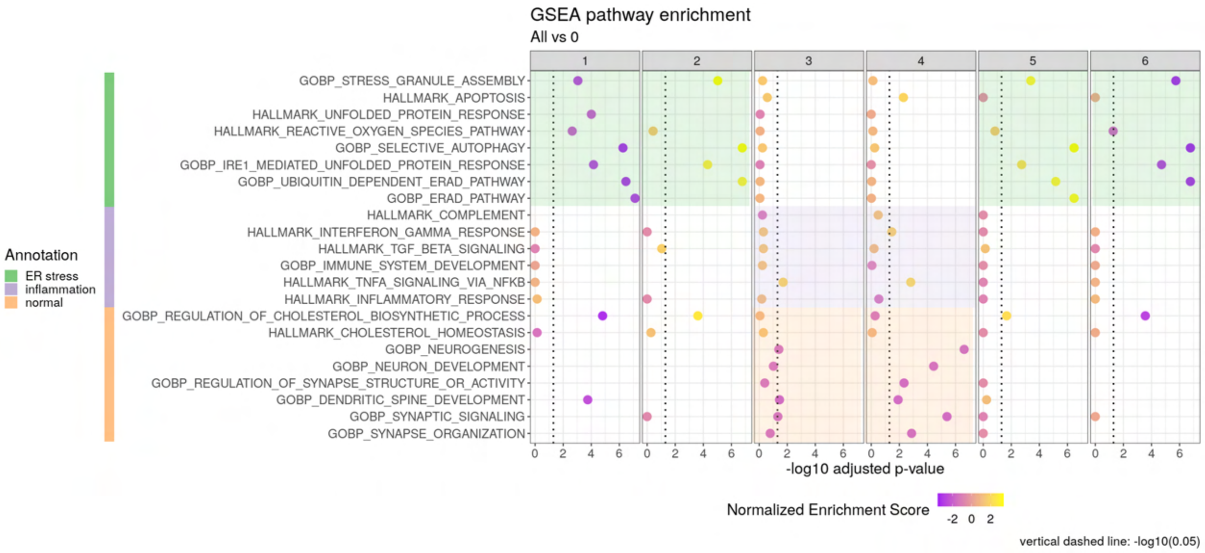


FIGURE 4.7: GSEA of ER stress, inflammation and normal astrocytic functions in clusters 1 to 6 GSEA of pathways, hallmarks and biological processes, associated with ER stress, inflammation and normal astrocytic functions in clusters 1 to 6, each compared to cluster 0.

cluster 4 has enriched neuroinflammation markers (figure 4.7) and down-regulated markers associated with neuronal support functions and synaptic homeostasis. Furthermore, cluster 3 appears to be more enriched in normal astrocytic functions when compared to cluster 4 (figure 4.8, panel ”3vs4”). Given that these clusters are, together with cluster 5 (acute stress), discriminated along PC1, this might suggest that both clusters 3 and 4 are also acute stress responders characterised by down-regulation of normal astrocytic functions, with 3 being a milder version of 4.

Clusters 1 and 6 appear to have ER stress markers down-regulated (figure 4.7), when comparing to the baseline cluster. However, I could not find any other functional information and they are not associated with any meaningful axis of variance in gene expression. As these clusters have been problematic since the beginning of the analysis (with a high percentage of mitochondrial genes that may suggest poor quality cells) and are not associated with age, I chose to not further study them in detail.

The previously suggested cluster relationships and possible functions are consistent with the genes associated with each main component of PCA (i.e., the genes with greater weight in the two main axes of variance)[1]. The cluster progression 5 → 3 → 4 ((figure 4.9) (A)) is associated with some genes associated with structural remodelling and cell division. *DCLK1* is associated with radial migration and axon growth of cortical neurons, which may be a response to neuron injury that happens with ageing. *TNC* is associated with guidance of migrating neurons as well as axons during development, synaptic plasticity, and neuronal regeneration. *GPC6* is associated with the control of cell growth and division. This may suggest that the PC1 axis is in part associated with the attempt to recover after an insult (stress). Furthermore, this axis has at its opposite end ((figure 4.9) (C)) genes associated with tumour suppression (*LRRC3B*) and synapse function and homeostasis. *GRM3* encodes the glutamate metabotropic receptor 3, whose decrease in expression can increase glutamate signalling. *RIMS1* regulates synaptic vesicle exocytosis and its downregulation has been observed in AD samples [143]. Another gene asso-

[1] If not stated otherwise, the genes’ description was retrieved from the GeneCards Human Gene Database [142]

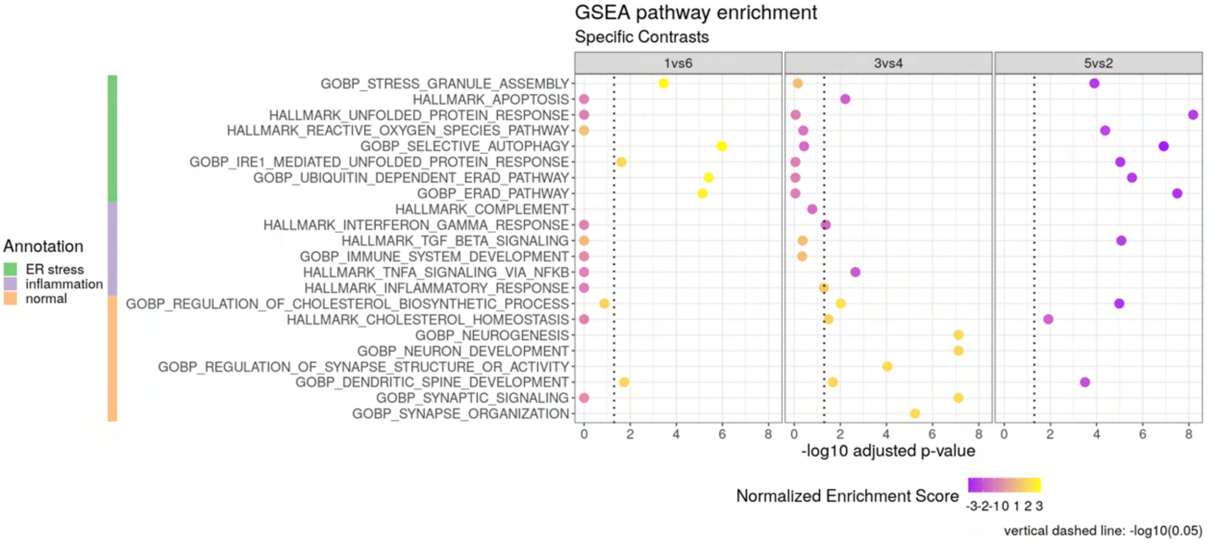


FIGURE 4.8: GSEA of ER stress, inflammation and normal astrocytic functions for specific cluster contrasts
GSEA of pathways, hallmarks and biological processes, associated with ER stress, inflammation and normal astrocytic functions of specific contrasts between clusters (1 against 6, 3 against 4, 5 against 2).

ciated with (+) PC1 is *ZNF98*, important for regulating apoptosis, protein folding and assembly. These genes are down-regulated in the progression axis 5 → 3 → 4, supporting the idea that the (-) PC1 axis of variance is mainly associated with a dysregulation of normal synaptic functions.

Finally, genes associated with both (-) PC1 and (-) PC2 ((figure 4.9) (B)), therefore potentially associated with both forms of stress, are linked to neuroinflammation (*CD44*) and neurotransmitter cycling and detoxification (*SLC38A1*). *MAN1C1* is not yet associated to major biologically relevant effects on the brain and astrocytes, but its overexpression is related to metabolic functions and has been associated to renal tumour suppression (less cell viability, colony formation, induced apoptosis, suppressed cell invasion and migration). At the opposite extreme, genes mutually associated with (+) PC1 and (+) PC2 ((figure 4.9) (D)) suggest a downregulation in the progression 5 → 3 → 4 of astrocyte proliferation, where decreased *CABLES1* (regulation of the cell cycle) may lead to increased number of apoptotic cells, and decreased *ERBB4* may restrain basal proliferative activity of hypothalamic astrocytes [144]. Furthermore, these axes are associated with *GPM6A*, which is highly expressed in mature neurons and is a major component of the axon growth cone during development and synaptogenesis. Finally, they are also associated with *FLRT2*, which can regulate memory functions in the adulthood. Both axes of stress (-PC1 and -PC2) therefore involve the down-regulation of genes necessary for astrocyte survival.

In summary, our analyses suggest that the main axes of variance in astrocytic gene expression are enriched in chronic (cluster 2) and acute (cluster 5) stress markers, with common downregulation of genes necessary for astrocyte survival. Furthermore, the main axis of variance is associated with a progression of clusters 5 → 3 → 4, which could suggest an acute stress response by clusters 3 and 4, enriched with age, and with potential dysregulation of some synaptic and neuronal support functions.

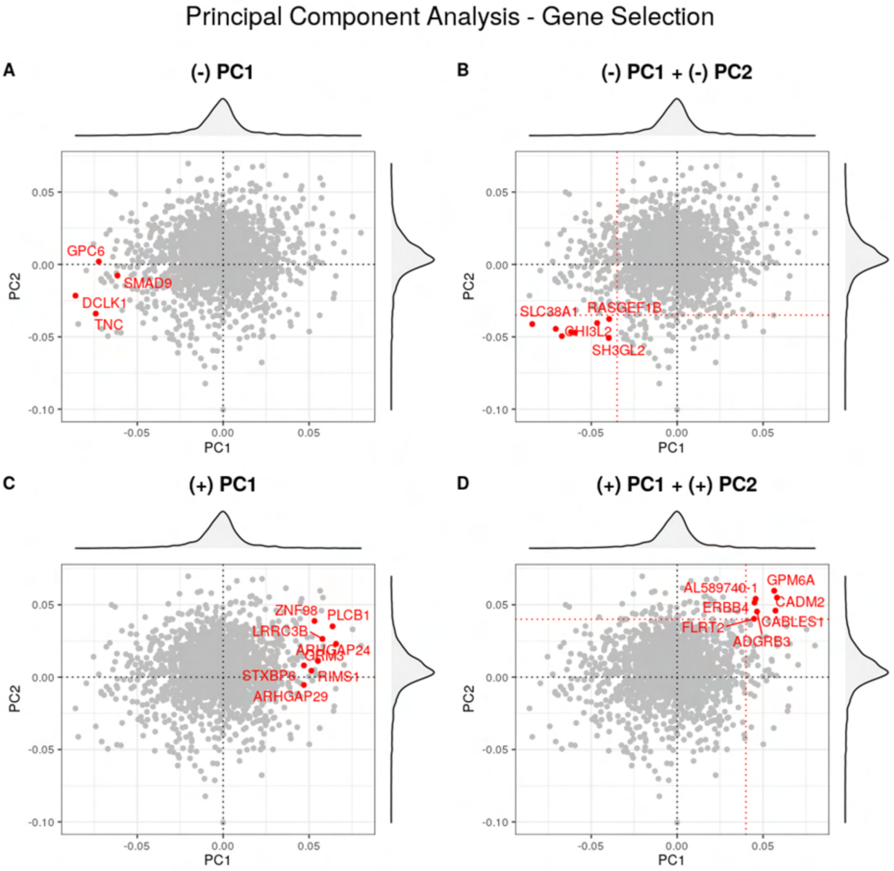


FIGURE 4.9: Genes with the highest and lowest weights in each PC
Representation of the weight of each gene in each of the data's largest variance axes (PC1 and PC2). In particular, there are highlighted, in red, genes whose biological function proves to be more interesting in light of the progression of clusters 5 → 3 → 4 **(A)**, or opposite **(C)**, as well as genes that are associated with both axes of progression, negative **(B)** or positive **(D)**, and which therefore may indicate genes associated with the progression of stress. The dashed lines in panel **(B)** and **(D)** (-0.35 and 0.4, respectively) indicate the thresholds chosen to select genes associated with both PCs.

4.3.2 Pseudo-bulk analysis validates the main axis of astrocytic gene expression variance

Pseudo-bulk RNA-seq data (i.e., simulated bulk RNA-seq data, by artificial pooling scRNA-seq samples) can be used in scRNA-seq studies, as it may dilute some of the noise present in scRNA-seq data (such as dropouts). For this purpose in the context of this project, the counts for each gene in all cells of an individual were summed for each cluster. This step required an initial filtering of cluster-individual pairs with less than 20 cells and genes with counts below 15 after the pooling of all cells of each individual, in order not to compromise the data normalization step using `edgeR` (figures E.1 and E.2).

Furthermore, the first three components seem to discriminate datasets, and the fourth only a single sample (figure E.4). Only the fifth component seems to be associated with variance in the data not given by a batch effect, and is the only component that correlates with the first principal component of the single-cell expression data (figure E.3). Therefore, a reconstruction of the projected data was made, removing the first 4 PCs (figure 4.10 (A)).

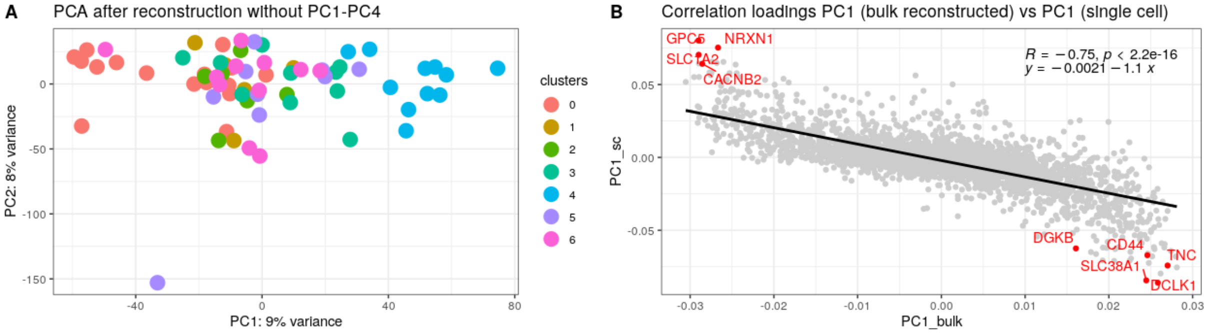


FIGURE 4.10: Pseudo-bulk RNA-seq data
(A) PCA of pseudo-bulk gene expression, resulting from pooling scRNA-seq data from all cells of each cluster from each individual. **(B)** Loadings/weights of genes associated with PC1 of pseudo-bulk (horizontal) and single cell (vertical) gene expression, with red highlighting of some of the top genes associated with single cell PC1 and cluster 4.

As can be seen in figure 4.10 (B), the signal of the main axis of variance in the single cell data remains in the pseudo-bulk data. Some of the main genes associated with +/- PC1 and cluster 4 are also represented at the extremes of this new axis of variance, which states for the robustness of this possibly biologically relevant signal.

4.3.3 Acute stress as a possible target for reversing loss of function in ageing astrocytes

So far, we have observed that cluster 2 of astrocytes is enriched in chronic stress markers and more abundant in older samples, and cluster 5 is enriched in acute stress markers and that does not show a particular association with age. Also, clusters 3 and 4 appear to become more prevalent in old ages and are involved in the inferred trajectory of clusters 5 $\rightarrow$ 3 $\rightarrow$ 4, aligned with the principal axis of gene expression variance (PC1). Given that clusters 3 and 4 are associated with a neuroinflammatory phenotype and depleted of markers of normal astrocytic functions, this might suggest that ageing astrocytes have greater difficulty adapting to acute stress, with cluster 3 being a “milder” state of cluster 4.

As clusters 2 and 4 are the extremes of the variance and associated with age, it will therefore be in the interest of this work to further study them, as they can be a potential factor for the deregulation of the normal functions of the CNS and predisposition to neurodegenerative diseases.

