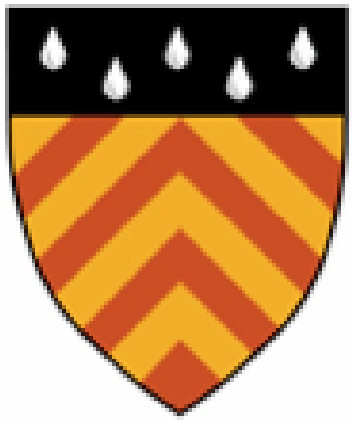
The Transcriptional Profile of Microglia: From Brain to Dish



Fiona Elizabeth Calvert

Clare Hall

December 2019

University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

Declaration of originality

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee

Fiona Calvert
December 2019

The Transcriptional Profile of Microglia: from Brain to Dish

Fiona Elizabeth Calvert

Microglia are the tissue resident macrophages of the central nervous system (CNS) and multiple lines of evidence indicate that microglia are a pathogenic cell type in Alzheimer's disease (AD). It is important to understand the transcriptional profiles of microglia, both from primary human cells and the *in-vitro* model systems used to study the cells at scale. In this thesis, I aim to build on previous small-scale studies of primary microglia and *in-vitro* model systems to answer three major questions: **1.** Can transcriptional data from fresh, primary human microglia be used to identify novel subpopulations of cells and understand how clinical phenotypes influence gene expression? **2.** How accurately do current simple *in-vitro* model systems of human microglia capture the profile of primary human cells? **3.** Do more complex model systems move cultured cells further along a trajectory towards the primary cell type?

I have utilised RNA-sequencing technology to build the most comprehensive transcriptional profile of primary human microglia to date, from over 100 neurosurgical patients. Using single-cell sequencing I have demonstrated that clinical pathology, particularly major trauma, causes specific gene expression changes within microglial transcriptomes. I have then shown that *in-vitro* models of primary microglia have significantly reduced expression of key marker genes and transcription factors, such as *P2RY12* and *SALL1*, when compared to primary cells. Using gene-set enrichment analysis tools, I have shown that many of the genes with higher expression in primary cells can be linked to neuronal processes such as CNS myelination. Data from the third chapter of this thesis identified the CNS environment as a major stimulating factor in the gene expression profile of primary microglia. Therefore, I used single cell analysis to understand how culturing stem cell derived microglia in the presence of neurons could move *in-vitro* systems closer towards the primary cell type. In summary, the work in this thesis has demonstrated that microglial transcriptomes are constantly reacting to stimuli within the local CNS environment, both to maintain their unique gene expression profiles and to respond to clinical conditions. I have also shown that current *in-vitro* model systems do not fully capture this transcriptional profile which largely appears to be driven by environmental stimuli within the CNS.

Acknowledgments

I sat staring at this page for a little while before writing this, I was unsure of how to put into words the thanks and gratitude I have for all the people who have helped me get to the point of completing this. I also have a tendency to overdo the soppy and felt pressured to write something witty and lighthearted. What follows, much like my PhD, I'm sure will be different from what I expected and planned but something that I will be proud of nonetheless.

First of all, I have to acknowledge Dan Gaffney, my ever wonderful and patient supervisor. I think this PhD has been a learning curve for both of us but I cannot stress enough how vital a part you have played in my journey here. You have pushed and challenged me in ways I couldn't have imagined when I started over four years ago. Throughout it all, you have also provided constant support and encouragement and have never made me feel like I couldn't do this. Thank you for taking a student who didn't even know the terminal existed and only knew how to perform a t-test and shaping me into an R-loving scientist with an annoying obsession for correct statistics. I would also like to acknowledge the wonderful lab you have built, a group of people who have been such a huge part of my PhD. To Andy, who acted as my supervisor within the lab and is truly one of the most remarkable scientists I have ever met. You made me a better scientist and opened my eyes to the wonderful world of moths and orchids. I hope you know that the Gaffney lab is only what it is, in part, because of you. To Julie and Clara, who I could not have survived the last couple of months without, thank you for listening to my gripes and moaning and for always giving the best life advice. To everyone else in the Gaffney lab (Natsuhiko, Nikos, Beata, Maria P, Maria I and Gerda) thank you for providing so many laughs and thoughtful discussions over lunch and coffee breaks. It has been a pleasure to work alongside you all and I hope our GIF-filled slack channels will live on forever!

To everyone I have had the pleasure of working with in these four years outside of the PhD - thank you for providing me with the most wonderful distractions when the PhD took its toll. To the members of staff at Clare Hall, who helped us put on an event that is one of my proudest achievements. The May Ball was a shining light of

my four years. To everyone I have met through the Story Collider, the producers and the storytellers, you are the most inspiring people I have ever met. You have changed the way I view science and have opened my eyes to a whole new wonderful world. Particularly to Erin, Liz and Steve - you have taught me so much in such a short space of time and I honestly feel privileged every day that I get to work with such ridiculously amazing people.

I could not have gotten to the point of finishing this PhD without my personal support network. To Lindsay, who I met at the Sanger interviews and knew we would be friends instantly - THANK YOU. You quickly became one of my best friends and I will always be grateful to have someone to cry in the toilets with. I write this full of pride that we both made it through, I wouldn't be here without you. To all my friends outside the PhD, both old and new: Emily, Becky, Emma, Madi, Guy, Luke, Sarah and Christy - thank you for providing wonderful relief from this process and filling my life with joy and excitement. My parents have always been, and I'm sure will continue to be, my biggest cheerleaders. I owe everything I have achieved to the strength and encouragement you have given me. I am so proud to be your daughter and don't know how to ever thank you for all you have done for me. I promise to only make you read this small part of my PhD and will not enforce proof-reading duties on you, even though I know you would if I asked. I love you both so much. Finally, the biggest acknowledgment of all. To Will, my long suffering partner who has been by my side for the majority of this PhD. You have had to listen to me complain about every failed experiment, every bug in my code and every frustration. You have seen me at my worst and provided me with every bit of support I need to pull myself out of those dark spaces. You made me laugh every day and have shown me how to not take life too seriously, something I desperately needed. You have been my rock, my escape and my source of happiness throughout every up and down of this PhD. I cannot thank you enough for being part of my life and for learning what microglia are for me.

Lindsay - we had many discussions about this very point so I include this as an acknowledgement of our journey. A large part of this acknowledgement goes to me, this thesis will forever be a reminder of what I can achieve even when I don't think I can.

Table of contents

Abbreviations

Alzheimer's disease	AD
Blood brain barrier	BBB
Central nervous system	CNS
Colony stimulating factor 1 receptor	CSF-1R
Embryoid body	EB
Expression quantitative trait loci	eQTL
Fluorescence-activated cell sorting	FACS
Genome-wide association studies	GWAS
Induced pluripotent stem cells	iPSCs
Knockout	KO
Late onset Alzheimer's disease	LOAD
Log fold change	LFC
Monocyte derived macrophages	MDMs
Multiple sclerosis	MS
Nuclease free water	NFW
Peripheral Blood Mononuclear Cells	PBMCs
Principal components analysis	PCA
Polymerase chain reaction	PCR
Quality control	QC
Quantile normalisation	QN
Single cell RNA-seq	scRNA-seq
Single nucleotide polymorphism	SNP
Traumatic brain injury	TBI
Transcription factor	TF
Transcripts per million	TPM
Uniform Manifold Approximation and Projection	UMAP
Variance stabilisation transformation	VST
Yolk sac	YS

Chapter 1: Introduction

1.1 Identification and characterisation of microglial cells in the brain

Microglia are the tissue resident macrophages of the central nervous system (CNS) and play an important role in it's immune defense[1]. Microglia were first described in the early 1900s, as scientists began to use developing microscopy techniques to study the brain. Santiago Ramón y Cajal, a Spanish neuroscientist famed for his descriptions and images of the CNS, dedicated much of his research to the non-neuronal cells within the brain, known as glial cells[2,3]. Within this glial cell population, Cajal identified the "third element" of the CNS describing the non-neuronal, non-astrocytic population of cells he observed. Río-Hortega divided this "third element" into two subdivisions: microglia and interfascicular glia, now known as oligodendrocytes[2]. Río-Hortega observed that microglia were relatively uniformly distributed in the brain, although noted a higher density in the grey matter, and described the cells as highly dynamic, often adapting their morphology to the features of the brain[3]. His later work focussed on microglial physiology following trauma to the brain where he described the cells taking on an ameboid shape and becoming highly phagocytic.

Since the early description of microglia, experimental tools have significantly improved and it is now easier to identify and observe microglial cells in a variety of systems, from primary cells across species to *in-vitro* models. Improved microscopy techniques have confirmed Rio-Hortega's initial observation that microglia have a highly ramified morphology (Figure 1.1), with dynamic processes that constantly survey the environment and maintain contact with neurons[4]. *In-vivo* time lapse imaging using zebrafish has suggested that this motility is not a random process[5] and that the cells are responding to ATP signals released from active neurons.

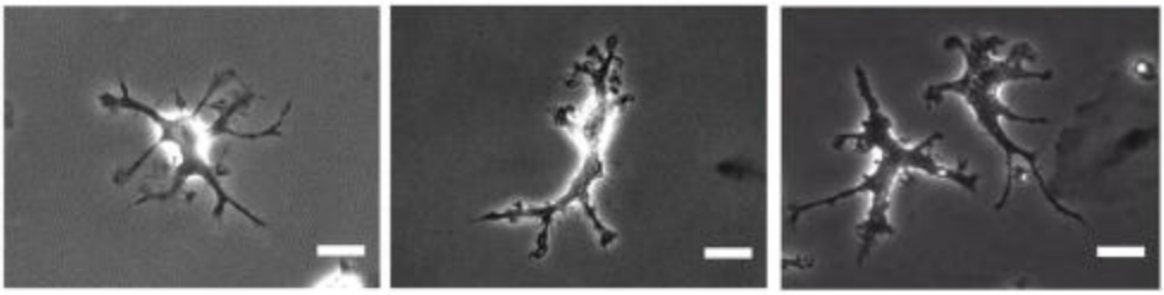


Figure 1.1 Microscopy images of mouse (left), fetal human (middle) and iPSC-derived microglia

Image taken from Muffat *et al*.[6], Figure 3 panel b.

In addition to describing the characteristics of microglial cells, Rio-Hortega was the first to theorise that microglia were of mesoderm origin[7]. For many years this theory was overlooked and instead it was argued that the cells were derived from neuro-ectoderm, along with other glial cell populations such as astrocytes[8–10]. However, evidence began to build that supported Rio-Hortega's original proposal: microglia were shown to have similar morphological features to macrophages[11] and were shown to express myeloid markers such as CD11b[12]. In mice, knockout (KO) of the *PU.1* gene, a key transcription factor (TF) in myeloid cell development, resulted in an absence of microglial populations in the brain[13].

1.2 Lineage of microglial populations in the brain

It is now well recognised that the microglial cells first described by Río-Hortega are tissue resident macrophages of the CNS. While the myeloid origin of these cells is no longer disputed, unique features of microglial development appear to distinguish them from other macrophage cells both in their initial origin and maintenance throughout adult life.

1.2.1 Microglial cell origin in embryonic development

Microglia-like cells have been identified in both rodent and human samples in the very early stages of embryonic development[14,15], suggesting they derive from a lineage independent of bone marrow hematopoiesis. In human fetal development,

Iba1+ (a myeloid cell marker) precursor cells have been observed in the developing nervous system as early as 4.5 gestational weeks[15], while hematopoietic stem cells don't seed the fetal liver until around gestational week 5[16].

Dissociation of fetal tissue samples from mice provided the first evidence that microglial progenitors are located in the yolk sac (YS) before moving into the developing brain as embryogenesis progresses[14]. More recently, a fate-mapping study has provided further evidence of the unique YS origin of microglial cells[17]. Fate-mapping relies on the ability to label cells from specific developmental origins and trace them through the developmental process. In the case of microglia, yellow fluorescent labelled protein (YFP) was linked to the *RUNX1* TF, which is specific to YS myeloid development. An estimated 32% of adult microglia cells were derived from YS precursors compared to only 3% of circulating monocytes. Specific erythro-myeloid progenitors within the mouse YS have since been identified[18] and it is these colony stimulating factor 1 receptor (CSF-1R) expressing-cells that appear to give rise to tissue resident macrophages such as microglia.

Mouse models have also been used to identify the pathways and molecules that regulate microglial differentiation from early progenitors. *Myb* is a TF which has previously been shown to be dispensable for yolk sac myelopoiesis but necessary for the creation of hematopoietic stem cells in the bone marrow. The initial production of microglia cells has been shown to be a *Myb* independent process[19,20], which further adds to the evidence behind the YS origin of microglia. Other TFs, like *PU.1* and *IRF8*, as well as protein coding genes, such as *MMP8* and *MMP9*, are required for the development of mature microglial cells[19,21]. The expression of CSF-1R by progenitor cells and a functional circulatory system is also necessary for microglial differentiation[17].

1.2.2 Maintenance of microglial populations throughout adulthood

The CNS has long been considered an "immune privileged" site, which limits immune reactions in the brain[22]. This, in part, is due to the presence of the blood brain barrier (BBB) that is thought to prevent circulating immune cells entering the brain. In most other tissues, circulating monocytes provide a progenitor cell for expanding

macrophage populations. It is known that even after the formation of the BBB, when monocytes theoretically cannot enter the brain, the population of microglia in the brain continues to grow with a large population surge two weeks after birth[14]. This evidence suggests that microglial cells have expansion potential and can self-maintain populations throughout adulthood. There are three proposed mechanisms for this continued growth of microglial populations: i) microglia are in fact replenished by circulating monocytes that cross the BBB, ii) there are populations of microglial progenitor cells that are present in the brain throughout life or iii) mature microglia themselves have the potential to proliferate.

Evidence for a significant contribution of circulating cells to the adult microglial population is controversial. Consistent with this hypothesis, *PU.1* KO mice lack any embryonically-derived microglia, but develop microglia-like cells within their CNS after receiving bone marrow transplants after birth[23]. However, fate-mapping studies have been used to demonstrate that up to 60% of microglia in adult mice are YS derived[20] and sublethal irradiation of mice followed by healthy hematopoietic cell transfer only gave rise to around 5% of donor derived microglia[17]. Parabiotic mouse models can be used to surgically join two mice and allow sharing of blood circulation, providing a useful tool for researchers to study how circulating cells contribute to certain populations[24]. If circulating monocytes contribute to the maintenance of homeostatic levels of microglia, one would expect to see similar levels of non-host cells in both the circulating system and the brain. However, multiple studies have demonstrated that parabiotic mice maintain higher levels of host-linked microglia[25–27] suggesting that monocyte cells do not contribute to the adult microglial population under normal conditions. It may be that under extreme conditions, such as a complete absence of microglia, brain injury or following significant neuroinflammation, circulating cells infiltrate the CNS. These cells may then contribute to the population of microglia-like cells in the brain, but this does not appear to be the case under homeostatic conditions[17,23,28].

The second theory of microglial repopulation is that there are progenitor cells within the brain that can differentiate into mature microglia. Following depletion of the microglial population in the adult mouse brain, using CSF-1R inhibitors, it has been

demonstrated that microglia rapidly repopulate the brain[29]. The rate of repopulation of microglia described in this study (from 600 cells/slice to >14,000 cells/slice in 72 hours) was determined to be too quick for repopulation to be explained by surviving cells. However, the presence of a progenitor population could explain these observations. Within the same study a population of Nestin and Ki67 positive cells were identified that appeared to be the source of repopulation. Initially, the nestin positive population had a distinct morphology to resident microglia, but then adopted the ramified morphology normally expected of native cells. However, since their initial description, the presence of microglia progenitor cells in the brain has remained controversial. Future studies have failed to identify a progenitor population[27] and noted that, while repopulating microglia may transiently express nestin, these cells derived solely from surviving cells. This suggests that adult microglia have proliferative potential and native cells are the driver behind population expansion.

1.3 Microglial function in development and the adult brain

There has been extensive research into the various roles microglia may play throughout the lifespan (Figure 1.2 [1,30–33]). As macrophage cells, microglia can clearly play an active role in the immune defense of the CNS. However, a growing body of evidence has shown that microglia are required for both neuronal development and normal brain function.

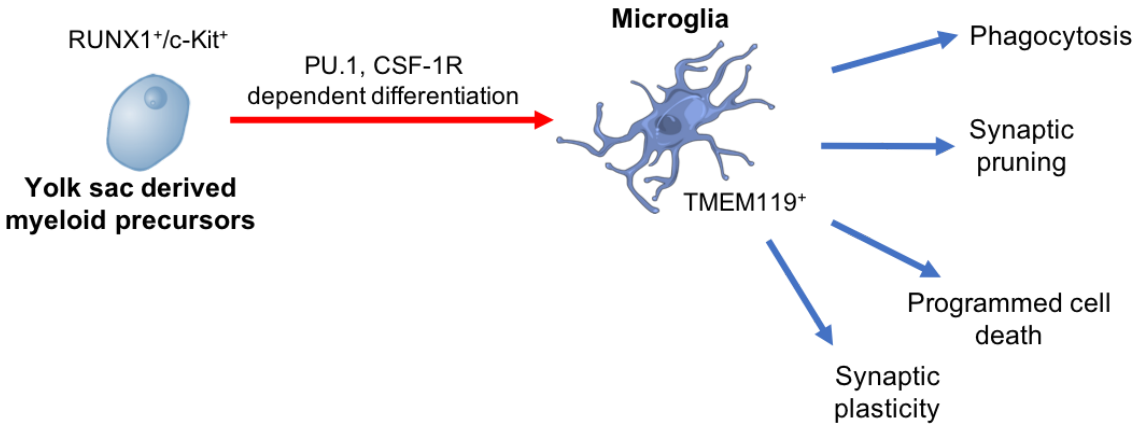


Figure 1.2 Overview of microglial development and function

A summary of microglial developmental pathways and functions in the healthy brain.

1.3.1 The role of microglia in the developing brain

Research in both humans and mice has demonstrated that microglia play an important role throughout brain development. Individuals with mutations in important regulators of microglial function, such as the *CSF-1R* gene, have profound neurological abnormalities[34] including abnormal arrangement of neurons and a lack of corpus callosum development. Studies like this provide direct evidence from human patients that microglial cells are required for normal brain development. However, these small scale patient studies cannot provide mechanistic details and so mouse models are often used as tools for studying microglia in development.

At the cellular level, microglia are able to phagocytose the early pool of neural precursor cells in order to control neurogenesis[35,36]. Studies have demonstrated that microglia play other important roles in brain development beyond their phagocytic function. Experimental evidence supports the idea that microglia provide trophic support to developing neurons in layer V cortical neurons in mice[37]. The cells accumulated close to the projection axons and, via a CX3CR1 dependent mechanism, produced IGF1 that maintained neuronal survival. Alongside trophic support for developing neurons microglial signalling has been shown to function in the programmed cell death of neurons. In the development of murine retina, prevention of microglial colonization of the tissue alleviated the production of nerve growth factor (NGF) and significantly reduced the level of normal programmed cell death[38]. More recent studies in both mouse Purkinje cells[39] and neurons in the mouse hippocampus[40] have implicated superoxide ions produced by microglia, through a CD11b/DAP12 dependent signalling pathway, in programmed cell death. Outside of their direct interactions with neurons, microglia also appear to be important for functional vasculature development *in-vivo* and, in the *in-vitro* based aortic ring model, addition of microglia cells to the culture stimulated vessel sprouting[41].

While the studies described above provide some evidence of the potential impact of microglia on neuronal development, one of the most well established and recognised functions of the cells in the developing brain is within the process of synaptic pruning. Synaptic pruning systematically removes weaker neurons and synaptic connections

to strengthen and improve the efficiency of the remaining connections within the brain. Experiments have shown that microglia closely co-localise to synaptic connections during active periods of pruning[42] and lysosomal markers have been used to highlight active engulfment of synaptic material[42,43]. Schafer *et al.*[42] studied microglia engulfment of synapses within mouse retinal ganglions and demonstrated that the cells preferentially digested "weaker" synaptic regions further supporting microglial involvement in the synaptic pruning process. Other studies have since established that the active engulfment of synapses by microglia is dependent on the activation of the classical complement cascade[42,44,45]. Disruptions of the CR3/C3 signaling cascade have been shown to cause deficits in synaptic connectivity[42] and *C1q* KO mice also have large disruptions in synapse elimination[44]. It is thought that complement protein tagged neurons provide the signal for phagocytosis by microglia[44].

1.3.2 Microglia in adulthood

Under normal conditions the brain is considered an "immune privileged" site, with the blood brain barrier (BBB) acting as a source of protection from infiltrating pathogens. While microglia may not have major immune functions under homeostatic conditions in the adult brain, it does not mean they remain inactive until disease or disruption occurs. Microglia are known to have a variety of homeostatic functions including phagocytosis of debris within the brain and monitoring of neuronal activity[1]. Many of the identified functions of microglial cells have been linked to CX3CR1 signalling. CX3CR1 is a receptor that is selectively expressed by microglia within the brain, which interacts with CX3CL1 ligand produced by neurons[46].