Combined with the GSEA results suggesting that cluster 4 may be associated with a loss of function by ageing astrocytes in response to acute stress, looking at individual differentially expressed genes therein could give some more specific functional insights into this cluster (figure 4.11 (A)). Astrocytes in cluster 4 have a deficiency in *SLC1A2* (important for synapse clearance and to prevent excitotoxicity) and *CADM1* (important to maintain functional excitatory synapses). Also, this cluster has an upregulation of *SLC38A1*, a gene encoding for the precursor of GABA and glutamate neurotransmitters. Additionally, this cluster has up-regulated *DCLK1* (axon growth and migration), *DPP10* (synapse homeostasis, binds to voltage-gated potassium channels), *KAZN* (cytoskeletal organization), and *CD44* and TNC (neuroinflammation). Finally, this cluster has down-regulated *NRXN1* (required for efficient neurotransmission), *GPC5* (control of cell division and growth regulation - AD samples have shown to be down-regulated in *GPC5* and *NRXN1* [145]), *CACNB2* (voltage dependent calcium channel protein) and *CABLES1* (important for cell cycle progression, knockdown leads to increased numbers of apoptotic cells). All of the above suggest that cluster 4 of astrocytes exhibits several characteristics known to be associated to pathological ageing (excitotoxicity, downregulation of specific genes, neuroinflammation, etc.).

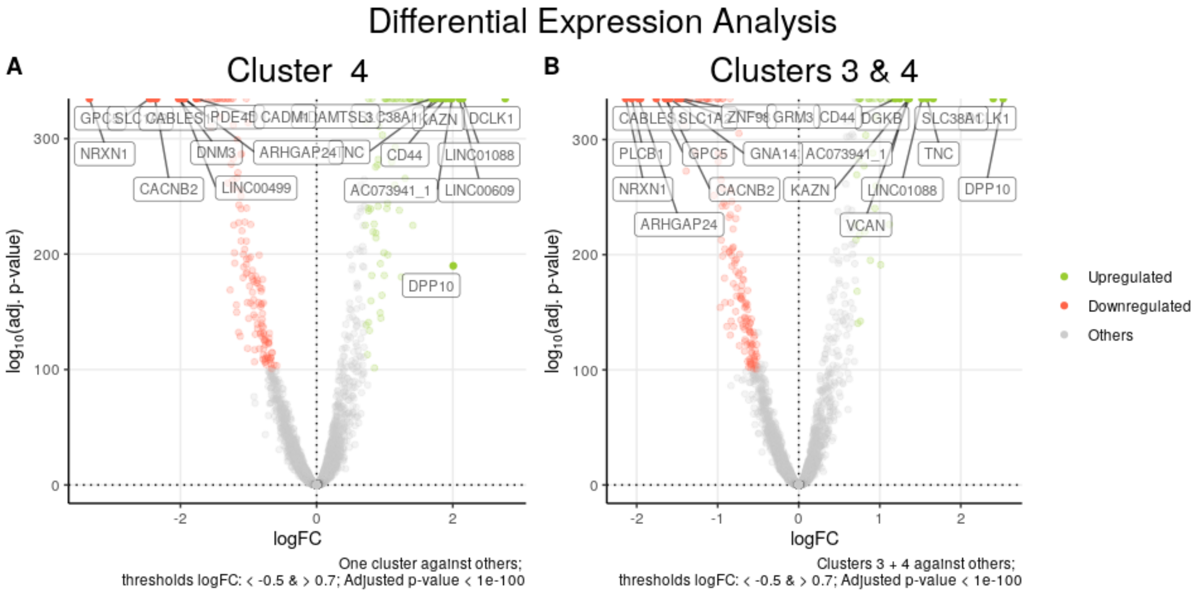


FIGURE 4.11: DEA of astrocytes in clusters 4 and 3&4
Volcano plots of differential expression analysis of astrocytes in **(A)** cluster 4 against the mean of all other clusters, and **(B)** cluster 3 + 4 against the mean of all other clusters. Some of the most differentially expressed genes are highlighted.

DEA in cluster 3 essentially suggests the same, having already been discussed that these clusters are quite similar in terms of differentially expressed genes, with cluster 3 showing a slightly more “normal” phenotype in terms of astrocytic functions (figure 4.8). Their combined analysis is also in accordance with this (figure 4.11 (B)).

Cluster 2 does not show enriched astrocyte-related processes in GSEA. Using PC2’s ordered list of genes by weights as input to GSEA was used in an attempt to discover more insights into the functions of astrocytes in cluster 2 (figure D.1). However, such results were not enlightening, and combined with the fact that PC2 is not exclusively associated with cluster 2, the functional characterization of cluster 2 was not possible.

Although clusters 2 and 4 are both at the extremes of the variance and associated with age, my analyses suggest it is more promising to focus on cluster 4 for subsequent validation and therapeutic exploration. Cluster 4 appears to have a stronger association with age, is at the end of the largest data variance axis and has a more coherent biological gene expression signal (unlike cluster 2, whose functional phenotype, in terms of astrocytic functions, could not be determined).

4.3.4 Candidate genes responsible for the main source of variance in ageing astrocytes

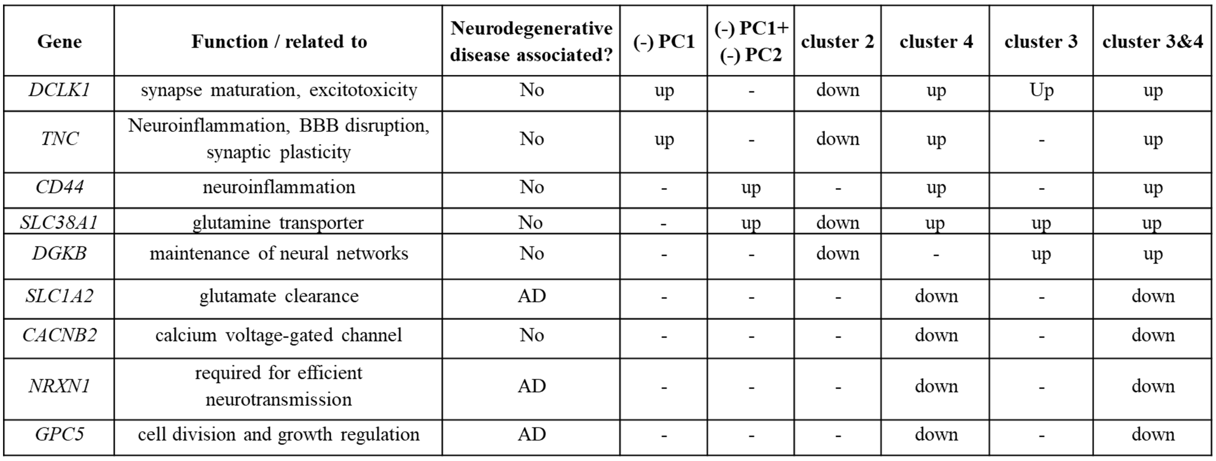
So far, it is known that chronic stress-associated immunosuppression affects brain homeostasis, being linked with ageing and neurodegenerative diseases, and that ER stress triggers an immunosuppressive reaction with implications for ageing and AD [31, 51]. However, there still seems to be no clear knowledge on resilience (i.e., the ability to recover) after acute stress in ageing astrocytes, only that it may be a way to predict healthy ageing and decreases with age [146].

My observation at this point is that cluster 4 is primarily enriched in aged samples, with increased markers of neuroinflammation and neuronal reconfiguration, as well as of synapse dysregulation and excitotoxicity, all of these being hallmarks of the ageing brain. Furthermore, pseudotime inference suggests that astrocytes in cluster 4 are acute stress responders. From a therapeutic point of view, it could be interesting to target marker genes of cluster 4 that directly or indirectly:

- Reduce neuroinflammation
- Regulate excitotoxicity
- Regulate synapse maturation
- Regulate the response to neuron injury

Some of these genes have already been described in section 4.3.3, and are summarized in table 4.1. They have been selected from the most up- or down-regulated genes in cluster 4 or clusters 3&4, with known functions associated with those previously listed. Their expression in the PCA space can also be found in figure 4.12 (A), and discriminated between clusters is illustrated in figure 4.12 (B).

TABLE 4.1: Table summarizing the main candidate genes in cluster 4 for phenotype emulation or reversal. Up = up-regulated; Down = down-regulated



Gene	Function / related to	Neurodegenerative disease associated?	(-) PC1	(-) PC1+ (-) PC2	cluster 2	cluster 4	cluster 3	cluster 3&4
DCLK1	synapse maturation, excitotoxicity	No	up	-	down	up	Up	up
TNC	Neuroinflammation, BBB disruption, synaptic plasticity	No	up	-	down	up	-	up
CD44	neuroinflammation	No	-	up	-	up	-	up
SLC38A1	glutamine transporter	No	-	up	down	up	up	up
DGKB	maintenance of neural networks	No	-	-	down	-	up	up
SLC1A2	glutamate clearance	AD	-	-	-	down	-	down
CACNB2	calcium voltage-gated channel	No	-	-	-	down	-	down
NRXN1	required for efficient neurotransmission	AD	-	-	-	down	-	down
GPC5	cell division and growth regulation	AD	-	-	-	down	-	down

Some of the expression of down-regulated genes in cluster 4 have already been perturbed in mice and associated with neuronal functions. Namely, disruptions in the expression of *Slc1a2* increase the susceptibility to neuronal degeneration [147], and *Nrxn1* knock-out mice show abnormal excitatory postsynaptic currents and a decrease in Ca^{2+} [148].

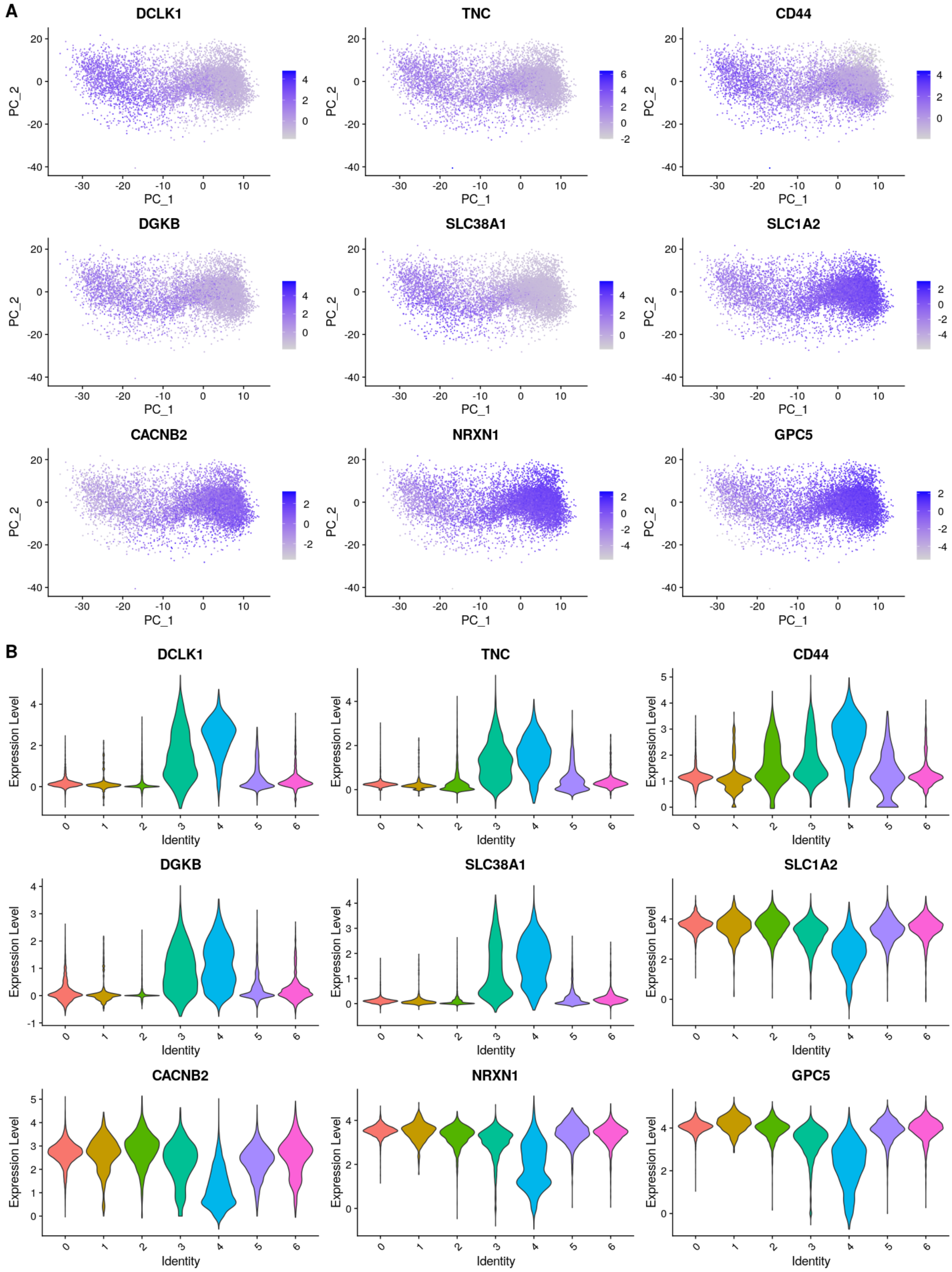


FIGURE 4.12: Marker genes of cluster 4
Expression of the main marker genes of cluster 4, through **(A)** representation of their expression in PCA space, and **(B)** through representation by violin plots of their expression in each cluster.

This set of genes has interesting biological functions that are consistent with the current hypothesis that there may be subtle changes in the ageing astrocyte transcriptome that can contribute to the predisposition of the ageing brain to neurodegenerative diseases.

4.3.5 Candidate compounds for phenotype reversal

Most drugs will have other mechanisms of action beyond those they were originally made for and thus the strategy of drug repurposing is a non-expensive and ethical way of surpassing the canonical time-consuming process of drug development [136, 137].

I used as input for `cTRAP` [134] the marker genes of astrocytes in cluster 4, obtained through differential expression analysis of cluster 4 against the others, ordered by t-statistic, and with adjusted p-value < 0.05 (BH correction). FDA-approved compounds that have shown simultaneously the best results for both `cTRAP` strategies, for each of the two metrics used (Spearman correlation coefficient and rank product[2]), were selected to be a basis for the discussion (figure 4.13):

1. For the *phenotype strategy*, I have selected compounds with negative Spearman coefficient ($<$ -0.01) between cluster 4's gene expression changes and those induced by CMap's compound perturbations, suggesting compounds that induce gene expression changes negatively correlated with cluster 4's phenotype, and thus being candidates for its reversal. Furthermore, compounds with low rank product coefficient ($>$ 60000) were also selected, i.e. compounds whose induced gene expression changes when compared to cluster 4's have the lowest combined ranks for correlation and functional enrichment.

2. For the *top gene strategy*, I have selected the compounds with positive Spearman coefficient ($>$ 0.05) between the differential gene expression results of cluster 4 and drug sensitivity results from *CTRP 2.1* database, and thus being the more likely compounds to target the marker genes of cluster 4. Similarly, compounds with high rank product coefficient ($<$ 100) were also selected.

This approach suggests Trifluoperazine, Niclosamide, Foretinib, and Olaparib as good candidates for phenotype reversal while targeting cells that express cluster 4's marker genes.

Trifluoperazine is a drug used for the treatment of schizophrenia for over 50 years [137]. It works by decreasing abnormal excitement – possible excitotoxicity – in the brain [149]. This drug is taken orally and has the capability of passing the BBB. It falls in the group of antipsychotic medications and is approved by FDA for these conditions. Zhang and colleagues (2017) have studied the potential of Trifluoperazine in preventing PD progression and showed that this drug can slow neurodegeneration by enhancing autophagy in response to stress [150].

Niclosamide is an orally-taken drug mainly used for the treatment of parasitic infections but it has shown preclinical potential in disease models of cancer and other infections [151]. Its proposed mechanism works by reducing the potential of the inner mitochondrial membrane to inhibit oxidative phosphorylation [152]. Some studies have proposed Niclosamide as a way of attenuating pro-inflammatory and

[2] The **rank product** summarises the individual rankings from cTRAP's comparison methods (Spearman, Pearson and GSEA-based scores)[134].