Recent evidence has also shown that microglia are important in the process of learning and memory in adults[47–49]. Learning and memory occur through the strengthening of synaptic and neuronal connections via processes of synaptic plasticity and long-term potentiation (LTP). *CX3CR1* KO mice have an impairment in measurable LTP alongside significant deficits in behavioural learning tests like fear conditioning and the Morris Water Maze[47]. ATP released by microglia in mice appears to modulate synaptic transmission by acting on $P2X_4$ and adenosine A1 receptors[48]. Using a selective eye closure mouse model, Sipe *et al.*[49] demonstrated

that microglia actively contribute to experience dependent plasticity through P2RY12 signalling.

There is also some evidence that external environmental factors can modulate microglial function. For instance, a high fat diet appeared to increase the number of microglia present within the hypothalamic region and was accompanied by an increased anti-inflammatory phenotype[50]. Obese humans studied within the same paper also showed cell type specific differences, including microglial dystrophy. Germ-free mice have also been used to study the impact of microbiome variation on microglial function[51]; without manipulation the mice showed global microglial defects including an immature phenotype and an impaired innate immune response. Recolonisation of germ-free mice partially restores microglia function, suggesting the influence of the gut microbiome on the brain is a dynamic process. However, these studies often do not provide evidence of specific molecular mechanisms that may drive these effects. Therefore, further research would need to be carried out to fully develop the scientific theories.

1.4 Microglia and disease

As the only major population of immune cells within the brain, microglia act as a first line of defence against infiltrating pathogens and are responsible for the clearance of cellular debris. However, microglia can also play a role in the development and progression of many disorders not immediately thought of as immune related [1,31,52]. When discussing microglia and disease it is important to distinguish between examples where microglia appear to play a causal role and those where the cells react to disease onset. The most well established causal link between microglial function and disease is Alzheimer's disease (AD) and as such this is discussed in more detail in section 1.5. The remainder of this section describes the evidence linking microglial function to a variety of other disorders and how the cells are involved in onset and progression.

1.4.1 Microglia in traumatic brain injury

Traumatic brain injury (TBI) is defined as “an alteration in brain function, or other evidence of brain pathology, caused by an external force”[53] and can often be further subdivided depending on the severity or outcome of the injury. As reactive immune cells within the brain, in the immediate aftermath of TBI microglial processes move rapidly to the site of injury, within minutes of damage[54]. Here, their primary function is to prevent disruption to the blood brain barrier[54–56]. Release of ATP from damaged tissue is thought to signal to microglia and stimulate the rapid movement of processes to the injury site, often without the movement of the cell body[54]. In mice it appears that microglial processes form specific honeycomb structures with single-process microglia dispersed throughout to assist with the sealing of the BBB[55]. A rapid increase in myeloid cell numbers occurs immediately in mice and can continue for up to four days[57]. Studies in human post-mortem brain samples have shown that the neuroinflammatory response that follows TBI can persist for months following injury[58].

TBI often has long term consequences including a potential increased risk of neurodegenerative disorders[1,59–63]. Meta analysis from 32 independent epidemiological studies, totalling “2,013,197 individuals, 13,866 dementia events and 8,166 AD events”, showed TBI increased the risk of any form of dementia by 1.6 times, with individuals showing a 1.5 times higher risk for AD specifically[64]. Many of the proteins associated with neurodegeneration have been shown to accumulate in the brain following TBI, including amyloid beta[65,66], tau[66] and α-synuclein[67]. Chronic traumatic encephalopathy (CTE), a neurodegenerative disorder characterised by the accumulation of hyperphosphorylated tau, has specifically been linked to consistent and repeated brain trauma[68].

Research into the molecular pathways that may drive this connection has suggested that chronic neuroinflammation driven by microglial responses may be responsible for the long term neurodegeneration risk associated with TBI[63,69]. Human brain autopsy samples from patients who have previously experienced a TBI have densely packed, reactive microglia that are not observed within aged matched control

samples[70]. The presence of these reactive microglia also appears to correlate with white matter degeneration, although only observational correlations were provided within this study. While some studies suggest that prolonged activation of microglia has a harmful impact on cognitive function there is also conflicting evidence that microglia may have a neuroprotective effect following TBI[63]. For instance, in a small randomised control study, TBI patients treated with the antibiotic minocycline showed a reduction in microglial activation but an increase in neurodegeneration compared to those patients not given the drug[71]. As well as the conflicting nature of some of the evidence around long-term microglial involvement in TBI, it should also be noted that neither side of the argument provides conclusive proof that microglia functions are driving the potential link between TBI and neurodegeneration.

The epidemiological studies linking TBI to dementia risk can also be difficult to interpret for a variety of reasons including misclassification of neurodegeneration and a lack of official clinical information[63]. It may also be that the link observed between TBI and AD could be driven by hidden factors that increase the risk of both AD and TBI without a causal link between the two. This means further work needs to be carried out on more controlled patient groups in order to fully understand the impact of TBI on dementia risk. It would also be worth building our understanding of how genetic risk factors can impact both TBI outcome and dementia risk. For instance, variants in the APOE gene linked to AD risk have been shown to impact TBI outcomes[72] but the interplay between the two is poorly understood.

1.4.2 Microglia in Multiple Sclerosis

Multiple sclerosis (MS) is a chronic neurological condition that is classified as both a neurodegenerative and autoimmune disorder. The immune system begins to attack the myelin sheath that surrounds neurons in the brain which leads to a multitude of symptoms including muscle weakness and coordination deficits. T-cells, primed to recognise myelin as forgien, are the driving immune cell type behind the development of MS.

While microglia are not associated with the onset of MS, the cells are present in the characteristic brain lesions of MS patients[73,74] and have been shown to be found near

to degenerating neurons in the brain[74]. The presence of the cells within diseased regions and their clear involvement in the immune response in the brain provides some evidence that microglia are involved in disease progression. However, as seen in TBI, different studies report opposing impacts of microglia function: either suggesting they further the progression of MS or that microglia play a neuroprotective role.

Production of reactive oxygen species (ROS) has been implicated in a variety of processes in MS [75] and microglia are often thought of as the major source of ROS within the brain. Microglia within the brain have been shown to express myeloperoxidase (MPO) and generate ROS as part of the myelin phagocytosis process[76]. Expression of MPO also significantly increased in MS patients compared to controls, with the highest level of expression seen in myeloid cells closest to lesion sites. The concept that microglia are the major source of ROS within MS has been further backed-up by more recent experimental data[77] and is thought to be due to Nox2 dependent oxidative burst. Microglia have also been shown to modulate neuronal activity in MS, further adding to described symptoms of the condition. In the Experimental Autoimmune Encephalomyelitis (EAE) mouse model of MS, activated microglia have been shown to release TNFα[78] which can in turn lead to enhanced glutamate function and synaptic degeneration.

On the other hand, a growing body of evidence has linked microglial function to protective disease processes, particularly remyelination[79,80]. *CX3CR1* KO mice, which have altered microglial functions, had a significantly reduced clearance of myelin debris in the EAE model which prevented remyelination[79]. It is also thought that anti-inflammatory microglia can aid the oligodendrocyte differentiation that is required for the remyelination process[80].

1.4.3 Microglial response in other neurological disorders

As the reactive immune cells within the brain, microglia have also been shown to respond to a variety of other neurological disorders, even though they may not play a causal role in the development of the disease. For instance, autism patients have increased microglia cell numbers when compared with healthy controls[81] and have

increased inflammatory profiles within the cerebrospinal fluid, including increased expression of macrophage chemoattractant protein (MCP)-1[82]. Microglia in autistic individuals may also be morphologically distinct. Morgan *et al.*[83] described a reduction in the number and length of distinctive microglial processes within the postmortem tissue from 13 male individuals with autism. Positron emission tomography (PET) scanning has revealed increased levels of microglial activation in autistic brains when compared to healthy controls[84]. Transcriptional profiling of brain tissue from autism patients has highlighted an increased expression of type 1 interferon genes compared to controls[85] and an enrichment of immune module genes within patient samples[86]. However, the genes linked to this immune module showed no enrichment for autism genome-wide association study (GWAS) genes. The lack of enrichment of immune genes within autism GWAS studies implies that the microglial response seen in patients is reactive rather than causal.

Microglia have also been linked to the symptoms associated with neuropathic pain[31,87,88], a chronic and debilitating pain caused by trauma, infection or pathology explicitly linked to peripheral nerve damage. As well as chronic pain symptoms, neuropathic pain also causes tactile allodynia: a disorder when pain hypersensitivity can be caused by what would normally be considered innocuous stimuli. While microglia are not involved in the initial pain stimuli or signalling, they have been shown to react to nerve damage associated with the disorder. Following initial peripheral injury there is marked neuroinflammation, microglial proliferation [89,90] and increased surveillance [91] by microglia. Crosstalk between neurons and microglia, through the CSF-1R signalling pathway, has also been linked to the onset of pain hypersensitivity[92]. Deletion of the *CSF1* gene from sensory neurons, which inhibits production of the signalling molecule, reduced pain hypersensitivity and microglial activation in mice.

1.5 Alzheimer's disease and microglia

Alzheimer's disease (AD) is the most common cause of dementia, a disease that affects around 850,000 people in the UK. Symptoms include progressive memory

loss and a reduction in general cognitive function. AD is also characterised by a general loss of neuronal mass. AD was first described by Dr. Alois Alzheimer in the early 1900s [93,94], where he noted plaques and tangles in patient autopsy samples that are now classically associated with AD pathology. AD is now clinically often split into two distinct categories: familial (early onset) and late onset AD (LOAD). It is thought that early onset AD makes up approximately 5% of all diagnosis' with this branch of the neurodegenerative disorder thought to be highly heritable[95]. Appearance of early onset AD symptoms often occur in patients in their 30s or 40s but are grouped up until the age of 65. Those that appear to sporadically develop symptoms after the age of 65, which is the more common condition, are classified as LOAD patients.

1.5.1 Early hypotheses in Alzheimer's disease research

The first major AD hypothesis focussed on the loss of cholinergic neurons within the brain[96]. Evidence of reduced acetylcholine release and its links with learning and memory further added to the theory[97]. The cholinergic hypothesis was the driver behind major pharmaceutical developments in AD treatments including the cholinesterase inhibitors that are still used in therapy today. However, since their approval as AD therapies, the cholinergic based treatments have appeared to only provide symptomatic relief with little to no effect on the progression of AD[98]. These observations suggest the specific loss of cholinergic neurons may not be driving the progression of the disease.

As understanding of the pathology of AD developed, the amyloid cascade hypothesis became the prevailing pathological theory. The amyloid cascade hypothesis states that it is the formation of the plaque like structures, seen within AD patient brains, that are the molecular drivers of the disease. It is now well accepted that the plaques first described by Alois Alzheimer are made up of aggregated amyloid protein (Aβ), specifically Aβ-42, and neurofibrillary tangles are composed of hyperphosphorylated tau. Hardy and Higgins were the first to coin the "amyloid cascade hypothesis" [99] and put forward the theory that the accumulation of plaques in the brain was the initiating stimulus that led to neuronal loss and the appearance of tau tangles. Since its development, amyloid and its role in the disease has been a major focus of AD research.