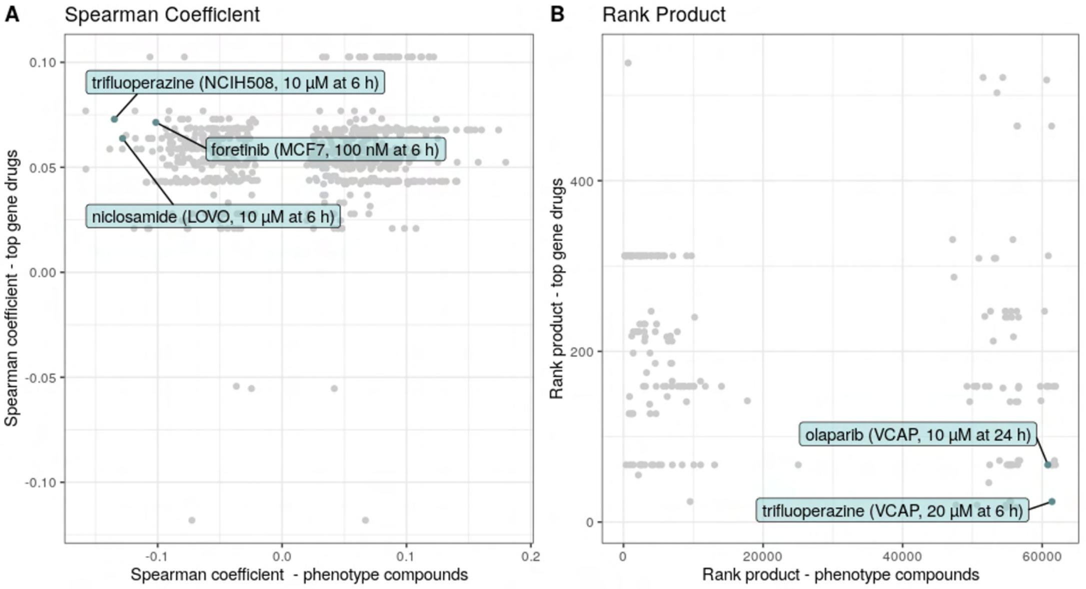


FIGURE 4.13: Identification of candidate compounds for reversal of cluster 4 phenotype
Scatter plots comparing the correlation between cluster 4's gene expression changes and those induced by each of CMap's compound perturbations (x axis) and the correlation between the differential gene expression results of cluster 4 and gene expression / drug sensitivity association across all cell lines from *CTRP 2.1* [134] (y axis). The comparisons are performed using **(A)** Spearman's correlation coefficient and **(B)** Rank product. Highlighted compounds are candidate reverters of cluster 4's phenotype that are FDA approved.

migratory phenotypes of microglia and astrocytes in ALS models [153]. There is also general evidence of Niclosamide as having a neuroprotective role, including prevention of neurodegeneration [154, 155].

Foretinib is a pan-kinase inhibitor, currently in clinical trials for the treatment of cancer. However, it has been proposed to prevent axon degeneration, via preservation of the mitochondria, being thus a candidate for many neurological diseases [156, 157].

Olaparib is a drug used in the treatment of several types of cancer, namely breast cancer or fallopian tube cancer. It was made to be taken orally and is demonstrated to fail passing the BBB in preclinical models. However there is evidence that these conventional models of the BBB may not predict clinical pharmacokinetics, and thus more studies should be performed on this possibility [158]. A study from 2020 suggested that the administration of Olaparib in a Huntington's disease model promoted neuroprotection and modulation of the inflammasome activation, resulting in the reduction of neurological deficits and improving the clinical outcomes in neurobehavioural tests [159]. Thus this drug could be also an interesting candidate for further studies.

Certainly, further validation of these *in silico* results is needed but they are a proof of concept and the basis for future research.

4.3.6 Computational validation of the enrichment in aged brains of acutely stressed astrocytes by digital cytometry

Digital cytometry is a technique that allows estimating the proportions of different cell types in bulk samples. This approach is particularly useful in this work, since single cell protocols may be biased in terms of the proportion of different cell types that are captured, and thus the proportions obtained from individual cell populations may not reflect the true composition of the human brain tissue.

When performing that cell type deconvolution with gene expression signatures for CNS cells including all types of astrocytes defined by our analysis, more than half of the cells in each of the bulk samples would correspond to type 6 astrocytes. Types 1 and 6 astrocytes were identified as problematic clusters throughout this analysis, with a high percentage of mitochondrial genes that may suggest poor quality cells. For this reason, I chose to exclude both clusters 1 and 6 from the analysis in order to achieve a more realistic resolution in astrocytic types, during the step of constructing the cell type signature matrix, which is shown in figure 4.14. Most cell types in the CNS appear to have strong signatures, i.e. groups of genes expressed exclusively therein. The astrocytic signature appears diluted among all astrocyte clusters, but it can be noticed that astrocytes of types 0 and 4 appear to have the most robust signature among all astrocytes.

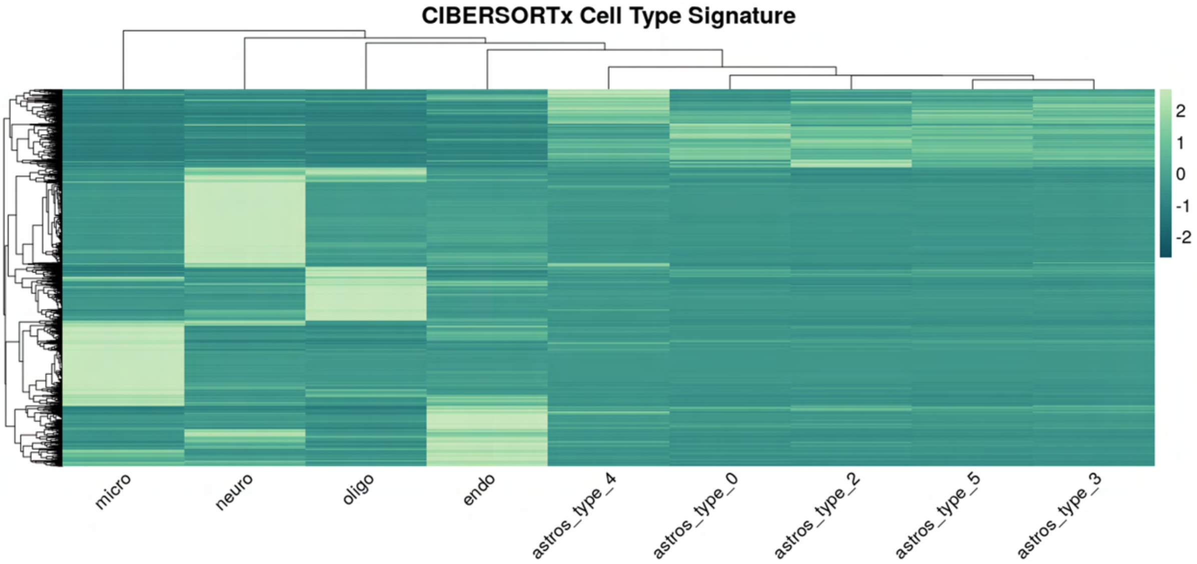


FIGURE 4.14: Cell type signature obtained through CIBERSORTx
Heatmap of the gene signature matrix for neural cell types, computed with CIBERSORTx, used to carry out the cell-type deconvolution task. Each row is a gene, and each column is a cell type. We can see that the main neural cell types have well-defined gene signatures.

By performing cell-type deconvolution in an independent dataset, consisting of bulk transcriptomes of cortex samples, the enrichment of cluster 4 in aged samples was validated (figures 4.15 and F.1), as well as a depletion of neurons in these samples. Oligodendrocytes appear to have constant proportions over age. Furthermore, type 5 astrocytes and microglia were not detected in these data. However, since microglia is expected to be one of the least abundant glial cell types, it may have been masked by the remaining cells. The same may have happened to type 5 astrocytes. Nevertheless, the enrichment of cluster 4 in aged samples is an important result of validation in independent data, demonstrating that the observation made with the scRNA-seq dataset was not due to chance.

It is important to notice that digital cytometry does not allow the comparison of absolute proportions between cell-types. Given that it estimates of the proportion of mRNA coming from each cell-type, the proportions of bigger cells (i.e., with a higher amount of mRNA) can be overestimated. Because of this, only the variation in proportions of each cell-type across samples were studied.

The cell-type gene signatures were derived from cortex data, and therefore direct extrapolation to other brain areas may not be possible. However, I have also performed this study in the cerebellum and hippocampus (figures F.2 and F.3). Type 4 astrocytes seem to be present also in these brain areas, but with a higher dispersion across the various ages (i.e., with an increasing trend not as obvious as in the cortex). At the same time, the proportions of neurons tend to decrease in older samples, maintaining the positive control used in cortex samples. This suggests a type of astrocytes that is not specific from the cortex, but whose enrichment in older samples is more pronounced therein.

In conclusion, the enrichment of type 4 astrocytes in aged cortices found with independent single cell and bulk transcriptomic datasets not only demonstrates the potential of transcriptomics and associated computational analysis tools for the study of the brain, but also gives the scientific community the transcriptomic profile of a candidate novel type of astrocytes for possible validation and therapeutic targeting *in vitro* and *in vivo*.

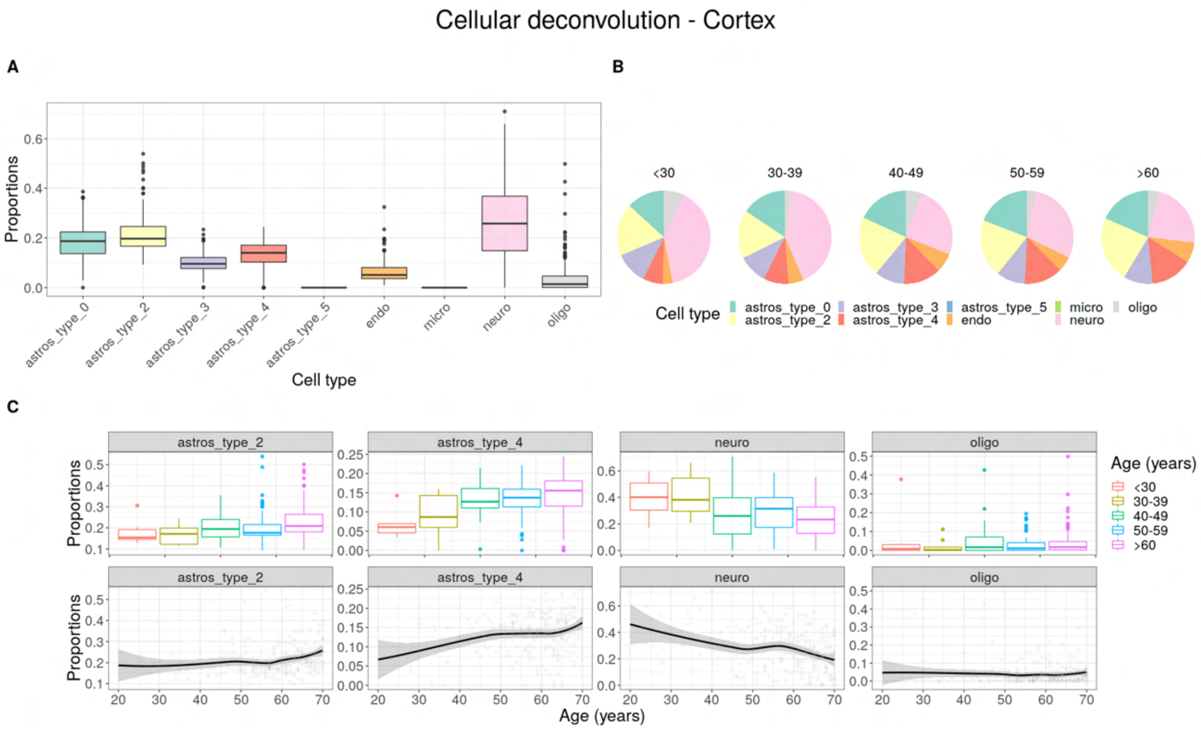


FIGURE 4.15: Cell-type deconvolution of cortex samples - highlights
Cellular proportions estimation for the various cell types in this work, including types 0, 2, 3, 4 and 5 astrocytes, neurons, microglia (micro), oligodendrocytes (oligo) and endothelial cells (endo), in cortex samples. **(A)** Distribution of cell-type proportions across cortex bulk RNA-seq samples (GTEx) represented in boxplots. **(B)** Distribution, by age group, of the proportions of the various cell types. **(C)** Distribution of proportions, by age group and through a general additive model along age (R `geom_smooth` function with default parameters), of cell types of greater interest.

Chapter 5

Concluding Remarks

Ageing is the strongest risk factor for numerous neurodegenerative diseases, yet the causes that underly the shift from physiological to pathological ageing remains nuclear. Several efforts have been made by the scientific community to discover those causal functional and molecular mechanisms.

Astrocytes are a very heterogeneous cell type from a functional and molecular point of view, being essential for neuronal survival and synapse homeostasis. However, occasionally these cells present pathological behaviours that are not protective of the central nervous system, namely in response to neuroinflammation concomitant with age. It is plausible that changes in astrocytes' phenotype can make the brain more vulnerable to injury and age-related diseases (pathological ageing). However, technologies currently applied to profile transcriptomes of bulk human brain tissues (such as RNA-seq) fail to detect subtle changes that may allow the identification of different astrocyte activity states. Single-cell RNA-seq allows the study of the transcriptomic profile of each cell individually. Given the complexity of the human brain, the transcriptomic resolution given by this technique allows to identify novel candidate genes and signalling pathways / biological processes characteristic of cells most relevantly contributing to the ageing of brain tissues.

In this work, I focused on implementing approaches for the analysis of publicly available human brain scRNA-seq data, in order to profile ageing-associated gene expression alterations in human astrocytes. This work culminated in the transcriptomic characterization of a group of astrocytes, named type 4 astrocytes or astrocytic cluster 4, whose enrichment with age was identified both in scRNA-seq data and in independent bulk RNA-seq data. Through relationships suggested by a variety of computational tools, such as PCA, pseudotime, DEA and GSEA, this cluster seems to be associated with a down-regulation of physiological astrocytic functions, in response to acute stress. This group of astrocytes appears to be enriched in markers of neuroinflammation and excitotoxicity, as well as loss of neuronal support and synaptic homeostasis functions, being therefore associated with known hallmarks of pathological ageing.

Furthermore, I was able to identify a set of candidate compounds, through a computational drug repurposing strategy, potentially capable of acting on the most differentially expressed genes in cluster 4 and reversing its phenotype.

This work scientifically contributed with the discovery of molecular targets for phenotype validation *in vitro* and *in vivo*, as well as candidate therapeutic compounds for the reversal of the pathologically aged astrocytes' phenotype.

5.1 Analysis Limitations

The analysis performed in this work followed several pipelines and tools shown to be state-of-the-art in benchmark studies. However, there are still some limitations that are worth putting into perspective.

First, given that the sample size in this work is relatively small, age may be counfounded with the biological invididuality from each sample. Secondly, the scRNA-seq data used in this work are only from the human cortex. Consequently, there is a lack of regional coverage. Both these caveats could be mitigated if I had access to a greater sample size comprising different areas of the brain.

Another caveat in this analysis was the scarcity of young brain samples. Due to the complexity of retrieving human brain samples in living subjects, this is usually done in post-mortem situations. There is a relatively high number of older brain samples because fortunately most deaths occur in older people. The ideal scenario would be to obtain samples of human brain tissue from healthy young individuals; however, biopsying healthy brains is impossible for obvious ethical reasons.

5.2 Suggestions and Future Work

The work described here consisted of the analysis of several healthy brain tissue samples (i.e., with no known neurodegenerative diseases), and the consequent identification of a group of astrocytes (cluster 4) with therapeutic potential for reversing some of their possibly pathological changes with age (e.g., loss of normal astrocytic functions). Although our analyses of gene expression alterations and therapeutic potential of astrocytes have focused mainly on cluster 4, it will still be interesting to further study cluster 3. This cluster seems to be associated not only with PC1 of astrocytic gene expression data but also to PC2. As both of these axes appear to convey different biological responses, the reason for cluster 3 to appear in both could also be interesting to further study in more detail.

Furthermore, it would be interesting to expand this study to other CNS cell types. For example, microglia are known to be in close contact with reactive astrocytes [67], and it is possible that microglia also have activation states, some even correlated with cluster 4's enrichment in older samples.

Despite an *in silico* validation of the enrichment of this group of astrocytes with age,functional validation *in vitro* or *in vivo* should also be performed on this group of aged astrocytes.