The earliest evidence implicating amyloid in AD came from studies of familial AD. Mutations within the amyloid precursor protein (*APP*) gene[100,101] and within the presenilin genes *PSEN1* and *PSEN2*[101–103] cause familial AD. APP, PSEN1 and PSEN2 are all involved in the production of the toxic Aβ-42 protein that forms the major component of plaques. The APP protein can be cleaved in different ways that lead to the production of a variety of forms of amyloid beta. It is thought that mutations associated with familial AD cause a bias towards the cleavage mechanism that generates the toxic Aβ-42. Further support came from early onset of AD in patients with Down's syndrome, who have three copies of the APP gene[104]. While mice do not spontaneously develop AD-like pathology or symptomatology as they age, *APP* and *PSEN* mutant mice have been shown to develop cognitive deficits, amyloid accumulation and synaptic loss[104].

Since the initial description of the amyloid cascade hypothesis, large bodies of research using a variety of molecular tools have been used to demonstrate that various forms of Aβ can initiate symptoms of AD[104–106]. For instance: in rat hippocampal cultures the addition of aggregated Aβ is neurotoxic[107]**,** APP transgenic mice have increased levels of Aβ oligomers and the same miceshow significant cognitive impairment compared to controls[108]. In mouse models of AD disrupting the amyloid pathway can result in a reversal of many of the cognitive phenotypes seen in the mice[109,110].

The growing evidence from *in-vitro* and *in-vivo* studies led to a push for drugs targeting the amyloid pathway. However, the amyloid cascade hypothesis is not without controversy [104,111,112]. One of the most significant problems with the theory that amyloid is the driver behind AD pathology is the repeated failure of anti-amyloid therapies in clinical trials[113]. These therapies fall into two broad categories: direct reduction of Aβ through antibody-style therapies and targeting of enzymes involved in the production of amyloid, such as BACE and γ-secretase. Many of the drugs targeting the enzymatic pathways have failed in clinical trials, either due to lack of efficacy[114] or significant off-target effects[115,116].

Despite the initial clinical safety failings of immunotherapies targeting Aβ[117], multiple therapies reached phase II and III trials[104]. However, the majority of these compounds have also dropped out of trials due to the failure to meet clinical endpoints[113]. In 2014, data was published from phase III trials of the anti-Aβ monoclonal antibody Bapineuzumab in which patients on the drug showed no significant improvement in AD-linked cognitive function compared to the placebo group[118]. The failure of Bapineuzumab in phase III trials came despite evidence from earlier phase II studies that long term treatment with the drug significantly reduced cortical amyloid fibrillar load[119].

The fact that immunotherapies targeting the amyloid pathway appear not to halt disease development despite reductions in amyloid load, has led to suggestions that targeting amyloid is the wrong strategy since it is not driving AD progression[113,120]. It is worth noting, however, that in late 2019 pharmaceutical company Biogen announced that they were seeking FDA approval for their anti-Aβ antibody despite earlier failure of the drug in trials[121]. The repeated failure of AD modifying drugs in clinical trials leads to questions not just about the validity of the targets but also practical factors about how trials are carried out[120] including whether patients are targeted for treatment too late in disease progression. There are also questions around the sensitivity of the major cognitive test used in AD clinical trials, the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-cogs), particularly in the early mild stages of disease[122].

1.5.2 Alzheimer's disease genetics and the neuroinflammation hypothesis

Although the amyloid cascade hypothesis has driven a large part of AD research, it is important to remember that the theory was founded on the genetics of early onset, familial, AD. The genetics behind LOAD is more complex and heterogeneous, not driven by single mutations in disease linked genes but by large numbers of variants of individually small effect sizes.

One of the first major genetic risk factors that was identified in LOAD is the *APOE* gene, a protein involved in cholesterol transport[123,124]. Specifically it has been demonstrated that the ε4 allele significantly increases AD risk, while the ε2 allele

confers a protective effect compared to the other alleles[125]. Early studies of the genetic risk factors for LOAD were carried out in relatively small patient numbers. This only allowed for the identification of single nucleotide polymorphisms (SNPs) which conferred relatively large increases in risk, such as *APOE*, or those within small targeted gene sets identified before analysis, such as *SORL1*[126]. However, genome-wide association studies (GWAS) have generated large scale datasets from case/control comparisons that can detect small effect size genetic links to complex disorders including AD[127].

While activation of the immune system, particularly microglia, was known to occur in AD as part of normal pathology[128,129], for many years this was thought to be a downstream effect of the disease. The results of AD GWAS provided the first indication that the innate immune system may have a causal role in the development of AD. Identification of SNPs near genes such as *CD33*, *CR1* and *MS4A6A*, which are classically considered immune related, suggests some role for the immune system within the disease. The identification of rare missense variants in genes, such as *TREM2*, *ABI3* and *PLCG2*, which are highly expressed in immune cells[130] has provided further evidence for the neuroinflammation theory. Table 1.1 lists the risk alleles identified in AD GWAS studies and the nearest gene to each SNP.

Lead SNP	Nearest gene	Publications
rs3851179	*PICALM*	131–134
rs10792832		135–137
rs11136000	*CLU*	131–133,138
rs9331896		134–136
rs4236673		137
rs3818361	*CR1*	132,139
rs6656401		135–138
rs4844610		134
rs744373	*BIN1*	133,139
rs6733839		134–136
rs4663105		137
rs3764650	*ABCA7*	139

rs4147929		[135,136]
rs3752246		[134]
rs111278892		[137]
rs610932	*MS4A6A*	[139]
rs983392		[135,136]
rs2081545		[137]
rs10948363	*CD2AP*	[135,136]
rs9473117		[134]
rs9381563		[137]
rs11771145	*EPHA1*	[135,136]
rs10808026		[134]
rs7810606		[137]
rs3865444	*CD33*	[135–137]
rs28834970	*PTK2B*	[135,136]
rs73223431		[134]
rs11218343	*SORL1*	[134–137]
rs10498633	*SLC24A4*	[135,136]
rs12881735		[134]
rs12590654		[137]
rs8093731	*DSG2/SUZ12P1*	[135,137]
rs35349669	*INPP5D*	[135,136]
rs10933431		[134,137]
rs1476679	*ZCWPW1*	[135,136]
rs1859788		[137]
rs17125924	*FERMT2*	[134–136]
rs7274581	*CASS4*	[135,136]
rs6024870		[134]
rs6014724		[137]
rs593742	*ADAM10*	[136]
rs442495		[137]
rs889555	*BCKDK/KAT8*	[136]
rs59735493		[137]
rs138190086	*ACE*	[134,136]
rs12444183	*PLCG2**	[136]
rs75932628	*TREM2**	[134]

rs187370608		[137]
rs7920721	*ECHDC3*	[134]
rs11257238		[137]
rs28394864	*ABI3**	[137]
rs179943	*ATXN1*	[140]
rs3826656	*NT_011109.848*	[140]
rs2049161	*BC040718*	[140]
rs597668	*EXOC3L2*	[133]
rs670139	*MS4A4E*	[139]
rs190982	*MEF2C*	[135]
rs2718058	*NME8*	[135]
rs10838725	*CELF1*	[135]
rs9381040	*TREML2*	[136]
rs59685680	*SPPL2A*	[136]
rs4985556	*IL-34*	[136]
rs3740688	*SPI1*	[134]
rs7933202	*MS4A2*	[134]
rs4575098	*ADAMTS4*	[137]
rs184384746	*HSEX1*	[137]
rs6448453	*CLNK*	[137]
rs114360492	*CNTNAP2*	[137]
rs117618017	*APH1B*	[137]
rs113260531	*SCIMP*	[137]
rs2632516	*BZRAP1-AS1*	[137]
rs76726049	*ALPK2*	[137]
rs76320948	*AC074212.3*	[137]

Table 1.1 Summary of reported AD GWAS hits

Lead SNPs and nearest genes identified in AD GWAS studies. Certain loci have differing lead SNPs identified by studies but are grouped by nearest gene. Loci with a * next to the gene name have previously been identified in rare variant studies.

The results of GWAS studies displayed here provide summaries of each locus, highlighting only the most associated SNP and the nearest gene to that SNP for each region. Linkage-disequilibrium (LD) within the human genome is a terminology that describes certain SNPs within a region that are found to be more associated with

each other than would be expected if they were inherited randomly. This means there are often multiple SNPs within a region in strong association with the “lead” SNP identified in a GWAS. It is, therefore, not possible to tell from standard GWAS analysis which of these SNPs is causal. Additionally, because disease associated variants are noncoding, there are many genes within a specific window of the associated SNPs that could be impacted by the variant. This means that it is also not possible to tell exactly which gene, and downstream signalling pathways, may be linked to disease risk.

To address these problems, methods to combine GWAS data with functional data, including transcriptomics (expression quantitative trait loci (eQTL) maps) and open chromatin assays (chromatin accessibility quantitative trait loci (caQTL) maps). It is then possible to run co-localisation analysis to identify variants affecting both disease risk and a functional output have been developed. Computation tools also provide methods to extend traditional GWAS analysis. For instance, GoShifter[141] prioritises functional annotations to identify causal variants by finding SNP enrichments in annotated regions.

In the case of AD, these combination approaches have further linked the immune system to disease risk. For instance, when eQTL maps of monocytes and T cells were colocalized with GWAS summary statistics from a variety of complex traits, significant co-localisations with AD GWAS SNPs were only identified within the monocyte eQTL map[142]. While this implied that the myeloid cell lineage of the immune system may be driving the neuroinflammatory component of AD, it did not fully rule out a role for neurons themselves. Integrative analysis of published GWAS summary statistics and whole-brain single cell RNA-sequencing data shows a significant enrichment of AD GWAS signal within the specific gene expression pattern of microglial cells, while no enrichment was seen in neurons[143]. AD risk SNPs are also significantly enriched in regions of open chromatin in myeloid cells, including microglia, but not in whole brain chromatin accessibility data[144]. Although AD genetics studies have now identified multiple risk loci these have not yet provided direct information on the biological role of microglia in neurodegeneration.

1.5.3 The role of microglia in Alzheimer's disease

Genetic studies have spurred a resurgence of research into how microglial function changes during AD. When Alois Alzheimer first described the brain pathology of AD, in addition to identifying amyloid plaques and tau tangles, he also observed alterations in the glia surrounding these abnormal proteins, including the development of "fibers" and "adipose saccules"[93]. Since this initial description, there has been a growing body of research that focuses on microglial involvement in AD. This has provided evidence that often falls into one of two categories: that promoting microglial activity will be beneficial in AD or that a reduction in activity will slow AD progression. However, these two ideas may not be mutually exclusive in that certain processes may be both beneficial or harmful depending on the context.