Moreover, the hypothesis put forward in this work is that these type 4 astrocytes are the result of a response to acute stress (i.e., ageing astrocytes have a greater difficultly adapting to this type of stress), shown by a downregulation of normal astrocytic functions. Due to this, it will be interesting, in addition to emulating their phenotype through genetic editing or under/overexpression of certain genes, to use various stressors (for example pharmaceutical ER stress inducers such as *tunicamycin* or *thapsigargin* [160], or physiologically-induced ER stress via glucose deprivation [160]) and to observe the response of astrocytic cell lines. It will also be interesting to study the phenotype associated with the transcriptomic profile of type 4 astrocytes in co-cultures of astrocytes and neurons, to emulate, as far as possible, the characteristic and essential neuronal support environment of astrocytes.

5.3 Concluding Remarks

Through this purely computational work, and using only **public data**, I identified, to my knowledge, a novel group of astrocytes associated with a down-regulation of physiological astrocytic functions and whose differentially expressed genes are candidates for further studies, such as phenotype validation and reversal. There are some authors who consider the practice of using public data unethical [161]. However, there is also a growing consensus that those entitled "research parasites" (i.e., scientists who use public data from other studies) make science move faster and more rigorously, without ever failing to credit the "hosts" of the data [162, 163]. My personal vision aligns with the latter, and the usage of only public data has proved to be one of the most fascinating parts of this whole research initiation project.

Although data generation is expensive, there is a lot of data publicly available, and an increasing importance given to data over new ideas and questions that can be addressed with the data [164]. In addition to being much cheaper and less time consuming than data generation, data sharing allows for rapid replication and validation of results, as well as new scientific discoveries beyond those for which the data were first obtained. It is possible to do science with public data, as long as the *right question* is asked. This project served as a fundamental part of my academic career, further stimulating my interest in the field of bioinformatics and computational biology, as well as my understanding of brain biology and neurodegenerative pathologies.

Biology, specifically the biology of the central nervous system, is tremendously complex and non-linear, and there are increasingly more tools that help to understand it. Over the years, additional functions performed by astrocytes have been discovered, and the concept of "cell type" may even be put into perspective. Certainly, in the coming years, more research will be done with a focus on this cell group, which has all the potential to explain these subtle shifts between physiological and pathological ageing.

Bibliography

[1] A. L. Palmer and S. S. Ousman. Astrocytes and aging. *Frontiers in Aging Neuroscience*, 10:337, 2018, doi:10.3389/fnagi.2018.00337.

[2] M. P. Mattson and T. V. Arumugam. Hallmarks of brain aging: Adaptive and pathological modification by metabolic states. *Cell Metabolism*, 27(6):1176–1199, 2018, doi:10.1016/j.cmet.2018.05.011.

[3] B. A. Yankner, T. Lu, and P. Loerch. The aging brain. *Annual Review of Pathology: Mechanisms of Disease*, 3(1):41–66, 2008, doi:10.1146/annurev.pathmechdis.2.010506.092044. PMID: 18039130.

[4] J. S. Lee, Y. H. Park, S. Park, U. Yoon, Y. Choe, B. K. Cheon, A. Hahn, et al. Distinct brain regions in physiological and pathological brain aging. *Frontiers in Aging Neuroscience*, 11:147, 2019, doi:10.3389/fnagi.2019.00147.

[5] World Population Ageing 2017 URL `https://www.un.org/en/development/desa/population/theme/ageing/WPA2017.asp`. Online; accessed 4 September 2021.

[6] C. Franceschi, P. Garagnani, P. Parini, C. Giuliani, and A. Santoro. Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nature Reviews Endocrinology*, 14(10):576–590, 2018, doi:10.1038/s41574-018-0059-4.

[7] I. Matias, J. Morgado, and F. C. A. Gomes. Astrocyte heterogeneity: Impact to brain aging and disease. *Frontiers in Aging Neuroscience*, 11, 2019, doi:10.3389/fnagi.2019.00059.

[8] M. Verkerke, E. M. Hol, and J. Middeldorp. Physiological and pathological ageing of astrocytes in the human brain. *Neurochemical Research*, 2021, doi:10.1007/s11064-021-03256-7.

[9] L. E. Clarke, S. A. Liddelow, C. Chakraborty, A. E. Münch, M. Heiman, and B. A. Barres. Normal aging induces A1-like astrocyte reactivity. *Proceedings of the National Academy of Sciences*, 115(8), 2018, doi:10.1073/pnas.1800165115.

[10] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009, doi:10.1038/nmeth.1315.

[11] I. P. Johnson. Age-related neurodegenerative disease research needs aging models. *Frontiers in Aging Neuroscience*, 7, 2015, doi:10.3389/fnagi.2015.00168.

[12] E. R. Kandel and M. N. Shadlen. *Principles of Neural Science*, chapter 1: The Brain and Behavior. McGraw Hill, 6 edition, 2021.

[13] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013, doi:10.1016/j.cell.2013.05.039.

[14] H. Clevers, S. Rafelski, M. Elowitz, A. Klein, C. Klein, J. Shendure, E. Lein, et al. What is your conceptual definition of "cell type" in the context of a mature organism? *Cell Systems*, 4(3):255–259, 2017, doi:10.1016/j.cels.2017.03.006.

[15] E. N. Marieb and K. Hoehn. *Human Anatomy & Physiology*. Pearson, 9 edition, 2013.

[16] E. R. Kandel and M. N. Shadlen. *Principles of Neural Science*, chapter 3: Nerve Cells, Neural Circuitry, and Behavior. McGraw Hill, 6 edition, 2021.

[17] New Cause of Schizophrenia Uncovered. Jul 2017. URL `https://neurosciencenews.com/glial-cells-schizophrenia-7139/`. Online; accessed 3 September 2021.

[18] S. Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3, 2009, doi:10.3389/neuro.09.031.2009.

[19] O. Gonzalez-Perez, V. Lopez-Virgen, and A. Quiñones-Hinojosa. Astrocytes: everything but the glue. *Neuroimmunology and Neuroinflammation*, 2(2):115, 2015, doi:10.4103/2347-8659.153979.

[20] M. C. Bordone. *The transcriptional landscape of Alzheimer's and Parkinson's diseases*. PhD thesis, Faculdade de Medicina, Universidade de Lisboa, 2020.

[21] C. S. V. Bartheld, J. Bahney, and S. Herculano-Houzel. The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology*, 524(18):3865–3895, 2016, doi:10.1002/cne.24040.

[22] S. Herculano-Houzel. Isotropic Fractionator: A simple, rapid method for the quantification of total cell and neuron numbers in the brain. *Journal of Neuroscience*, 25(10):2518–2521, 2005, doi:10.1523/jneurosci.4526-04.2005.

[23] C. Daskalopoulou, B. Stubbs, C. Kralj, A. Koukounari, M. Prince, and A. Prina. Physical activity and healthy ageing: A systematic review and meta-analysis of longitudinal cohort studies. *Ageing Research Reviews*, 38:6–17, 2017, doi:10.1016/j.arr.2017.06.003.

[24] J. Cohen and C. Torres. Astrocyte senescence: Evidence and significance. *Aging Cell*, 18:e12937, 02 2019, doi:10.1111/acel.12937.

[25] Y. Hou, X. Dan, M. Babbar, Y. Wei, S. G. Hasselbalch, D. L. Croteau, and V. A. Bohr. Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology*, 15(10):565–581, 2019, doi:10.1038/s41582-019-0244-7.

[26] D. Murman. The impact of age on cognition. *Seminars in Hearing*, 36(03):111–121, 2015, doi:10.1055/s-0035-1555115.

[27] A. A. Simen, K. A. Bordner, M. P. Martin, L. A. Moy, and L. C. Barry. Cognitive dysfunction with aging and the role of inflammation. *Therapeutic Advances in Chronic Disease*, 2(3):175–195, 2011, doi:10.1177/2040622311399145.

[28] M. T. Lin and M. F. Beal. The oxidative damage theory of aging. *Clinical Neuroscience Research*, 2(5-6):305–315, 2003, doi:10.1016/s1566-2772(03)00007-0.

[29] S. Fulda, A. M. Gorman, O. Hori, and A. Samali. Cellular stress responses: Cell survival and cell death. *International Journal of Cell Biology*, 2010:1–23, 2010, doi:10.1155/2010/214074.

[30] M. K. Brown and N. Naidoo. The endoplasmic reticulum stress response in aging and age-related diseases. *Frontiers in Physiology*, 3, 2012, doi:10.3389/fphys.2012.00263.

[31] A. Salminen. Increased immunosuppression impairs tissue homeostasis with aging and age-related diseases. *Journal of Molecular Medicine*, 99(1):1–20, 2020, doi:10.1007/s00109-020-01988-7.

[32] S. B. Chidambaram, A. Rathipriya, S. R. Bolla, A. Bhat, B. Ray, A. M. Mahalakshmi, T. Manivasagam, et al. Dendritic spines: Revisiting the physiological role. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 92:161–193, 2019, doi:10.1016/j.pnpbp.2019.01.005.

[33] D. Dickstein, C. Weaver, J. Luebke, and P. Hof. Dendritic spine changes associated with normal aging. *Neuroscience*, 251:21–32, 2013, doi:10.1016/j.neuroscience.2012.09.077.

[34] B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, et al. Preclinical Alzheimers disease: Definition, natural history, and diagnostic criteria. *Alzheimers & Dementia*, 12(3):292–323, 2016, doi:10.1016/j.jalz.2016.02.002.

[35] C. Franceschi, M. Bonafè, S. Valensin, F. Olivieri, M. De Luca, E. Ottaviani, and G. Benedictis. Inflamm-aging: An evolutionary perspective on immunosenescence. *Annals of the New York Academy of Sciences*, 908:244–54, 07 2000, doi:10.1111/j.1749-6632.2000.tb06651.x.

[36] N. C. Berchtold, D. H. Cribbs, P. D. Coleman, J. Rogers, E. Head, R. Kim, T. Beach, et al. Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proceedings of the National Academy of Sciences*, 105(40):15605–15610, 2008, doi:10.1073/pnas.0806883105.

[37] I. H. Salas, J. Burgado, and N. J. Allen. Glia: victims or villains of the aging brain? *Neurobiology of Disease*, 143:105008, 2020, doi:10.1016/j.nbd.2020.105008.

[38] R. Peters. Ageing and the brain. *Postgraduate Medical Journal*, 82(964):84–88, 2006, doi:10.1136/pgmj.2005.036665.

[39] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary. Brain age and other bodily ‘ages’: implications for neuropsychiatry. *Molecular Psychiatry*, 24(2):266–281, 2018, doi:10.1038/s41380-018-0098-1.

[40] Role of neuroinflammation in neurodegenerative diseases (Review). Apr 2016. URL https://www.spandidos-publications.com/10.3892/mmr.2016.4948?text=fulltext. Online; accessed 3 September 2021.

[41] Dementia. URL https://www.who.int/en/news-room/fact-sheets/detail/dementia. Online; accessed 3 September 2021.

[42] Y. Mu and F. H. Gage. Adult hippocampal neurogenesis and its role in Alzheimers disease. *Molecular Neurodegeneration*, 6(1):85, 2011, doi:10.1186/1750-1326-6-85.

[43] G. W. V. Hoesen, B. T. Hyman, and A. R. Damasio. Entorhinal cortex pathology in Alzheimers disease. *Hippocampus*, 1(1):1–8, 1991, doi:10.1002/hipo.450010102.

[44] R. Xia and Z.-H. Mao. Progression of motor symptoms in Parkinson’s disease. *Neuroscience Bulletin*, 28(1):39–48, 2012, doi:10.1007/s12264-012-1050-z.

[45] J. A. Obeso, M. C. Rodríguez-Oroz, B. Benitez-Temino, F. J. Blesa, J. Guridi, C. Marin, and M. Rodriguez. Functional organization of the basal ganglia: Therapeutic implications for parkinsons disease. *Movement Disorders*, 23(S3), 2008, doi:10.1002/mds.22062.

[46] B. R. Foerster, R. C. Welsh, and E. L. Feldman. 25 years of neuroimaging in amyotrophic lateral sclerosis. *Nature Reviews Neurology*, 9(9):513–524, 2013, doi:10.1038/nrneurol.2013.153.

[47] L. P. Rowland and N. A. Shneider. Amyotrophic Lateral Sclerosis. *New England Journal of Medicine*, 344(22):1688–1700, 2001, doi:10.1056/nejm200105313442207.

[48] FDA’s Decision to Approve New Treatment for Alzheimer’s Disease. URL https://www.fda.gov/drugs/news-events-human-drugs/fdas-decision-approve-new-treatment-alzheimers-disease. Online; accessed 12 October 2021.

[49] C. L. Masters, R. Bateman, K. Blennow, C. C. Rowe, R. A. Sperling, and J. L. Cummings. Alzheimers disease. *Nature Reviews Disease Primers*, 1(1), 2015, doi:10.1038/nrdp.2015.56.

[50] F. Corlier, G. Hafzalla, J. Faskowitz, L. H. Kuller, J. T. Becker, O. L. Lopez, P. M. Thompson, et al. Systemic inflammation as a predictor of brain aging: Contributions of physical activity, metabolic risk, and genetic risk. *NeuroImage*, 172:118–129, 2018, doi:10.1016/j.neuroimage.2017.12.027.

[51] A. Salminen, K. Kaarniranta, and A. Kauppinen. ER stress activates immunosuppressive network: implications for aging and Alzheimer’s disease. *Journal of Molecular Medicine*, 98(5):633–650, 2020, doi:10.1007/s00109-020-01904-z.

[52] F. Andromidas, S. Atashpanjeh, A. J. Myers, B. E. Mackinnon, M. M. Shaffer, and A. O. Koob. The astrogenic balance in the aging brain. *Current Neuropharmacology*, 19, 2021, doi:10.2174/1570159x19666210420095118.

[53] NCI Dictionary of Cancer Terms URL `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/astrocytoma`. Online; accessed 17 October 2021.

[54] R. Gargini, B. Segura-Collar, B. Herránz, V. García-Escudero, A. Romero-Bravo, F. J. Núñez, D. García-Pérez, et al. The IDH-TAU-EGFR triad defines the neovascular landscape of diffuse gliomas. *Science Translational Medicine*, 12(527), 2020, doi:10.1126/scitranslmed.aax1501.

[55] World Population Prospects - Population Division URL `https://population.un.org/wpp/`. Online; accessed 4 September 2021.

[56] Global Health and Aging Report. https://www.who.int/ageing/publications/global_health.pdf, October 2011. Online; accessed 18 October 2021.

[57] V. L. Feigin, E. Nichols, T. Alam, M. S. Bannick, E. Beghi, and et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016. 18(5):459–480, may 2019, doi:10.1016/s1474-4422(18)30499-x.

[58] Indicator Metadata Registry Details - Disability-adjusted life years (DALYs) URL `https://www.who.int/data/gho/indicator-metadata-registry/imr-details/158`. Online; accessed 18 October 2021.

[59] R. D. Fields, A. Araque, H. Johansen-Berg, S.-S. Lim, G. Lynch, K.-A. Nave, M. Nedergaard, et al. Glial biology in learning and cognition. *The Neuroscientist*, 20(5):426–431, 2013, doi:10.1177/1073858413504465.

[60] H. Tabata. Diverse subtypes of astrocytes and their development during corticogenesis. *Frontiers in Neuroscience*, 9, 2015, doi:10.3389/fnins.2015.00114.

[61] E. M. Hol, R. F. Roelofs, E. Moraal, M. A. F. Sonnemans, J. A. Sluijs, E. A. Proper, P. N. E. D. Graan, et al. Neuronal expression of GFAP in patients with Alzheimer pathology and identification of novel gfap splice forms. *Molecular Psychiatry*, 8(9):786–796, 2003, doi:10.1038/sj.mp.4001379.

[62] Y. Kim, J. Park, and Y. K. Choi. The role of astrocytes in the central nervous system focused on BK channel and heme oxygenase metabolites: A review. *Antioxidants*, 8(5):121, 2019, doi:10.3390/antiox8050121.

[63] F. Vasile, E. Dossi, and N. Rouach. Human astrocytes: structure and functions in the healthy brain. *Brain Structure and Function*, 222(5):2017–2029, 2017, doi:10.1007/s00429-017-1383-5.

[64] C.-Y. Liu, Y. Yang, W.-N. Ju, X. Wang, and H.-L. Zhang. Emerging roles of astrocytes in neuro-vascular unit and the tripartite synapse with emphasis on reactive gliosis in the context of Alzheimer's Disease. *Frontiers in Cellular Neuroscience*, 12, 2018, doi:10.3389/fncel.2018.00193.