Microglial phagocytosis is a good example of the above phenomenon. Initially, research focussed on microglial phagocytosis of amyloid plaques within the brain[129,145,146], in part due to the observed physical association of microglia with the plaques. It has been suggested that microglial recruitment to plaque sites promotes phagocytosis and lowers plaque burden[147]. However, as the disease progresses the phagocytic capability of microglia reduces[148] and in fact the cytokines produced by the process are part of a negative feedback loop that reduces phagocytosis[147]. The evidence from these mouse studies implies that promoting microglial phagocytosis could be a viable therapeutic target as it reduced amyloid load. However, selective reduction in microglial populations in an AD mouse model may reduce neuronal loss without impacting amyloid load[149] which suggests microglial phagocytosis of amyloid is not necessarily required for the reversal of AD symptoms. In fact, microglial phagocytosis, via a complement dependent mechanism, has since been linked to excessive engulfment of healthy synapses[150]. This means that increasing microglial phagocytic capabilities may in turn lead to further neuronal loss.

Outside of phagocytosis, microglia have been linked to a variety of other molecular processes in AD. For instance CSF-1R inhibition in the 5XFAD mouse model of AD has been shown to significantly reduce the seeding of plaques within the brain[151,152], although Aβ accumulation still appears in cortical blood vessels. Other work suggests that microglia may form a barrier around developing plaques which reduces further

accumulation of Aβ[146]. In tauopathy mouse models, microglia aid the propagation of tau across the brain via the secretion of previously phagocytosed tau in exosomes[153].

Further insights into microglial functions in AD have come from studying mutations identified by GWAS. For example, multiple studies have functionally characterised mutations in *TREM2*[154,155]. Triggering receptor expressed on myeloid cells 2 (TREM2) is a receptor that signals through a TYROBP/DAP12 dependent mechanism to activate a variety of signalling pathways and downstream functions, such as phagocytosis and chemotaxis[156]. A variety of approaches have shown that disease-associated missense mutations in *TREM2* can alter microglial phagocytosis, survival and proliferation[156]. The soluble form of TREM2, produced following cleavage of the receptor, has also been implicated in AD[157–159]. There is evidence that TREM2 may function in conjunction with other GWAS risk genes during AD including *APOE*[160,161], *CD33*[162] and *MS4A*[163].

Alternative experimental approaches have examined how microglial functions change in AD patients compared to age matched healthy controls, particularly at the level of gene expression. In mice, two studies have identified microglial populations that only appear in diseased states[164,165] and identify a loss of homeostatic gene expression (*P2RY12*, *CX3CR1* and *TMEM119*) alongside an increase in inflammatory markers such as *AXL*, *CLEC7A* and *CST7*. Additionally, activation of TREM2 signalling pathways were required for the formation of this disease associated subtype of microglia cells in mice. In human samples, single cell analysis of AD post-mortem brain samples also identified a disease specific population of microglial cells[166]. Like the populations identified in mice, these cells had increased expression of genes like *SPP1* and *APOE*. The disease specific microglia also showed an increased expression of HLA and complement linked genes, compared to non-disease linked microglia.

In summary, it is clear that microglia play a significant role in how our brains function in health and disease but exactly how microglial processes change in disease and precisely how to target the same pathways in treatments remains unclear. Much of this complexity often arises because microglia seem to play both detrimental and

beneficial roles in many diseases depending on the stage, activation pattern or model system being studied.

1.6 Studying human microglia

While significant advances have been made in microglial research, many of the studies that have been used to understand microglia function in health and disease have been carried out in mice. Mouse models are an invaluable tool, enabling large scale studies, manipulation of the cells and providing a way to study microglia throughout the lifespan of an organism. However, studies in mice are not without limitations and controversies[167–169]. There are significant differences in the fundamental functions of microglia in mice and humans, including differences in marker expression, such as IFNγ and TLR4, and differences in response to pharmacological compounds. In mouse models of AD, microglia are often described as taking on an activated phenotype while in human autopsy samples the cells appear to degenerate with age, often referred to as dystrophic or senescent[170]. This can lead to opposing theories about the role microglia play in disease.

However, primary human microglia are extremely difficult to source and come with experimental caveats. Many commercially available human microglia sources are fetal samples which may behave differently to fully developed microglia. Additionally, commercially available cells are often cultured which can impact microglial expression[171]. Protocols for accessing human adult microglia cells from both post-mortem and surgical tissues have been refined and appear to yield relatively pure samples[172–174]. Although isolated human microglia may have high purity, there are multiple experimental factors to consider when using these cells. Even small periods of culturing can alter the profile of human microglia[171,175] and little is known about how the isolation protocols (dissociation and cell marker expression based sorting) may impact microglial profiles. Small scale microarray analysis of sorted murine mammary glands has suggested that fluorescence activated cell sorting (FACS) has minimal impact on gene expression[176]. However, full comparisons have

not been carried out to understand how FACS sorting may impact immune cell expression, particularly microglia.

While it is possible to isolate fresh primary adult human microglia from neurosurgical patients, in order to study microglia from healthy individuals, samples must be acquired from post-mortem tissue. As microglial phenotypes have been shown to be heavily dependent on the active neuronal environment[171], it is therefore difficult to know how much post-mortem delay impacts microglia. A study comparing isolated microglia from brains with differing lengths of post-mortem delay demonstrated that disease state had a greater impact on microglia than the time between death and collection[175]. However, it is difficult to directly compare fresh microglia to post-mortem samples while controlling for confounding factors. Therefore, it is impossible to definitively know the impact of post-mortem collection on microglial phenotype.

1.6.1 Transcriptomic studies in primary human microglia

RNA-sequencing technology enables the study of the whole transcriptome of cells and whole tissues. Statistical analysis of the resulting data can be used to compare the transcriptional profiles of samples across a variety of conditions. As isolation protocols for human primary microglia have improved RNA-sequencing has become widely used to understand differing aspects of microglia. This includes comparisons between human and mouse samples[171], identifying microglia-specific marker genes[177,178], comparison of transcriptomes across ages[179], highlighting region and disease specific changes in gene expression[180] and understanding the role environment plays in microglial gene expression[171].

While RNA-sequencing at a bulk level has provided tools to study large scale gene expression and generated vast amounts of data, the ability to use the technology at the single-cell resolution has provided a tool to study gene expression at a much finer resolution[181,182]. Single-cell RNA-sequencing (scRNA-seq) allows identification of individual populations of cells *in silico,* obviating the need for prior knowledge of cell markers, and enabling comparisons of tissue composition between experimental groups.

scRNA-seq has allowed researchers to take whole brain tissue and identify multiple cell types, such as neurons and microglia[166,183,184]. Whole brain single cell analysis has been used to investigate changes that occur to different cell types in the brain during development[183] and disease[166,184]. Being able to identify microglia from whole brain samples also removes the cell sorting step required for bulk RNA-sequencing, which in turn reduces the chances of experimental processes impacting microglial gene expression. However, within whole brain single cell analysis the fraction of microglia is relatively low (3% reported by Mathys *et al.*[166]) and smaller numbers of cells per subgroup makes statistical comparisons more difficult. Therefore, it is also possible to use single-cell sequencing on sorted primary human microglia[185,186], in order to better capture subtle microglial population changes. This has been used to further our understanding of microglial populations across ages[185] and disease[185,186].

An extended review of how transcriptional analysis of primary microglia has impacted our understanding of the cell type can be found in section 2.1. While current published datasets have provided an insight into microglial transcriptomes, many are still based on relatively small patient numbers. This is largely because access to primary human microglial samples is still difficult. Growing brain bank collections have allowed access to larger numbers of post-mortem samples but these studies are still limited by patient number (with the largest reported at 48 collections[166]) and often cover only specific disease states. Fresh human microglia are even more difficult to access, coming from either fetal samples or neurosurgical patients.

1.6.2 Modelling human microglia

While studying primary human microglia is important for understanding the cells in health and disease, there are clear limitations with these studies particularly around scale and the ability to experimentally manipulate the cells. Therefore, a clear challenge has been to develop ways to model human cells in the lab. Induced pluripotent stem cells (iPSCs) are proliferating cells that have been reverted back to a stem cell like state from adult cells and they have the potential to differentiate into any cell[187–190]. This means iPSC based cell model systems provide researchers with a useful tool for studying human disease in a dish[191]: they are able to be used at scale, can be manipulated experimentally and allow for repeated sampling. Large scale

banks of iPSC lines, such as the HipSci consortium, mean that researchers can also run iPSC based experiments using large numbers of both healthy and diseased cell lines.

As iPSC cells can technically be differentiated into any cell in the body, methods have been developed to differentiate these cells along a myeloid lineage. Initially these studies focussed on the development of macrophage models and their utilisation for studying immune response[192–195]. Many of these iPSC-derived macrophage differentiation protocols make use of the induction of embryoid bodies (EB) from stem cells. These EB structures are made up of cells from all three germ layers[196] that can then further differentiate into more specialised cells.

However, in more recent years there has also been a focus on pushing the myeloid cells closer towards the specialised microglia-like phenotype. These protocols range from simple monoculture based systems[197–201], similar to those used to generate macrophage-like cells, to more complex co-cultured[198,202] and organoid systems[200,203–206]. These more complex model systems build on the idea that much of the unique microglial transcriptional signature comes from the environmental stimulation they receive from neurons and other parts of the CNS[171].

A major factor to consider when using *in-vitro* models for human cells is understanding how accurately the cell culture systems capture the primary cell type. Often this comparison is limited to marker gene expression and functional capabilities. For a detailed analysis of how the iPSC models described above have been compared to primary cells see Chapters 3 and 4. For microglia particularly, comparison is complex, as the primary cells are difficult to access and therefore transcriptional comparisons are often made across studies. This can often lead to confounding batch effects, especially when running small scale comparisons. Systematic comparisons of model systems to the primary microglia can be used to highlight potential signalling pathways that are not switched on *in-vitro* and could be manipulated to move cells closer towards the primary cell type.

1.7 Thesis overview

The overarching theme of the following thesis builds on section 1.6 and the difficulties around studying human microglia. I aim to answer three major questions throughout the thesis: **1.** How does microglial composition and gene expression profile change across a population? **2.** How accurately do current simple *in-vitro* model systems of human microglia capture the profile of primary human cells? **3.** Does culturing stem cell derived microglia with neurons move the model systems closer to the primary phenotype?

The analysis in the second chapter of my thesis forms part of a large-scale project in collaboration with Dr Adam Young and Dr Natsuhiko Kumasaka studying the genetic architecture of human primary microglia. As part of the project we collected and processed the largest number of fresh, primary human microglia samples to date from a wide variety of clinical phenotypes. In this chapter I used single cell RNA-sequencing to identify different subpopulations of primary microglia and identified how the likelihood of finding cells within these populations is influenced by clinical phenotypes. I then used bulk and single cell RNA sequencing data from the same patient population to further understand how clinical phenotypes such as age, pathology and sex influcenced microglial transcriptomes.

In the third chapter of my thesis, I focus on the transcriptional profiles of *in-vitro* models of microglia and how closely they match the transcriptional profile of the primary human cell type. I collected publicly available data and combined it with available in-house datasets to generate a large scale analysis project to compare primary human microglia with monocyte-derived macrophages, cancer-cell lines, iPSC-derived macrophages and iPSC-derived microglia. For all the data, I used raw sequencing files that were all processed through the same pipeline and I ensured that I collected data from multiple studies for each cell type. Both of these decisions were made to reduce the batch effects that can occur when comparing sequencing data across different studies[207–209]. I used the processed data to understand how the

different *in-vitro* systems capture the gene expression of the primary cells and which signalling pathways may not be switched on in these *in-vitro* systems.

In the final results chapter of my thesis, I will focus on more complex stem cell based model systems, including co-culture and orgainoid based models. This forms part of a collaboration with Dr Phil Brownjohn and Dr Moritz Haneklaus, from the Livesey Lab, working with their published microglia differentiation protocols[200]. I initially used bulk RNA-sequencing to add the complex model systems to the large dataset generated in Chapter 3 in order to understand how the more complex model systems compared to the monoculture systems described in Chapter 2. I then used single cell sequencing, and particularly single cell trajectory analysis, to understand how microglial cells from each of the model systems fit on a developmental pathway that ultimately ends with the primary cell type.

Chapter 2: Heterogeneity in primary adult microglial transcriptomes

Collaboration note

The work described in the following chapter forms part of a collaborative project. Patient samples were collected and primary microglia were isolated by Dr Adam Young and colleagues at the Division of Clinical Neurosciences based at Cambridge University Hospital and the Wellcome Trust Medical Research Council Cambridge Stem Cell Institute. Single cell sequencing preparation was carried out by the single cell sequencing facility at the Wellcome Sanger Institute. Myself and Dr Andrew Knights worked collaboratively to process the bulk primary microglia samples for sequencing. Dr Natsuhiko Kumasaka ran the initial quality control analysis across the dataset. For the bulk data, he used genotype information to identify any sample swaps and mixes and for the single cell analysis he ran the initial processing to remove poor quality samples.

Initial analysis of the single cell dataset was carried out by myself including visualisation and clustering of single cell data, links to clinical metadata and Alzheimer's disease. It was then determined that the analysis needed to be updated to be corrected for potential batch effects or confounding factors. Due to an injury, and a 3 month medical intermission of my PhD, Dr Natsuhiko Kumasaka ran the re-analysis of the data in order to prepare a manuscript for submission[210]. The single cell work discussed in this chapter is from the analysis run by Dr Natsuhiko Kumasaka and some extended work by myself. Any figures taken directly from the analysis are noted in the figure legend.

2.1 Introduction

As interest in microglia has developed it is important to fully characterise the gene expression profile of primary microglia, both to understand how they are perturbed in disease and how we can be modeled *in-vitro*. To date, most studies of primary microglia have been in mice, with validation in small numbers of human samples.

Many studies have used RNA-sequencing to identify transcriptional markers of microglia, with a focus on differentiating the native cell from classical macrophages and other tissue resident macrophages.

2.1.1 Marker gene identification in mice and human samples

Microarray analysis has been used to compare tissue resident dendritic cells (from the spleen, liver and lung) and tissue resident macrophages (spleen, lung and peritoneal macrophages and microglia) in C57BL/6J mice in order to identify markers of each cell type[177]. Microglia were shown to have a lower expression of hundreds of transcripts that were expressed in other tissue resident macrophages. The paper also identified gene expression that is specific to microglia in comparison to the other tissue resident cells, notably the transcription factor *SALL1* and cell surface marker *CX3CR1*. More recently[178] a six-gene microglial transcriptional signature (*P2RY12*, *GPR34*, *PROS1*, *GAS6*, *C1QA* and *MERTK*) has been identified which appears to distinguish microglia from other immune cells, including other myeloid cell types, and other brain cells, such as astrocytes and neurons. As well as validating the unique signature within primary human cells, the group also cultured adult mouse microglia in the presence or absence of TGF-β and demonstrated that the signature they described is TGF-β dependent.

Two independent studies[211,212] have since pinpointed *TMEM119*, a protein coding gene originally linked to bone formation, as a marker that distinguishes native microglia cells from infiltrating myeloid progenitors. It is currently unclear whether resident microglia cells and infiltrating cells play differing roles in disease, such as AD, and the studies described above suggest that finding markers for each cell type may help future researchers to follow the role of each cell type.

2.1.2 Fresh, primary human microglia bulk RNA-sequencing

The most extensive bulk RNA-sequencing dataset of fresh human primary microglia to-date profiled the cell type across 19 individuals between the ages of 5 and 15 and also included chromatin accessibility studies of the same samples[171]. Here it was shown that broad clinical diagnosis (acute ischemia, epilepsy and tumour), age and sex had no observable impact on microglial gene expression and highlighted that

pathology did not significantly affect expression of the most highly expressed microglial genes in their dataset (e.g. *SPP1*, *CD74* and *ACTB*). Using ATAC-seq and ChIP-seq, they detected the most enriched transcription factor recognition motif associated with open chromatin and highlighted a dominant signature for the *PU.1* transcription factor. The group also ran RNA, ATAC and ChIP-seq on matched samples from fresh collections and cells that had been cultured for varying lengths of time. They noted that expression of microglia marker genes such as *CX3CR1* and *P2RY12* as well as transcription factors such as *SALL1*, decreased after a period of only 6 hours in culture and continued to decline over 7 days in cell culture.

The authors also demonstrated that the addition of TGF-β to the *in-vitro* culture media of the primary cells had a modest effect on gene expression, with expression of certain genes, such as *SALL1*, increasing back towards the levels seen in the fresh primary cells. Although, it was noted that none of the genes whose expression increased in the presence of TGF-β returned to fully match the levels seen in the primary cells. As had been suggested in earlier studies[178], this provided further evidence that TGF-β signalling is, at least in part, important for maintaining microglial transcriptional identity.

2.1.3 Single cell sequencing and primary microglia

Advances in technology means that it is now possible to study transcriptomes at a single cell level, which allows researchers to study heterogeneity of cell types in a population. Single cell profiling of 16,000 CD45 and CD11b sorted microglial cells from 15 individuals (7 autopsy and 8 biopsy samples) identified 14 unique microglial populations within the brain[185]. Within the 14 subpopulations identified, the authors noted that the three largest clusters were transcriptionally similar with no differentially expressed transcription factors between groups. It was, therefore, suggested that these subpopulations represented cells of the same class but in different activation states. The remaining, more transcriptionally distinct, microglial clusters were considered more specialised subtypes of microglial cells.

Single cell transcriptomics can also be used to understand dynamic changes in cell expression or cell proportions in health and disease across whole tissues. In

microglial research this is of particular interest when looking at changes that occur during Alzheimer's disease (AD). Single cell analysis of whole brain tissue has identified AD specific microglia gene expression changes in both mice[164] and human[166,184] samples. Although it is worth noting that as microglia represent a small fraction of cells within the brain, there are limitations in the ability to understand heterogeneity within the cell type due to low cell numbers.

2.1.4 The impact on age and sex on microglial transcriptomes

As microglia have a distinct origin and are not replenished by circulating monocytes under normal conditions[17], previous work has also focused on how microglial transcriptomes change with age. Comparison of 10 aged (average age at death = 95) bulk post-mortem microglia RNA-sequencing profiles to a publicly available dataset of primary microglia from middle-aged individuals (mean age = 53) identified 1060 upregulated and 1174 downregulated genes in the aged microglia[179]. Pathway enrichment analysis showed that upregulated genes were enriched for amyloid fiber formation and those genes with decreased expression in aged microglia were enriched for TGF-β signaling. The loss of TGF-β signaling in aged cells was suggested to represent a loss of the homeostatic function of microglia during aging.

While comprehensive aging studies in human microglia are complex, due to the lack of accessibility of the cell type, it is possible to monitor changes in microglial transcriptomes across the lifespan of mice[213]. Using single cell sequencing, researchers were able to identify populations of microglia enriched for cells from aged mice and showed that the gene expression profile of these cells was shifted towards a more active state, due to increased expression of inflammatory markers. However, the authors noted that the proportion of the cells in this increased active state was only a small fraction of the total cells in these aged mice. It was suggested in the study that this may be because the activated cells were responding to local disruptions, such as blood brain barrier compromise[214] or microinfarcts[215], that can be associated with aging as opposed to representative of a global change in expression profile.

Previous work has also focused on whether microglial transcriptomes differ between sexes. Evidence from mouse studies is often conflicting. One study[216], noted large numbers of differentially expressed genes between male and female adult mice and the authors highlighted that male microglia show an increased inflammatory phenotype. The researchers also showed that female microglia are protective during ischemia within mice and suggested that it was due to the fact that the microglia were able better control excessive inflammation. Further studies in mice have also highlighted how microglial gene expression can be impacted in sex specific ways during development[217] and as part of the interaction with the microbiome[218]. However, Hammond *et al.*[213], compared single cell microglial gene expression in male and female mice across three major developmental ages (E14.5, P4/P5, and P100) and highlighted only a small difference between the sexes. While, as expected, genes on the sex chromosomes were differentially expressed between male and female mice there was only a small fraction of cells (~0.5% of microglia) that appeared to cluster in a sex specific way. The cluster was enriched for female cells of the P4/P5 developmental age and showed increased expression of genes such as *CD74* and *ARG1*. In human studies, the evidence for sex-specific expression of genes in microglia is limited. Using bulk RNA-sequencing, Gosselin *et al.*[171] observed that a small set of genes, most located on the sex chromosomes, showed sex-specific differences.

One limitation of the studies discussed above are their small sample sizes. This means that previous observations of correlations between microglial transcriptional profiles and life-history or clinical pathology are based on phenotypes from small numbers of individuals. In this chapter, I describe the analysis of bulk and single cell RNA-sequencing data from a cohort of 141 patients samples of fresh primary adult human microglia, the largest cohort to date. I describe how heterogeneous primary microglia were across patients and identified markers for individual subpopulations of the cell type. I highlight how clinical pathology was a major driver of heterogeneity across microglia and how this information can be used in conjunction with subpopulation markers to infer biological relevance of clusters. Using both single cell and bulk data I investigate how various other clinical phenotypes, such as age, sex and brain region, can affect microglial transcriptomes.