[65] C. Escartin, E. Galea, A. Lakatos, J. P. O'Callaghan, G. C. Petzold, A. Serrano-Pozo, C. Steinhäuser, et al. Reactive astrocyte nomenclature, definitions, and future directions. *Nature Neuroscience*, 24(3):312–325, 2021, doi:10.1038/s41593-020-00783-4.

[66] M. Pekny and M. Pekna. Reactive gliosis in the pathogenesis of CNS diseases. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1862(3):483–491, 2016, doi:10.1016/j.bbadis.2015.11.014.

[67] S. A. Liddelow, K. A. Guttenplan, L. E. Clarke, F. C. Bennett, C. J. Bohlen, L. Schirmer, M. L. Bennett, et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*, 541(7638):481–487, 2017, doi:10.1038/nature21029.

[68] S. A. Liddelow and B. A. Barres. Reactive astrocytes: Production, function, and therapeutic potential. *Immunity*, 46(6):957–967, 2017, doi:10.1016/j.immuni.2017.06.006.

[69] Two Families of Postsynaptic Receptors. Jan 1970. URL `https://www.ncbi.nlm.nih.gov/books/NBK10855/`.

[70] S. Guerra-Gomes, N. Sousa, L. Pinto, and J. F. Oliveira. Functional roles of astrocyte calcium elevations: From synapses to behavior. *Frontiers in Cellular Neuroscience*, 11, 2018, doi:10.3389/fncel.2017.00427.

[71] R. D. Fields. Release of neurotransmitters from glia. *Neuron Glia Biology*, 6(3):137–139, 2010, doi:10.1017/s1740925x11000020.

[72] K. Guimarães, D. Q. Madureira, and A. L. Madureira. Interactions between astrocytes and the reward-attention circuit: A model for attention focusing in the presence of nicotine. *Cognitive Systems Research*, 50:15–28, 2018, doi:10.1016/j.cogsys.2018.03.001.

[73] A. Armada-Moreira, J. I. Gomes, C. C. Pina, O. K. Savchak, J. Gonçalves-Ribeiro, N. Rei, S. Pinto, et al. Going the extra (synaptic) mile: Excitotoxicity as the road toward neurodegenerative diseases. *Frontiers in Cellular Neuroscience*, 14, 2020, doi:10.3389/fncel.2020.00090.

[74] A. N. Early, A. A. Gorman, L. J. V. Eldik, A. D. Bachstetter, and J. M. Morganti. Effects of advanced age upon astrocyte-specific responses to acute traumatic brain injury in mice. *Journal of Neuroinflammation*, 17(1), 2020, doi:10.1186/s12974-020-01800-w.

[75] T. N. Bhatia, D. B. Pant, E. A. Eckhoff, R. N. Gongaware, T. Do, D. F. Hutchison, A. M. Gleixner, et al. Astrocytes do not forfeit their neuroprotective roles after surviving intense oxidative stress. *Frontiers in Molecular Neuroscience*, 12, 2019, doi:10.3389/fnmol.2019.00087.

[76] F. Chaudhry, J. Isherwood, T. Bawa, D. Patel, K. Gurdziel, D. E. Lanfear, D. M. Ruden, et al. Single-cell RNA sequencing of the cardiovascular system: New looks for old diseases. *Frontiers in Cardiovascular Medicine*, 6, 2019, doi:10.3389/fcvm.2019.00173.

[77] N. Habib, C. Mccabe, S. Medina, M. Varshavsky, D. Kitsberg, R. Dvir-Szternfeld, G. Green, et al. Disease-associated astrocytes in Alzheimer's disease and aging. *Nature Neuroscience*, 23(6):701–706, 2020, doi:10.1038/s41593-020-0624-8.

[78] L. Soreq, J. Rose, E. Soreq, J. Hardy, D. Trabzuni, M. R. Cookson, C. Smith, et al. Major shifts in glial regional identity are a transcriptional hallmark of human brain aging. *Cell Reports*, 18(2):557–570, 2017, doi:10.1016/j.celrep.2016.12.011.

[79] Transcriptome: Connecting the Genome to Gene Function. 2008. URL `https://www.nature.com/scitable/topicpage/transcriptome-connecting-the-genome-to-gene-function-605/`. Online; accessed 25 October 2021.

[80] T. V. Lanz, A.-K. Pröbstel, I. Mildenberger, M. Platten, and L. Schirmer. Single-cell high-throughput technologies in cerebrospinal fluid research and diagnostics. *Frontiers in Immunology*, 10, 2019, doi:10.3389/fimmu.2019.01302.

[81] T. Tammela and J. Sage. Investigating tumor heterogeneity in mouse models. *Annual Review of Cancer Biology*, 4(1):99–119, 2020, doi:10.1146/annurev-cancerbio-030419-033413.

[82] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1), 2017, doi:10.1186/s13073-017-0467-4.

[83] E. Papalexi and R. Satija. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45, 2017, doi:10.1038/nri.2017.76.

[84] R. Salomon, D. Kaczorowski, F. Valdes-Mora, R. E. Nordon, A. Neild, N. Farbehi, N. Bartonicek, et al. Droplet-based single cell RNAseq tools: a practical guide. *Lab on a Chip*, 19(10):1706–1727, 2019, doi:10.1039/c8lc01239c.

[85] Single-Cell RNA Sequencing FAQs. URL `https://web.genewiz.com/single-cell-faq`. Online; accessed 16 September 2021.

[86] Linked Read Sequencing – 10X Genomics Chromium Technology URL `https://dnatech.genomecenter.ucdavis.edu/linked-read-sequencing-10x-genomics-gemcode/`. Online; accessed 16 September 2021.

[87] Single-Cell RNA-Seq: An Introductory Overview and Tools for Getting Started. URL `https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started`. Online; accessed 16 September 2021.

[88] P. Qiu. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, 11(1), 2020, doi:10.1038/s41467-020-14976-9.

[89] Analysis of single cell RNA-seq data. Aug 2021. URL `https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html`. Online; accessed 22 September 2021.

[90] J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology*, 38(6):737–746, 2020, doi:10.1038/s41587-020-0465-8.

[91] S. R. Krishnaswami, R. V. Grindberg, M. Novotny, P. Venepally, B. Lacar, K. Bhutani, S. B. Linker, et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nature Protocols*, 11(3):499–524, 2016, doi:10.1038/nprot.2016.015.

[92] M. Slyper, C. B. M. Porter, O. Ashenberg, J. Waldman, E. Drokhlyansky, I. Wakiro, C. Smillie, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature Medicine*, 26(5):792–802, 2020, doi:10.1038/s41591-020-0844-1.

[93] M. R. Alkaslasi, Z. E. Piccus, S. Hareendran, H. Silberberg, L. Chen, Y. Zhang, T. J. Petros, et al. Single nucleus RNA-sequencing defines unexpected diversity of cholinergic neuron types in the adult mouse spinal cord. *Nature Communications*, 12(1), 2021, doi:10.1038/s41467-021-22691-2.

[94] snRNA-Seq URL `https://www.illumina.com/science/sequencing-method-explorer/kits-and-arrays/snrna-seq.html?fbclid=IwAR1E1JTnfr2daJxKcsnQEyQ8e8LhusGIE6ypd4GJ2o09hVagaXgf2EboREE`. Online; accessed 6 October 2021.

[95] M. Alvarez, E. Rahmani, B. Jew, K. M. Garske, Z. Miao, J. N. Benhammou, C. J. Ye, et al. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier diem. *Scientific Reports*, 10(1), 2020, doi:10.1038/s41598-020-67513-5.

[96] Gene Expression Omnibus URL `https://www.ncbi.nlm.nih.gov/geo/`. Online; accessed 20 September 2021.

[97] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, May 2013, doi:10.1038/ng.2653.

[98] The R Project for Statistical Computing URL `https://www.r-project.org/`. Online; accessed 21 September 2021.

[99] Open source & professional software for data science teamsAug 2021. URL `https://www.rstudio.com/`. Online; accessed 21 September 2021.

[100] Create Elegant Data Visualisations Using the Grammar of Graphics URL `https://ggplot2.tidyverse.org/`. Online; accessed 21 September 2021.

[101] R. Amezquita, A. Lun, E. Becht, V. Carey, L. Carpp, L. Geistlinger, F. Marini, et al. Orchestrating single-cell analysis with bioconductor. *Nature Methods*, 17:137–145, 2020, doi:10.1038/s41592-019-0654-x.

[102] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. M. III, S. Zheng, A. Butler, M. J. Lee, et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021, doi:10.1016/j.cell.2021.04.048.

[103] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016, doi:10.1098/rsta.2015.0202.

[104] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, pages 2579–2605.

[105] How to Use t-SNE Effectively. Oct 2016. URL `https://distill.pub/2016/misread-tsne/?_ga=2.135835192.888864733.1531353600-1779571267.1531353600`. Online; accessed 11 September 2021.

[106] Statistics Archieven: A step-by-step guide to statistical analysis URL `https://www.scribbr.com/category/statistics/`. Online; accessed 12 September 2021.

[107] S. Bhalerao and P. Kadam. Sample size calculation. *International Journal of Ayurveda Research*, 1(1):55, 2010, doi:10.4103/0974-7788.59946.

[108] Value and Statistical Significance: Simply Psychology. URL `https://www.simplypsychology.org/p-value.html`. Online; accessed 12 September 2021.

[109] R. B. Dell, S. Holleran, and R. Ramakrishnan. Sample size determination. *ILAR Journal*, 43(4):207–213, 2002, doi:10.1093/ilar.43.4.207.

[110] 2 Introduction to single-cell RNA-seq: Analysis of single cell RNA-seq data. Jul 2019. URL `https://scrnaseq-course.cog.sanger.ac.uk/website/introduction-to-single-cell-rna-seq.html`. Online; accessed 22 September 2021.

[111] Orchestrating Single-Cell Analysis with Bioconductor URL `http://bioconductor.org/books/release/OSCA/quality-control.html#choice-of-qc-metrics`. Online; accessed 15 September 2021.

[112] Analysis of single cell RNA-seq data. Jul 2019. URL `https://scrnaseq-course.cog.sanger.ac.uk/website/cleaning-the-expression-matrix.html`. Online; accessed 15 September 2021.

[113] P.-L. Germain. *scDblFinder: scDblFinder*, 2021. R package version 1.7.5.

[114] C. S. Mcginnis, L. M. Murrow, and Z. J. Gartner. Doubletfinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Systems*, 8(4), 2019, doi:10.1016/j.cels.2019.03.003.

[115] N. M. Xi and J. J. Li. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Systems*, 12(2), 2021, doi:10.1016/j.cels.2020.11.008.

[116] C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1), 2019, doi:10.1186/s13059-019-1874-1.

[117] A. T. L. Lun, D. J. McCarthy, and J. C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Res.*, 5:2122, 2016, doi:10.12688/f1000research.9501.2.

[118] J. T. Leek, W. E. Johnson, H. S. Parker, E. J. Fertig, A. E. Jaffe, Y. Zhang, J. D. Storey, et al. *sva: Surrogate Variable Analysis*, 2021. R package version 3.40.0.

[119] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015, doi:10.1093/nar/gkv007.

[120] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, et al. Comprehensive integration of single-cell data. *Cell*, 177(7), 2019, doi:10.1016/j.cell.2019.05.031.

[121] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21(1), 2020, doi:10.1186/s13059-019-1850-9.

[122] Find DEGs after doing integration - Issue #1256 - satijalab/seurat. URL `https://github.com/satijalab/seurat/issues/1256`. Online; accessed 14 September 2021.

[123] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008, doi:10.1088/1742-5468/2008/10/p10008.

[124] A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, 2020, doi:10.12688/f1000research.15666.3.

[125] A. T. Mckenzie, M. Wang, M. E. Hauberg, J. F. Fullard, A. Kozlenkov, A. Keenan, Y. L. Hurd, et al. Brain cell type specific gene expression and co-expression network architectures. *Scientific Reports*, 8(1), 2018, doi:10.1038/s41598-018-27293-5.

[126] M. D. Robinson, D. J. Mccarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009, doi:10.1093/bioinformatics/btp616.

[127] C. W. Law, M. Alhamdoosh, S. Su, X. Dong, L. Tian, G. K. Smyth, and M. E. Ritchie. RNA-seq analysis is easy as 1-2-3 with limma, glimma and edgeR. *F1000Research*, 5:1408, 2018, doi:10.12688/f1000research.9005.3.

[128] J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1), 2021, doi:10.1038/s41467-021-25960-2.

[129] W. Chang and B. Borges Ribeiro. *shinydashboard: Create Dashboards with 'Shiny'*, 2018. R package version 0.7.1.

[130] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005, doi:10.1073/pnas.0506580102.

[131] G. Korotkevich, V. Sukhov, and A. Sergushichev. Fast gene set enrichment analysis. 2019, doi:10.1101/060012.

[132] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014, doi:10.1038/nbt.2859.

[133] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, page 477, 2018, doi:10.1186/s12864-018-4772-0.

[134] B. P. de Almeida, N. Saraiva-Agostinho, and N. L. Barbosa-Morais. *cTRAP: Identification of candidate causal perturbations from differential gene expression data*, 2021. https://nuno-agostinho.github.io/cTRAP, https://github.com/nuno-agostinho/cTRAP.

[135] Connectivity Map (CMAP)Jul 2018. URL `https://www.broadinstitute.org/connectivity-map-cmap`. Online; accessed 15 September 2021.

[136] S. M. Strittmatter. Overcoming drug development bottlenecks with repurposing: Old drugs learn new tricks. *Nature Medicine*, 20(6):590–591, 2014, doi:10.1038/nm.3595.

[137] X. Zhang, R. Xu, C. Zhang, Y. Xu, M. Han, B. Huang, A. Chen, et al. Trifluoperazine, a novel autophagy inhibitor, increases radiosensitivity in glioblastoma by impairing homologous recombination. *Journal of Experimental & Clinical Cancer Research*, 36(1), 2017, doi:10.1186/s13046-017-0588-z.

[138] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, 2019, doi:10.1038/s41587-019-0114-2.

[139] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, 2015, doi:10.1038/nmeth.3337.

[140] CIBERSORTx URL `https://cibersortx.stanford.edu/`. Online; accessed 17 September 2021.

[141] K. Li, J. Li, J. Zheng, and S. Qin. Reactive astrocytes in neurodegenerative diseases. *Aging and disease*, 10(3):664, 2019, doi:10.14336/ad.2018.0720.

[142] GeneCards: The Human Gene Database URL `https://www.genecards.org/`. Online; accessed 28 September 2021.

[143] A. Grubman, G. Chew, J. F. Ouyang, G. Sun, X. Y. Choo, C. Mclean, R. Simmons, et al. A single cell brain atlas in human Alzheimer's disease. 2019, doi:10.1101/628347.

[144] A. Sharif, V. Duhem-Tonnelle, C. Allet, M. Baroncini, A. Loyens, J. Kerr-Conte, F. Collier, et al. Differential erbb signaling in astrocytes from the cerebral cortex and the hypothalamus of the human brain. *Glia*, 57(4):362–379, 2009, doi:10.1002/glia.20762.

[145] P. Preman, M. Alfonso-Triguero, E. Alberdi, A. Verkhratsky, and A. M. Arranz. Astrocytes in Alzheimer's disease: Pathological significance and molecular pathways. *Cells*, 10(3):540, 2021, doi:10.3390/cells10030540.

[146] L. Zhu, Y. Dou, M. Bjorner, and W. Ladiges. Development of a cyclophosphamide stress test to predict resilience to aging in mice. *GeroScience*, 42(6):1675–1683, 2020, doi:10.1007/s11357-020-00222-z.

[147] Slc1a2 MGI Mouse Gene Detail - MGI:101931 - solute carrier family 1 (glial high affinity glutamate transporter), member 2 URL `http://www.informatics.jax.org/marker/MGI:101931`. Online; accessed 28 September 2021.

[148] Nrxn1 MGI Mouse Gene Detail - MGI:1096391 - neurexin I URL `http://www.informatics.jax.org/marker/MGI:1096391`. Online; accessed 28 September 2021.

[149] Trifluoperazine: MedlinePlus Drug Information URL `https://medlineplus.gov/druginfo/meds/a682121.html`. Online; accessed 28 September 2021.