2.2 Methods

2.2.1 Experimental design and sample collection

Human brain tissue was obtained with informed consent under protocol REC 16/LO/2168 approved by the NHS Health Research Authority. All collections were completed by Dr Adam Young and his colleagues at the Division of Clinical Neurosciences based at Cambridge University Hospital. Samples were collected from neurosurgical patients undergoing scheduled procedures where tissue would normally be removed. Patient pathologies were grouped into four major categories: control, haemorrhage, hydrocephalus, trauma and tumour. Control samples include tissue where the site of sampling is a site further away from the site of injury or disease (i.e. tumour biopsy where the tissue sampled is considered pathologically normal). Figure 2.1 summarises the metadata for all patient samples collected and includes the experimental design of the study. Tissue samples were used for both bulk and single cell RNA-sequencing. Paired blood samples were also taken from each patient at the induction of anaesthesia for genotyping. However, genotype information was not used in the analysis described in this chapter.

Once collected tissue was immediately transferred to Hibernate A low fluorescence (HALF) supplemented with 1x SOS (Cell Guidance Systems), 2% Glutamax (Life Technologies), 1% P/S (Sigma), 0.1% BSA (Sigma), insulin (4g/ml, Sigma), pyruvate (220 g/ml, Gibco) and DNase 1 Type IV (40 g/ml, Sigma) on ice and transported to a dedicated CL2 laboratory.

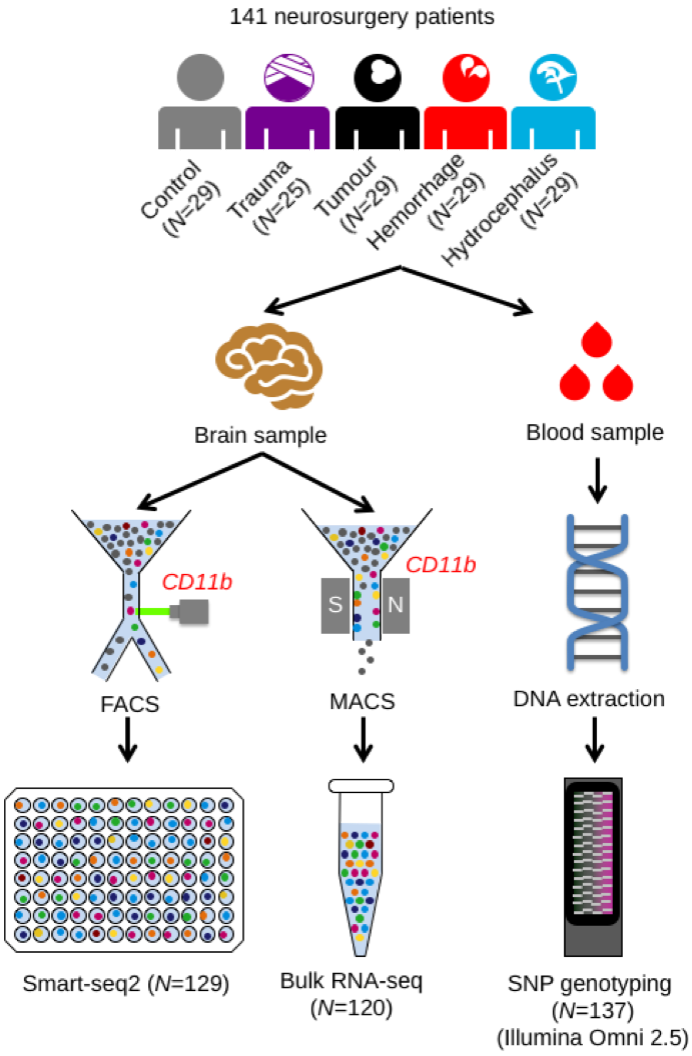


Figure 2.1 Schematic of experimental design

Experimental protocol for all (141) samples collected as part of the 16/LO/2168 linked study. Plot created by Dr Natsuhiko Kumasaka.

2.2.2 Tissue processing and cell sorting

All tissue processing was completed by Dr Adam Young colleagues at the Division of Clinical Neurosciences based at Cambridge University Hospital and the Wellcome Trust Medical Research Council Cambridge Stem Cell Institute.

Brain tissue was mechanically digested in fresh ice-cold HALF supplemented with 1x SOS (Cell Guidance Systems), 2% Glutamax (Life Technologies), 1% P/S (Sigma), 0.1% BSA (Sigma), insulin (4g/ml, Sigma), pyruvate (220 g/ml, Gibco) and DNase 1 Type IV (40 g/ml, Sigma). The prepared mix was spun in HBSS+ (Life Technologies)

at 300g for 5 mins and supernatant discarded. The digested tissue was rigorously triturated at 4°C and filtered through a 70 m nylon cell strainer (Falcon) to remove large cell debris and undigested tissue. Filtrate was spun in a 22% Percoll (Sigma) gradient with DMEM F12 (Sigma) and spun at 800g for 20 mins. Supernatant was discarded and the pellet was resuspended in ice cold supplemented HALF.

For single cell smartseq2 sequencing, human microglia were sorted using fluorescence-activated cell sorting (FACS). The isolated cell suspension was incubated with conjugated PE anti-human CD11b antibody (BioLegend) for 20 mins at 4°C. Cells were washed twice in ice cold supplemented HALF and stained with Helix NP viability marker. Cell sorting was performed on BD AriaIII cell sorter (Becton, Dickinson and Company, Franklin Lakes, New Jersey, US) at the University of Cambridge Cell Phenotyping Hub at Cambridge University Hospital, Cambridge, UK. Cells were sorted into 96 well plates, prepared by the Wellcome Sanger Institute for the purposes of single cell sequencing.

To avoid sustained stress on microglia as a result of prolonged sorting times for bulk sequencing magnetic-activated cell sorting was performed on these cells. Isolated cell suspensions were incubated with anti-CD11b conjugated magnetic beads (Miltenyi) for 15 mins at 4°C. Cells were washed twice with supplemented HALF and passed through an MS column (Miltenyi). Each sample was washed three times in the column and then extracted. Samples were added to a 1.5ml Eppendorf to which 300 l of RNAlater (Qiagen) was added, samples were stored at -80C prior to library preparation and sequencing.

2.2.3 RNA handling

For single cell sequencing, 96 well plates were prepared and sequenced by the Wellcome Sanger Institute single cell core facility using the SmartSeq2 protocol [219]. Extraction and library preparation of bulk samples was completed by Dr Andrew Knights and myself. Total RNA from the bulk primary microglia samples was extracted with the Qiagen AllPrep DNA/RNA micro kit. This was carried out according to the manufacturer's instructions. Following extraction samples were analysed using

an Agilent Technologies Bioanalyser RNA Pico kit for quality (RIN number) and quantification. Extracted RNA was stored at -80 °C until library preparation.

The amount of total RNA extracted from these samples was incredibly varied, ranging from > 300 ng to 0.5 ng of approximate yield, with the majority of samples producing less than 10 ng of total RNA. This is a much lower input RNA level than is required for traditional bulk sequencing and, therefore, we used a low RNA input library preparation pipeline developed in-house by Dr Andrew Knights which is a modified version of the SmartSeq2 protocol protocols developed for single cell sequencing. For samples with large amounts of RNA yields, 10 ng was used as a maximum input for the protocol. Samples with lower than 10 ng of RNA input were processed in the same way, although the number of PCR amplification cycles was increased for certain samples to compensate for the low input amounts (Figure 2.2). In total 120 of the 141 collected samples were prepared for sequencing, the 21 samples that were not included in sequencing pools were discarded due to either having no quantifiable RNA or large amounts of RNA degradation, to the point where no RIN value could be calculated.

25 µL of lysis binding buffer (Table 2.1) was added to the extracted RNA, that had been diluted to 25 µL with nuclease free water. 20 µL of oligo-DT beads were added to wells of a 96-well plated and washed once with 100 µL of lysis binding buffer while on a magnetic plate. The pelleted bead plate was removed from the magnet and the beads were resuspended with the 50 µL RNA samples. The wells were pipette-mixed and incubated at room temperature for 15 minutes, with shaking (1100 rpm Mixmate). The plates were then placed back on the magnet for supernatant removal and two washes with 150 µL of wash buffer A (Table 2.1). Samples were then transferred to a fresh plate before washing twice with 50 µL of wash buffer B (Table 2.1).

The samples were washed again with 50 µL of elution buffer before RNA is eluted from the beads by re-suspension in 9.5 µL of elution buffer and incubating at 75 °C for 2 minutes. Plates were then immediately transferred back to the magnetic plate and 7 µL of eluted solution was transferred to a fresh plate on ice. 2 µL 10 µM oligo $dT_{30}VN$ and 2.34 µL 10 mM dNTPs (Thermo) were added to each well of the 96-well

plates and samples were heated at 72 °C for 3 minutes before being rapidly chilled on ice. 13.65 µL of reverse transcription (RT) master mix (Table 2.1) was added to each well of the plate and following mixing the samples were placed on a PCR block for RT (Figure 2.2).

Lysis binding buffer (100 mL)	Wash buffer A (250 mL)	Wash buffer B (100 mL)	RT master mix (per reaction)
20 mL of 1 M Tris-HCl pH 7.5 (FC = 200 mM)	2.5 mL 1 M Tris-HCl pH 7.5 (FC = 10 mM)	1 mL 1 M Tris-HCl pH 7.5 (FC = 10 mM)	5 µL 5x SmartScribe FS Buffer
12.50 mL 8 M LiCl (FC = 1 M)	4.69 mL 8 M LiCl (FC = 0.15 M)	1.88 mL 8 M LiCl (FC = 0.15 M)	0.63 µL SUPERase Inhibitor (Thermo Fisher AM2696)
4 mL 500 mM EDTA pH 8 (FC = 20 mM)	500 µL 500 mM EDTA pH 8.0 (FC = 1 mM)	200 µL 500 mM EDTA pH 8.0 (FC = 1 mM)	1.25 µL 0.1 M dithiothreitol
2 g LiDS (L9781-5G) (FC = 2 % w/v)	0.25 g LiDS (FC = 0.1 % w/v)	96.92 mL NFW	5 µL 5 M betaine (Sigma PCR-grade B0300-5VL)
1 mL 1 M DTT (P2325) (FC = 10 mM)	242.31 mL NFW		0.15 µL 1 M MgCl2
62.5 mL NFW			0.38 µL 100 µM TSO
			1.25 µL SMARTScribe reverse transcriptase (Takara Clontech 639538)

Table 2.1 Reaction mixes used in low-input RNA-sequencing library preparation

Following RT of the samples, 25 µL of nuclease-free water (NFW) was added to each well of the 96-well plate and a 0.8:1 Ampure XP clean-up (Beckman Coulter A663882) was performed using a Zephyr (PerkinElmer). The material was then eluted in 10 µL of 10 mM Tris-HCl (pH 7.5) and 13 µL PCR master mix was added to the solution (12.5 µL of 2x KAPA HiFi hotstart and 0.5 µL of 10 µM ISPCR primer). A further PCR reaction was carried out for amplification (Figure 2.2); due to the

variability in input RNA quantity for this reaction, the number of PCR cycles used was increased for low input samples (see Figure 2.2 for range).