[150] Y. Zhang, D. T. Nguyen, E. M. Olzomer, G. P. Poon, N. J. Cole, A. Puvanendran, B. R. Phillips, et al. Rescue of pink1 deficiency by stress-dependent activation of autophagy. *Cell Chemical Biology*, 24(4), 2017, doi:10.1016/j.chembiol.2017.03.005.

[151] W. Chen, R. A. Mook, R. T. Premont, and J. Wang. Niclosamide: Beyond an antihelminthic drug. *Cellular Signalling*, 41:89–96, 2018, doi:10.1016/j.cellsig.2017.04.001.

[152] J. Deitrick and W. Pruitt. Wnt/beta catenin-mediated signaling commonly altered in colorectal cancer. *Progress in Molecular Biology and Translational Science Molecular and Cellular Changes in the Cancer Cell*, page 49–68, 2016, doi:10.1016/bs.pmbts.2016.09.010.

[153] A. Serrano, S. Apolloni, S. Rossi, S. Lattante, M. Sabatelli, M. Peric, P. Andjus, et al. The S100A4 transcriptional inhibitor niclosamide reduces pro-inflammatory and migratory phenotypes of microglia: Implications for Amyotrophic Lateral Sclerosis. *Cells*, 8(10):1261, 2019, doi:10.3390/cells8101261.

[154] K. C. Bermea, E. A. Casillas, L. D. Morales, L. L. Valdez, B. B. Su, A. Tsin, and B. Cheng. Evidence of a neuroprotective function for niclosamide in human sh-sy5y neuroblastoma and rat PC12 neural cells. *Acta Scientific Neurology*, 3(9):85–94, 2020, doi:10.31080/asne.2020.03.0235.

[155] O. Cerles, E. Benoit, C. Chéreau, S. Chouzenoux, F. Morin, M.-A. Guillaumot, R. Coriat, et al. Niclosamide inhibits oxaliplatin neurotoxicity while improving colorectal cancer therapeutic response. *Molecular Cancer Therapeutics*, 16(2):300–311, 2016, doi:10.1158/1535-7163.mct-16-0326.

[156] Researchers at SickKids identify an anti-cancer drug as a candidate to inhibit the degeneration of neurons.Nov 2017. URL `https://can-acn.org/researchers-at-sickkids-identify-an-anti-cancer-drug-as-a-candidate-to-inhibit-the-degeneration-of-neurons/`. Online; accessed 28 September 2021.

[157] K. Feinberg, A. Kolaj, C. Wu, N. Grinshtein, J. R. Krieger, M. F. Moran, L. L. Rubin, et al. A neuroprotective agent that inactivates prodegenerative trka and preserves mitochondria. *Journal of Cell Biology*, 216(11):3655–3675, 2017, doi:10.1083/jcb.201705085.

[158] C. Hanna, K. M. Kurian, K. Williams, C. Watts, A. Jackson, R. Carruthers, K. Strathdee, et al. Pharmacokinetics, safety, and tolerability of olaparib and temozolomide for recurrent glioblastoma: results of the phase i oparatic trial. *Neuro-Oncology*, 22(12):1840–1850, 2020, doi:10.1093/neuonc/noaa104.

[159] E. Paldino, V. D'Angelo, D. Laurenti, C. Angeloni, G. Sancesario, and F. R. Fusco. Modulation of inflammasome and pyroptosis by olaparib, a PARP-1 inhibitor, in the R6/2 mouse model of huntington's disease. *Cells*, 9(10):2286, 2020, doi:10.3390/cells9102286.

[160] C. M. Oslowski and F. Urano. Measuring er stress and the unfolded protein response using mammalian tissue culture system. *The Unfolded Protein Response and Cellular Stress, Part B Methods in Enzymology*, page 71–92, 2011, doi:10.1016/b978-0-12-385114-7.00004-0.

[161] D. L. Longo and J. M. Drazen. Data sharing. *New England Journal of Medicine*, 374(3):276–277, 2016, doi:10.1056/nejme1516564.

[162] C. S. Greene, L. X. Garmire, J. A. Gilbert, M. D. Ritchie, and L. E. Hunter. Celebrating parasites. *Nature Genetics*, 49(4):483–484, 2017, doi:10.1038/ng.3830.

[163] Y. Park and C. S. Greene. A parasites perspective on data sharing. *GigaScience*, 7(11), 2018, doi:10.1093/gigascience/giy129.

[164] P. Nurse. Biology must generate ideas as well as data. *Nature*, 597(7876):305–305, 2021, doi:10.1038/d41586-021-02480-z.

Appendix A

Sample metadata and Number of Nuclei

TABLE A.1: Public metadata associated with each of the single-cell RNA sequencing data samples used in this work.

Dataset name (ID)	GEO sample ID	Sample name	Sex	Age	Cause of Death	Brain Area	Sequencing
youngoldish	GSM4206906_Control_647C	youngoldish4	M	69	NA	PFC	NovaSeq6000
	GSM4206907_Control_777C	youngoldish3	M	58	NA	PFC	NovaSeq6000
	GSM4206904_Control_598C-1	youngoldish2	M	56	NA	PFC	NovaSeq6000
	GSM4206903_Control_335C	youngoldish1	M	44	NA	PFC	NovaSeq6000
young	GSM4654473_Nuc-RM95-2	young4	F	24	not dead (epilepsy)	TC	HiSeq4000
	GSM4654471_Nuc-RM77-2	young3	M	7	not dead (epilepsy)	TC	HiSeq4000
	GSM4654469_Nuc-RM102-2	young2	F	20	not dead (epilepsy)	TC	HiSeq4000
	GSM4654467_Nuc-RM101-2	young1	F	50	not dead (epilepsy)	TC	HiSeq4000
oldish	GSM4848449_s03_09	oldish4	M	82	COPD w/ interstitial lung disease & emphysema	M_PFC	NovaSeq6000
	GSM4848450_s04_09	oldish3	F	79	Breast cancer	M_PFC	NovaSeq6000
	GSM4848448_s03_03	oldish2	M	77	Small cell lung cancer	M_PFC	NovaSeq6000
	GSM4848447_s01_01	oldish1	M	58	Coronary artery disease, endstage cardiomyopathy	M_PFC	NovaSeq6000
oldish	GSM4886763_IGF112653	old6	F	91	NA	EC	HiSeq4000
	GSM4886758_IGF112648	old5	M	81	NA	EC	HiSeq4000
	GSM4886765_IGF112655	old4	F	80	NA	EC	HiSeq4000
	GSM4886754_IGF112644	old3	M	77	NA	EC	HiSeq4000
	GSM4886759_IGF112649	old2	M	74	NA	EC	HiSeq4000
	GSM4886746_IGF112636	old1	M	73	NA	EC	HiSeq4000

TABLE A.2: Number of cells associated with each sample, in the main data filtering steps. Entries with (*) mark samples that were removed in the astrocytic data processing phase.

sample	**# total nuclei (raw)**	**# total nuclei after filtering**	**# total astrocytic nuclei**
young3	4896	4215	407
young2	5744	5432	197
young4	4911	4633	447
youngoldish1	61679	3269	250
young1	5491	5212	212
youngoldish2	23160	2352	352
oldish1	2530	2260	411
youngoldish3	4798	594	31*
youngoldish4	63513	6536	957
old1	2902	2789	935
old2	3720	3452	1494
oldish2	4885	4416	468
old3	4653	4151	1709
oldish3	2817	2708	465
old4	2379	2263	1181
old5	2673	2543	1139
oldish4	2341	2291	99*
old6	6095	5746	3070
TOTAL	**209187**	**64862**	**13694**

Appendix B

Neural cell-type gene markers

TABLE B.1: Number and percentage of each clusters' marker genes ($log_2FC > 0.25$ and adjusted p-value < 0.05) that are present in the list of known cell type markers (1000 for each cell type [125]), in the detailed analysis.

	Markers												
Cluster	**Neurons**		**Astrocytes**		**Oligodendrocytes**		**Microglia**		**Endothelial cells**		**Undefined**		**Total**
	#	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**
0	7	4.00	3	1.71	79	45.10	1	0.57	6	3.43	79	45.10	175
1	9	2.93	4	1.30	3	0.98	117	38.10	13	4.23	161	52.40	307
2	9	2.45	153	41.70	0	0.00	2	0.55	20	5.45	183	49.90	367
3	124	52.10	6	2.52	1	0.42	2	0.84	11	4.62	94	39.50	238
4	15	3.82	152	38.70	1	0.25	2	0.51	30	7.63	193	49.10	393
5	250	35.70	14	2.00	13	1.86	4	0.57	27	3.86	392	560	700
6	9	2.93	4	1.30	3	0.98	109	35.50	14	4.56	168	54.70	307
7	89	21.00	23	5.42	27	6.37	2	0.47	19	4.48	264	62.30	424
8	273	45.20	11	1.82	8	1.32	4	0.66	19	3.15	289	47.80	604
9	19	4.61	146	35.40	6	1.46	1	0.24	33	8.01	207	50.20	412
10	289	46.80	11	1.78	7	1.13	1	0.16	15	2.43	294	47.60	617
11	260	37.40	13	1.87	7	1.01	4	0.58	31	4.45	381	54.70	696
12	186	43.80	6	1.41	6	1.41	3	0.71	15	3.53	209	49.20	425
13	29	16.20	2	1.12	1	0.56	60	33.50	6	3.35	81	45.30	179
14	8	2.31	14	4.05	6	1.73	9	2.60	150	43.40	159	46.00	346
15	227	34.80	20	3.07	13	1.99	3	0.46	19	2.91	370	56.70	652
16	226	36.60	14	2.27	12	1.94	6	0.97	18	2.92	341	55.30	617

TABLE B.2: Number and percentage of each clusters' marker genes ($log_2FC > 0.25$ and adjusted p-value < 0.05) that are present in the list of known cell type markers (1000 for each cell type [125]), in the general analysis.

	Markers												
Cluster	**Neurons**		**Astrocytes**		**Oligodendrocytes**		**Microglia**		**Endothelial cells**		**Undefined**		**Total**
	#	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**
neuro	341	43.00	12	1.51	6	0.58	0	0.00	26	3.28	408	51.50	793
astro	17	3.92	160	36.90	2	0.46	2	0.46	34	7.83	219	50.50	434
oligo	7	4.00	3	1.71	79	45.10	1	0.57	6	3.43	79	45.10	175
micro	7	2.23	4	1.27	2	0.63	117	37.30	15	4.78	169	53.80	314
endo	8	2.31	14	4.05	6	1.73	9	2.60	150	43.40	159	46.00	346
undef	71	21.50	12	3.63	21	6.34	4	1.21	11	3.32	212	64.00	331

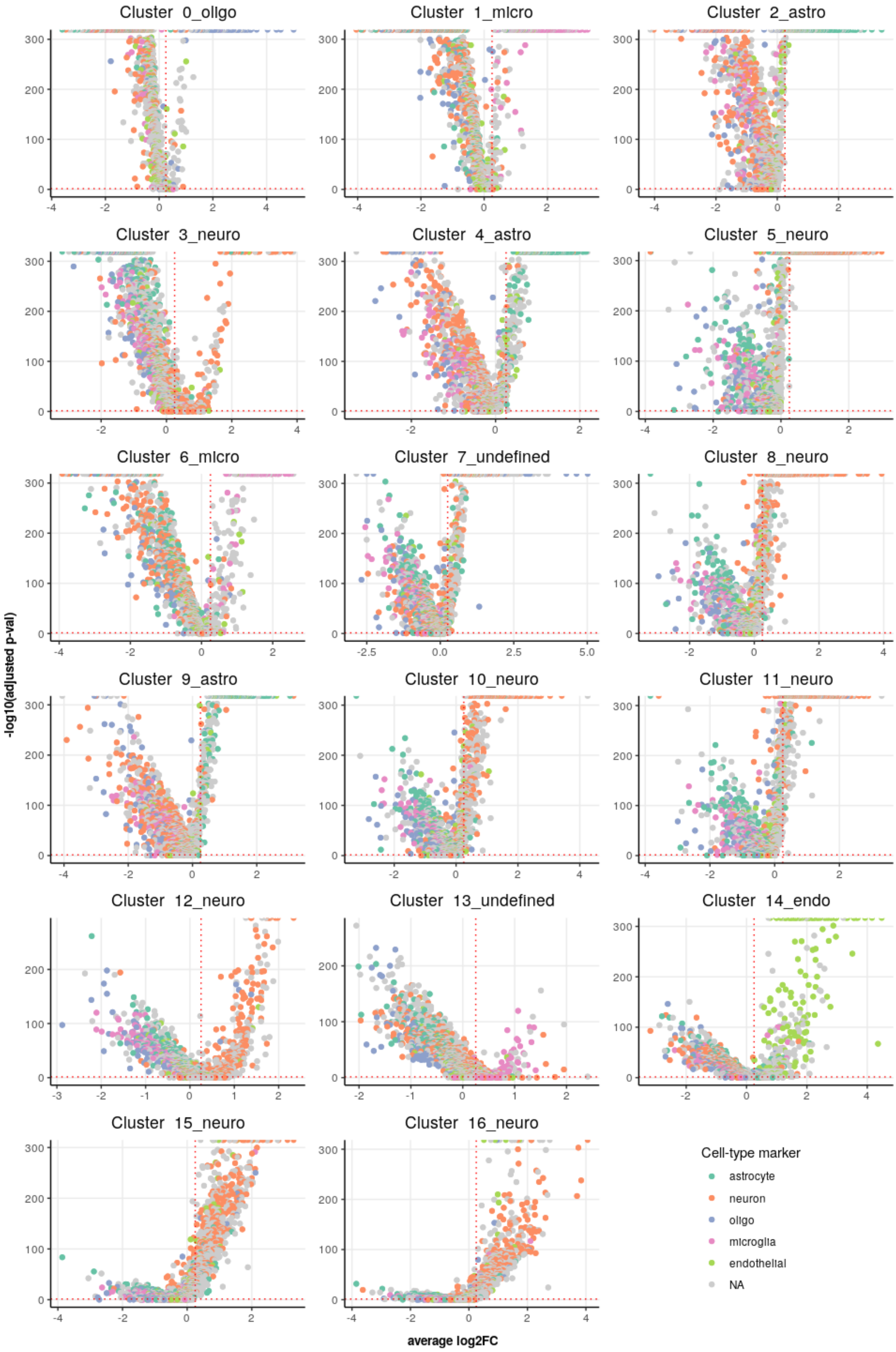


FIGURE B.1: Volcano plots of differential expression in each cluster compared to the others in the detailed analysis

Adjusted p-val $<$ 0.05 (Y-axis) and $log_2FC >$ 0.25 (X-axis), identified in dashed red lines, were used to define the marker genes of each cluster.

Appendix C

Gene markers of astrocytic clusters

TABLE C.1: 50 most up-regulated and 50 most down-regulated genes in cluster 4 compared to the average of the others.

Upregulated					Downregulated				
genes	**logFC**	**Avg Expr**	**t-stasistic**	**adj. p-val**	**genes**	**logFC**	**Avg Expr**	**t-statistic**	**adj. p-val**
NRXN1	-3,34	11,77	-103,61	0	**DCLK1**	2,77	7,4	81,16	0
GPC5	-2,45	12,01	-76,42	0	**LINC01088**	2,15	7,19	64,31	0
SLC1A2	-2,43	11,92	-69,57	0	**LINC00609**	2,1	7,66	59,8	0
CACNB2	-2,38	9,91	-65,38	0	**DPP10**	2,01	11,08	30,1	1,47E-190
CABLES1	-2,35	8,88	-66,18	0	**AC073941_1**	1,99	7,34	61,08	0
DNM3	-2,03	9,18	-58,69	0	**KAZN**	1,96	7,51	57,27	0
LINC00499	-2,01	9,18	-46,5	0	**CD44**	1,9	7,53	50,46	0
ARHGAP24	-1,95	8,74	-54,05	0	**TNC**	1,82	7,09	62,17	0
PDE4D	-1,94	9,7	-57,23	0	**SLC38A1**	1,74	6,96	65,77	0
CADM1	-1,77	10,87	-59,29	0	**ADAMTSL3**	1,7	6,84	72,39	0
SLC1A3	-1,72	11,48	-54,65	0	**TSHZ2**	1,65	6,86	67,27	0
GRM3	-1,72	8,34	-46,71	0	**VCAN**	1,6	7,43	51,51	0
GNA14	-1,64	9	-45,43	0	**SNED1**	1,52	7,44	46,51	0
ADGRV1	-1,58	11,66	-46,58	0	**GALNT15**	1,51	6,89	66,44	0
LRRC4C	-1,55	8,85	-38,88	3,06E-311	**LINC00836**	1,5	6,68	78,98	0
PTN	-1,54	8,41	-45,09	0	**PLEKHA5**	1,41	8,51	32,85	1,37E-225
SLC35F1	-1,49	8,77	-42,46	0	**TTN**	1,4	6,88	62,06	0
UNC5C	-1,47	8,01	-40,89	0	**GPC6**	1,38	7,01	48,72	0
GABRB1	-1,46	10,33	-44,65	0	**HSPA1A**	1,37	7,31	40,08	0
ADGRB3	-1,43	11,16	-49,97	0	**L3MBTL4**	1,37	7,09	48,98	0
CADM2	-1,43	11,42	-47,78	0	**NFASC**	1,37	7,12	53,66	0
LRRC3B	-1,42	8,59	-39,42	1,74E-319	**ADAMTS9_AS2**	1,3	7,61	35,82	2,49E-266
ADGRL3	-1,38	10,73	-47,18	0	**SYNPO2**	1,3	6,86	54,51	0
LRP1B	-1,37	12,03	-44,95	0	**KCNJ3**	1,27	6,78	56,29	0
CTNNA2	-1,34	12,39	-42,54	0	**GFAP**	1,24	9,09	29,3	6,77E-181
GPM6A	-1,31	12,26	-50,22	0	**STXBP5L**	1,23	6,81	53,17	0
PLCB1	-1,3	9,42	-35,53	3,21E-262	**MARCH3**	1,22	7,79	33,83	1,08E-238
SYNE1	-1,3	9,47	-42,01	0	**AQP1**	1,18	6,98	43,35	0
ZNF98	-1,27	8,25	-28,34	1,17E-169	**SLC24A4**	1,17	6,84	52,31	0
ERBB4	-1,26	11,48	-41,22	0	**CERS6**	1,16	7,54	37,69	2,43E-293
SLC4A4	-1,25	9,96	-36,68	1,20E-278	**MAN1C1**	1,14	7,26	35,42	1,49E-260
GLUL	-1,24	8,91	-35,69	1,92E-264	**HSPB8**	1,12	7,14	42,96	0
PCDH9	-1,23	12,72	-43,82	0	**UBC**	1,12	7,3	39,95	0
LINC00511	-1,23	8,62	-35,17	4,04E-257	**DGKB**	1,1	7,13	36,81	1,53E-280
CARMIL1	-1,22	9,44	-37,01	2,51E-283	**ROBO2**	1,07	6,82	43,4	0
PREX2	-1,22	9,81	-38,96	1,71E-312	**HSPB1**	1,04	7,41	30,86	4,91E-200
DNAH7	-1,21	8,54	-32,88	6,23E-226	**COL21A1**	1,04	7,29	34,29	6,25E-245
AC092691_1	-1,21	10,71	-31	8,94E-202	**ABI3BP**	1,03	6,9	41,5	0
MACROD2	-1,2	8,75	-30,28	7,44E-193	**CRYAB**	1,03	7,25	33,94	3,85E-240
AF279873_3	-1,18	8,05	-27,2	1,18E-156	**SLCO3A1**	1,02	6,74	48,65	0
RNF219_AS1	-1,17	10,12	-26,63	2,43E-150	**CP**	1,01	6,68	51,85	0
NHSL1	-1,17	9,46	-35,86	8,99E-267	**IGFBP5**	1	6,69	48,24	0
MERTK	-1,16	8,28	-38,12	9,85E-300	**WDR49**	0,99	7,65	29,41	3,28E-182
GRIA2	-1,16	8,24	-36,17	2,89E-271	**DAAM2**	0,99	7,87	31,01	7,06E-202
AL589740_1	-1,15	9,3	-25	6,97E-133	**FRY**	0,97	6,96	38,39	7,73E-304
TMEM132C	-1,14	7,73	-33,02	7,29E-228	**DPP6**	0,96	7,52	27,94	4,70E-165
RORB	-1,14	8,53	-34,16	3,48E-243	**HSP90AA1**	0,94	8,14	26,16	3,11E-145
TENM2	-1,12	9,3	-24,67	1,52E-129	**LINC01094**	0,94	7,29	30,25	2,03E-192
RYR3	-1,12	10,84	-35,13	1,57E-256	**PRRX1**	0,93	7,42	26,59	7,64E-150
TNIK	-1,12	10,68	-37,25	6,75E-287	**SMAD9**	0,91	6,91	38,89	1,93E-311

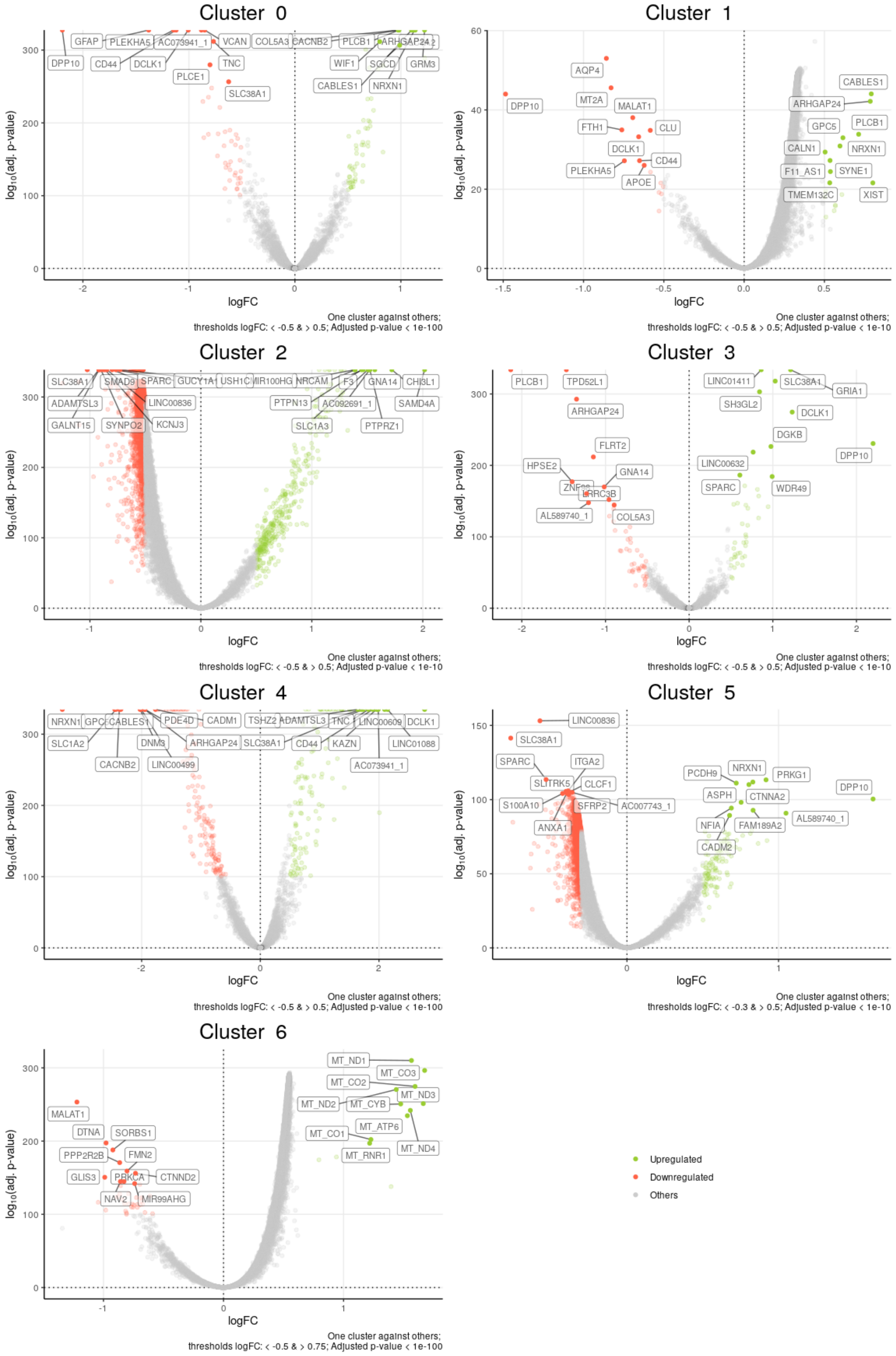


FIGURE C.1: DEA of one cluster against the remaining ones
Differential expression analysis of each cluster against the mean of all other clusters. Some genes with the highest absolute logFC are highlighted, at the extremes of each of the “volcano plots”.

Appendix D

Functional enrichment of the main principal components of astrocytic gene expression

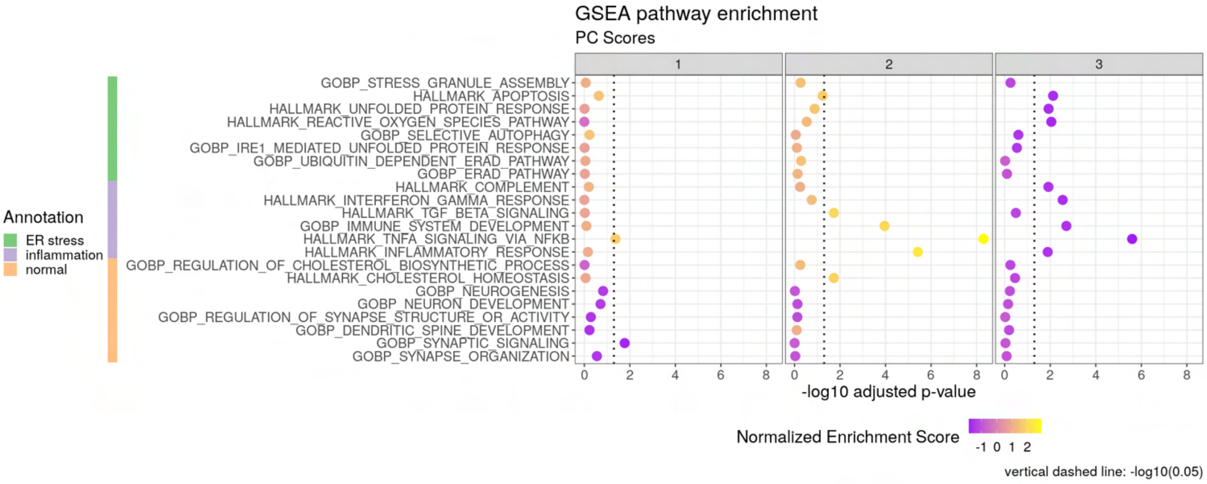
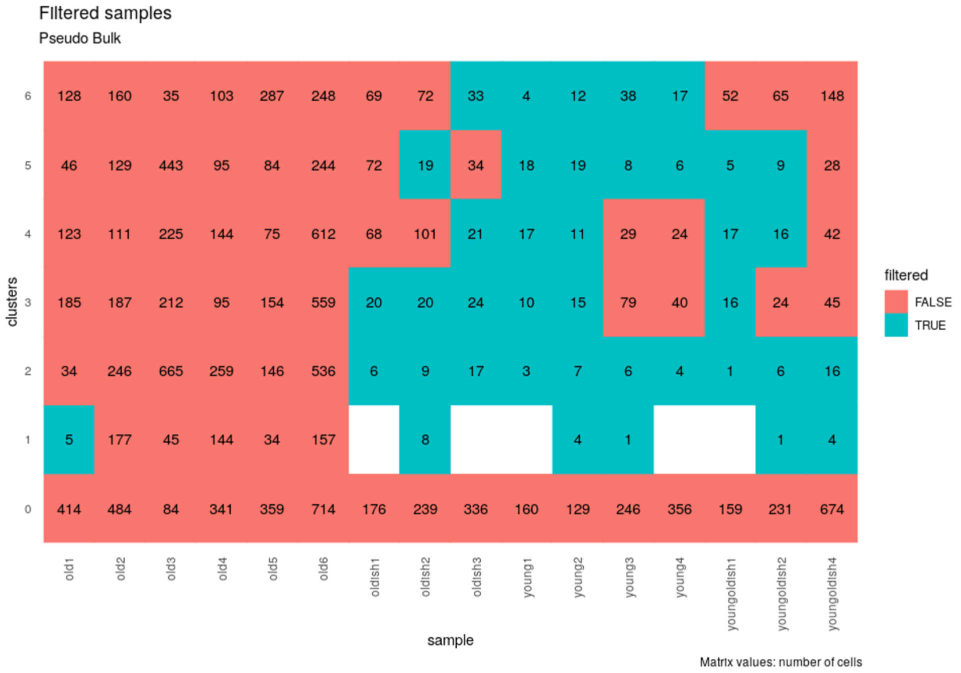


FIGURE D.1: GSEA of the three main PCs of astrocytic expression
Selection of the pathways, hallmarks and biological processes associated with ER stress, inflammation and normal astrocytic functions (neuronal support and synaptic homeostasis), and respective enrichment along the first three principal components (PCs) of astrocytic gene expression.

Appendix E

Analysis of pseudo-bulk astrocytic data



clusters	old1	old2	old3	old4	old5	old6	oldish1	oldish2	oldish3	young1	young2	young3	young4	youngoldish1	youngoldish2	youngoldish4
6	128	160	35	103	287	248	69	72	33	4	12	38	17	52	65	148
5	46	129	443	95	84	244	72	19	34	18	19	8	6	5	9	28
4	123	111	225	144	75	612	68	101	21	17	11	29	24	17	16	42
3	185	187	212	95	154	559	20	20	24	10	15	79	40	16	24	45
2	34	246	665	259	146	536	6	9	17	3	7	6	4	1	6	16
1	5	177	45	144	34	157		8			4	1			1	4
0	414	484	84	341	359	714	176	239	336	160	129	246	356	159	231	674

FIGURE E.1: Pseudo-bulk sample filtering criteria
Matrix showing the number of astrocytes from each cluster in each sample. Cluster/sample combinations having 20 or fewer cells (blue) were removed from the pseudo-bulk RNA-sequencing analysis, as to not compromise the normalization step. Three more samples (oldish3 and young3 from cluster 6, and oldish3 from cluster 3) were removed (also in blue), as they were compromising the normalization step. Blank entries correspond to samples that were not present in a specific cluster in the astrocytic scRNA-seq data.

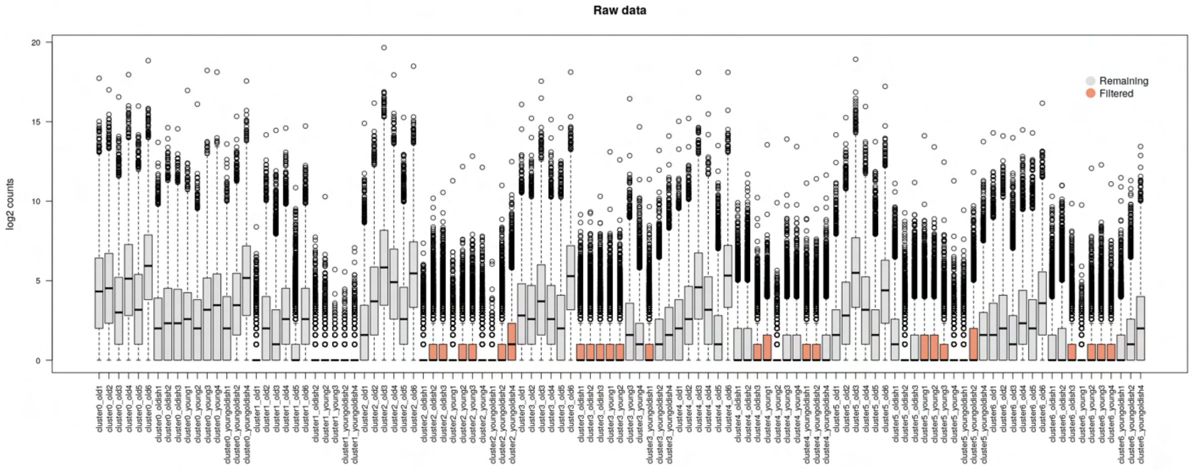


FIGURE E.2: Library sizes of pseudo-bulk samples
Boxplots summarising the distributions of read coverage across pseudo-bulk RNA-seq samples (removed samples - see figure E.1 - in orange)

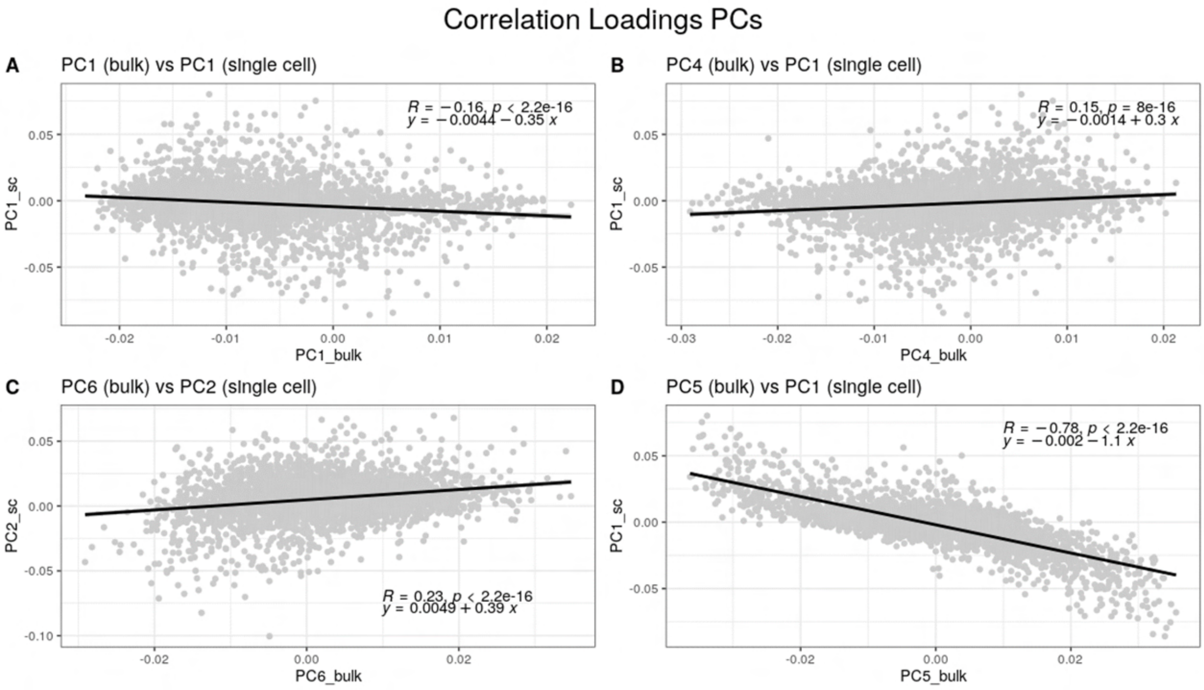


FIGURE E.3: Correlation of principal components from pseudo bulk versus single cell
Linear regression (black solid line) applied to relation between the loadings of each gene (grey dot) in the principal components of the pseudo-bulk and single-cell gene expression. There is a clear tendency for the pseudo-bulk PC5 to align with the single-cell PC1 (panel **D**).

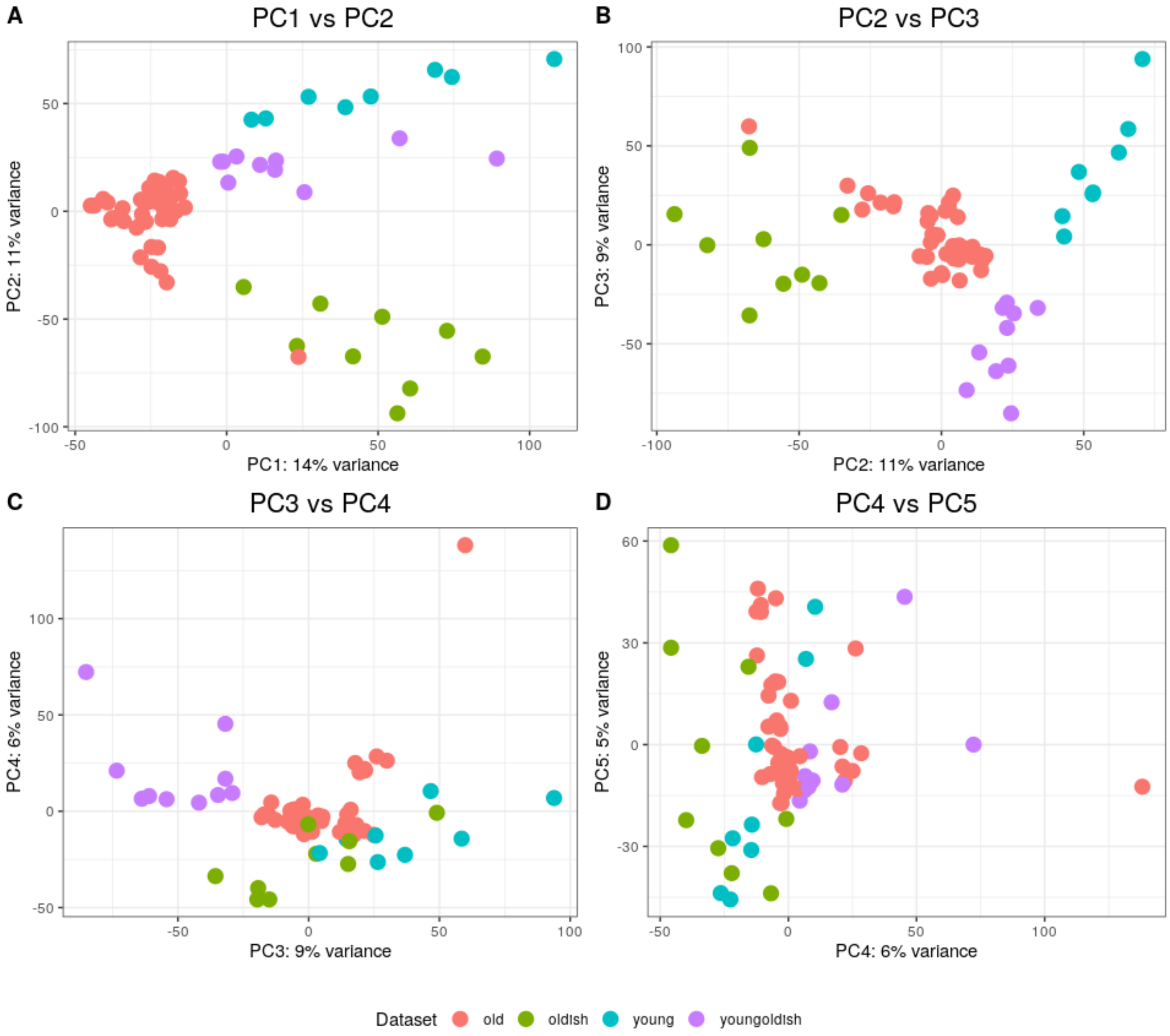


FIGURE E.4: Principal Components of non-filtered pseudo-bulk RNA-seq data
PCA of pseudo-bulk gene expression, resulting from pooling scRNA-seq data from all cells of each cluster from each individual, for **(A)** PC1 and PC2, **(B)** PC2 and PC3, **(C)** PC3 and PC4 and **(D)** PC4 and PC5

Appendix F

Cell Deconvolution

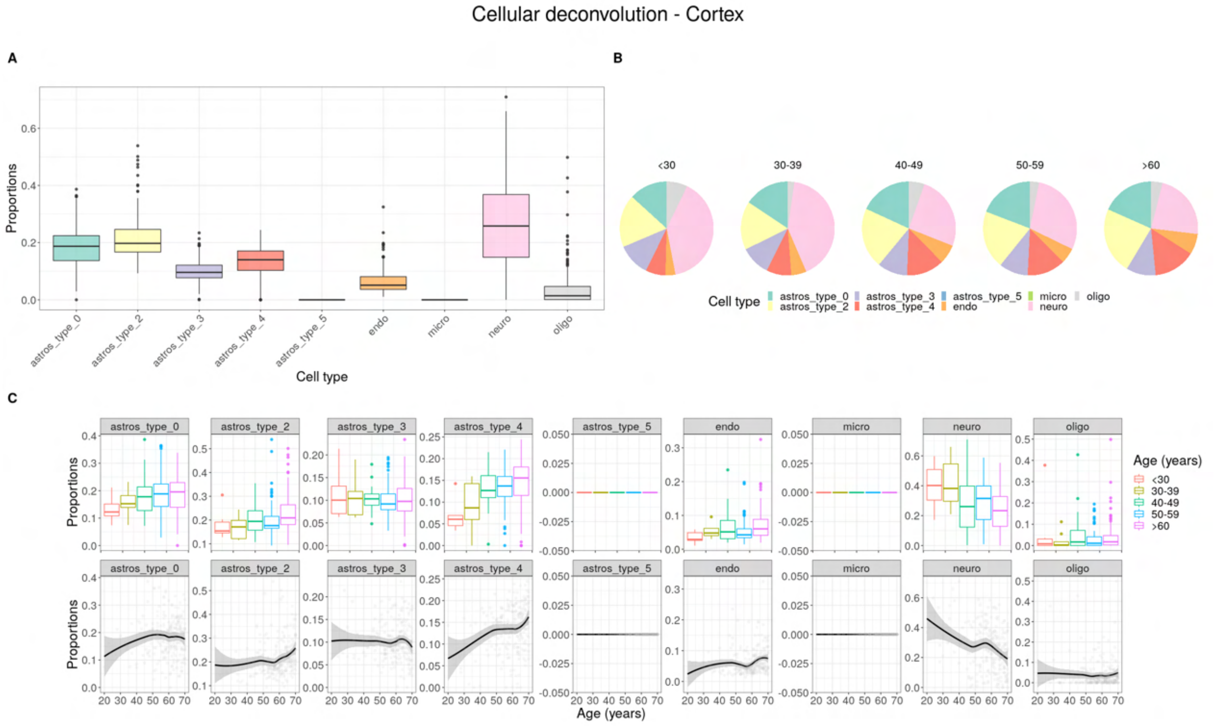


FIGURE F.1: Cell-type deconvolution of cortex samples
(A) Boxplots of distributions of CIBERSORTx estimates of proportions of the various cell types in this work, including types 0, 2, 3, 4 and 5 astrocytes, neurons, microglia (micro), oligodendrocytes (oligo) and endothelial cells (endo), in GTEx cortex samples. **(B)** Distribution, by age group, of the proportions of the various cell types. **(C)** Distribution of proportions, by age group and through a general additive model along age (R `geom_smooth` function with default parameters).

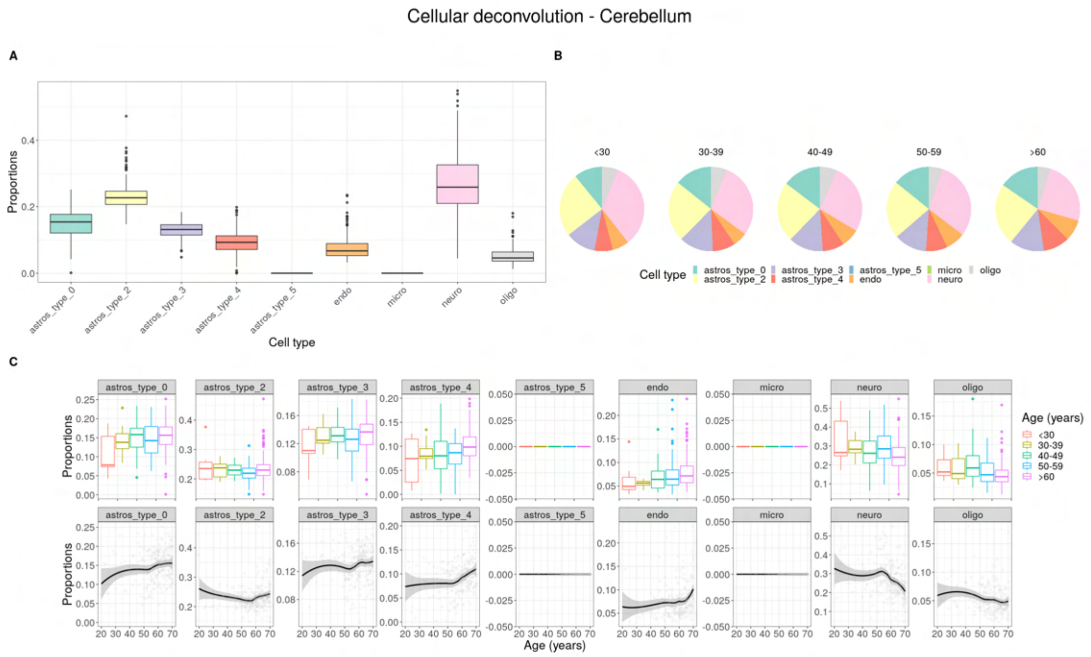


FIGURE F.2: Cell-type deconvolution of cerebellum samples
(A) Boxplots of distributions of CIBERSORTx estimates of proportions of the various cell types in this work, including types 0, 2, 3, 4 and 5 astrocytes, neurons, microglia (micro), oligodendrocytes (oligo) and endothelial cells (endo), in GTEx cerebellum samples. **(B)** Distribution, by age group, of the proportions of the various cell types. **(C)** Distribution of proportions, by age group and through a general additive model along age (R `geom_smooth` function with default parameters).

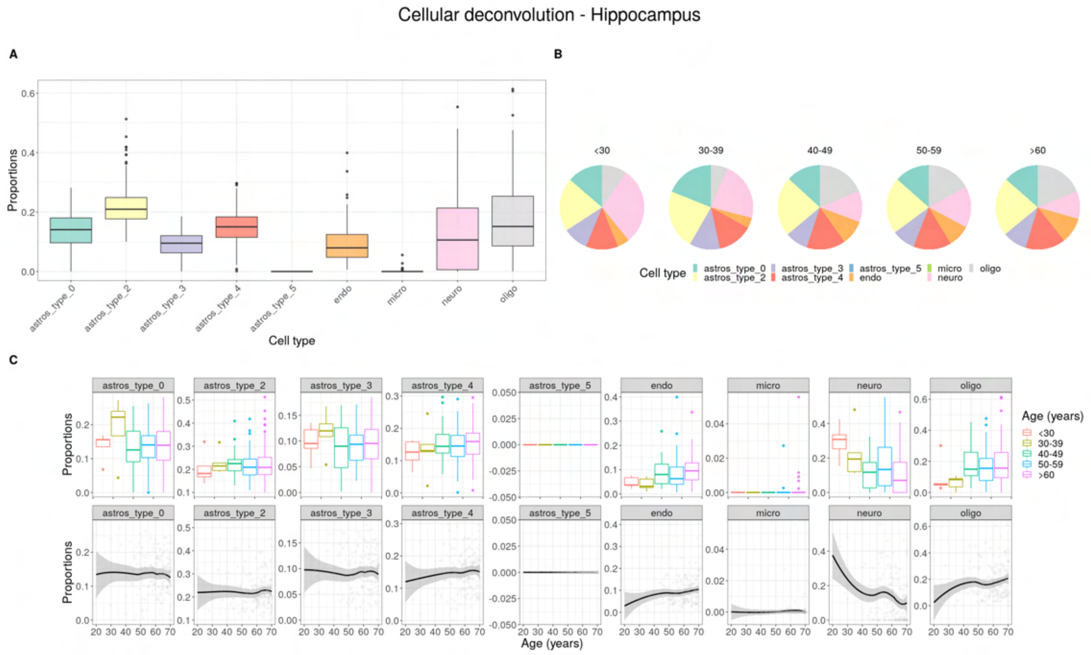


FIGURE F.3: Cell-type deconvolution of hippocampus samples
(A) Boxplots of distributions of CIBERSORTx estimates of proportions of the various cell types in this work, including types 0, 2, 3, 4 and 5 astrocytes, neurons, microglia (micro), oligodendrocytes (oligo) and endothelial cells (endo), in GTEx hippocampus samples. **(B)** Distribution, by age group, of the proportions of the various cell types. **(C)** Distribution of proportions, by age group and through a general additive model along age (R `geom_smooth` function with default parameters).