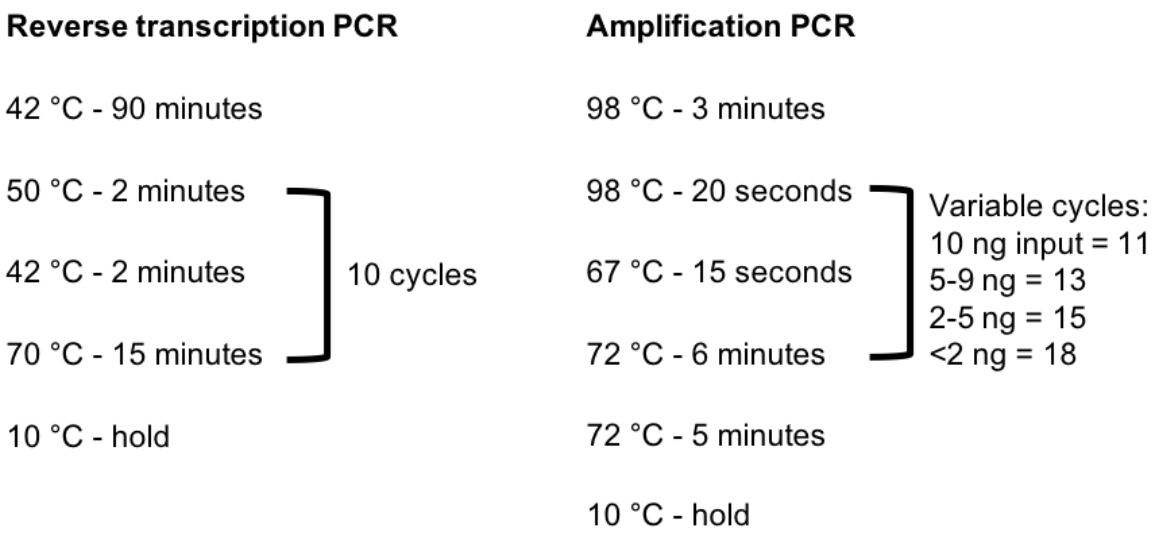


Figure 2.2 PCR reactions in low-input RNA-sequencing library preparation

After the PCR reaction, a further 25 µL of NFW was added to samples and a 0.8:1 Ampure XP clean-up was carried out before elution in 20 µL of 10 mM Tris-HCl (pH 8.0). cDNA was then quantified with the Quant-iT High Sensitivity kit (Thermo Fisher Q33120), according to the manufacturer's instructions. Samples were read on a BMG Pherastar. 4 ng of cDNA was diluted with 10 mM Tris-HCl (pH 7.5) to a volume of 9.5 µL. 5 µL of a 3x tagmentation buffer (99 mM Tris acetate, 198 mM potassium acetate, 30 mM magnesium acetate and 48 % v/v N,N-dimethylformamide) and 0.5 µL of TDE1 were then added and mixed before samples were incubated at 55 °C for 5 minutes. Tagmentation was then halted by the addition of 2.5 µL of stop buffer (220 mM EDT and 1.1% w/v sodium dodecyl sulphate), with samples then incubated at room temperature for 10 minutes. Tagmented cDNA was then diluted to a volume of 50 µL with 10mM Tris-HCl (pH 7.5) and purified with a 2:1 ratio of Ampure XP beads. The cDNA samples were eluted in 7 µL of 10mM Tris-HCl (pH 7.5) and then amplified and sample indexed using PCR. Briefly, the eluted 7 µL of tagmented cDNA was added to 2.5 µL of i5 index adapter and 2.5 µL of i7 index adapter from the Nextera XT index kit v2 set A , 0.25 µL of 50 µM PC1 primer, 0.25 µL of 50 µM PC2 primer and 12.5 µL of 2x KAPA HiFi polymerase. Mixed samples were then incubated

at 72 °C for 3 minutes, 98 °C for 30 seconds, followed by 9 cycles at 98 °C for 15 seconds, 62 °C for 30 seconds and 72 °C for 30 seconds, followed by a final extension at 72 °C for 3 minutes. Libraries were purified using a 0.8:1 ratio of Ampure XP beads and the final individual libraries were eluted in 20 µL of 10mM Tris-HCl (pH 7.5). Samples were then pooled together (three independent pools) at equal cDNA concentrations and submitted for 75 bp paired-end sequencing.

2.2.4 Initial processing and quality control of sequencing data

Initial processing of sequencing was carried out by Dr Natsuhiko Kumasaka. Prior to alignment adapter trimming of Tn5 transposon and PCR primer sequences was carried out using the skewer package[220]. Both bulk and smart-seq2 sequencing data were aligned using the STAR package[221], version 2.5.3a, using ENSEMBL human gene assembly 90 as the reference transcriptome. Samples were then quantified with featureCounts[222], version 1.5.3. Genotype information collected from patients was then used to check for sample swaps or mixing of samples that may have occurred during processing. Following QC for sample swaps and mixes, 109 patient samples were used in bulk analysis.

For single-cell analysis each individual cell was passed through a further quality control pipeline to remove poor quality cells from the dataset. The final thresholds used were: number of expressed genes > 500, number of fragments > 10000, < 20 % mitochondrial genes and the percentage of fragments mapped to the top 100 highly expressed genes is < 70 %. Demuxlet [223] was used to remove doublets from two different patients with different genetic backgrounds from within the sample. Following QC analysis 9538 cells from 129 patients were taken forward for further analysis.

2.2.5 Comparison of bulk data to publicly available datasets

Processed bulk microglia RNA-sequencing data was combined with publicly available datasets from other cell types: brain tissue from The Genotype-Tissue Expression (GTEx) Project (The data used for the analyses described in this thesis were obtained from the GTEx Portal), monocytes from the BLUEPRINT consortium (this study makes use of data generated by the BLUEPRINT Consortium) and a collection

of publicly available *in-vitro* model data (see section 3.2.1 for data references). Count tables were combined and converted into counts per million (CPM) and Uniform Manifold Approximation and Projection (UMAP) analysis was run using Seurat's RunUMAP function with the following parameters: 5 PCs, 30 nearest neighbours and a minimum distance set to 0.3.

2.2.6 Classification of microglial cells using publicly available datasets

Full descriptions of the single cell data analysis carried out by Dr Natsuhiko Kumasaka can be found in the preprint of the manuscript describing this work [210] but the methodology will be summarised below.

Gene count data for single cell datasets of 68k peripheral blood mononuclear cells (PBMCs)[224] and 15K unsorted brain cells[225] were downloaded from publicly available sources and all datasets (including the data collected for this study) were converted to Counts Per Million (CPM).

A latent factor linear mixed model was used, with the 3 studies treated as random effects, to obtain 12 latent factors. These factors were then used to run Uniform Manifold Approximation and Projection (UMAP) analysis. The publicly available datasets also included pre-determined cell type classification and these classifications were then used to identify microglia cells from within our unclassified dataset. 8,662 cells were identified as microglia and taken forward for further analysis.

2.2.7 Variance components analysis

Variance components analysis was used to determine how clinical and technical factors within the dataset impacted gene expression. Count data (log(TPM+1)) across all genes whose TPM>0 for at least 10% of cells was used in a linear mixed model to estimate variation. 13 known factors (patient, number of expressed genes per cell, pathology, plate ID, ERCC percentage, number of fragments, plate position, age, mitochondria RNA percentage, brain region, brain hemisphere, ethnicity and sex) were fitted as random effects with idependent variance parameters.

2.2.8 Clustering of single cell data, differential expression and clinical metadata links

A latent linear mixed model was again used to estimate latent factors for downstream dimensionality reduction and clustering on only the microglia cells identified through the methodology described in section 2.2.6. The 13 factors described in section 2.2.7 were included in the model to control for potential confounding between the known factors and unknown heterogeneity within the dataset. The first 15 latent factors were then used within Shared Nearest Neighbour Clustering (as run in Seurat version 3.0.2) with a resolution parameter of 0.2. The first 15 latent factors were also used to run UMAP analysis.

The same linear mixed model used for variance component analysis was also used for differential expression analysis, with the addition of the four subpopulations fitted as a random effect. The model was fit on a gene-by-gene basis and across each factor. If the factor of interest was numerical (i.e. age) Bayes factor of effect size was computed by comparing the full model and the model without the factor of interest. If the factor of interest was categorical with x levels (i.e. pathology with 5 levels), samples were partitioned into any of two groups. There were 2^x-1 contrasts which were tested against outputs when removing the factor of interest from the model to calculate Bayes factors. Bayes factors were then used within a finite mixture model to calculate the posterior probability as well as the local true sign rate (*lstr*). Lstr values were used to identify differentially expressed genes (*lstr* > 0.5 unless stated otherwise)

2.2.9 Pathway enrichment analysis

I then used gProfiler[226], version e94_eg41_p11_36d5c99 with significance determined at a 5% FDR, to estimate the significance of enrichment across defined pathways, through a hypergeometric distribution model. Gene lists were established from the differential expression studies described above (section 2.2.8).

2.3 Quality control analysis across datasets

2.3.1 Bulk RNA-sequencing quality control

Before running downstream analysis pipelines, extended quality control analysis was run on all samples that passed the technical quality control (109 samples in bulk dataset). In bulk data initially correlation analysis was run between all samples (averaged across all genes for each sample). These correlations were then compared to those observed in BLUEPRINT monocytes and a small primary microglia dataset. Figure 2.3 is a heatmap of the correlation coefficients across all samples. While correlation coefficients between the monocyte and peadiatric microglial samples are high and consistent across all samples, within the adult primary microglia dataset there is a much larger amount of variability amongst samples.

After looking at variability amongst the samples collected as part of this study, I wanted to compare global expression patterns in our bulk RNA-seq dataset to other large scale datasets in other similar cell types. I used UMAP analysis to understand the transcriptional similarities between primary microglia, brain tissue from GTEx, monocyte data from BLUEPRINT and a selection of *in-vitro* models (note: for detailed analysis of primary microglia versus *in-vitro* models please refer to Chapter 3, sect). The UMAP analysis plot (UMAP 1 vs UMAP 2) highlights how samples group together based on their transcriptional similarities (Figure 2.4).

At the top of the plot the brain tissue samples split into two distinct groups, with cerebellum tissue on the left and the remaining regions on the right. The three remaining distinct clusters represented: monocytes, primary microglia and *in-vitro* models. The separation of the microglia samples from other large scale datasets suggested a transcriptional signature in microglia that is not captured by other available datasets. The primary microglia data collected as part of this study, also clustered with small numbers of samples from other fresh human primary microglia datasets. This highlights that despite higher levels of variation between samples

(Figure 2.3), the microglia collected as part of this study were transcriptionally similar to other publicly available datasets.

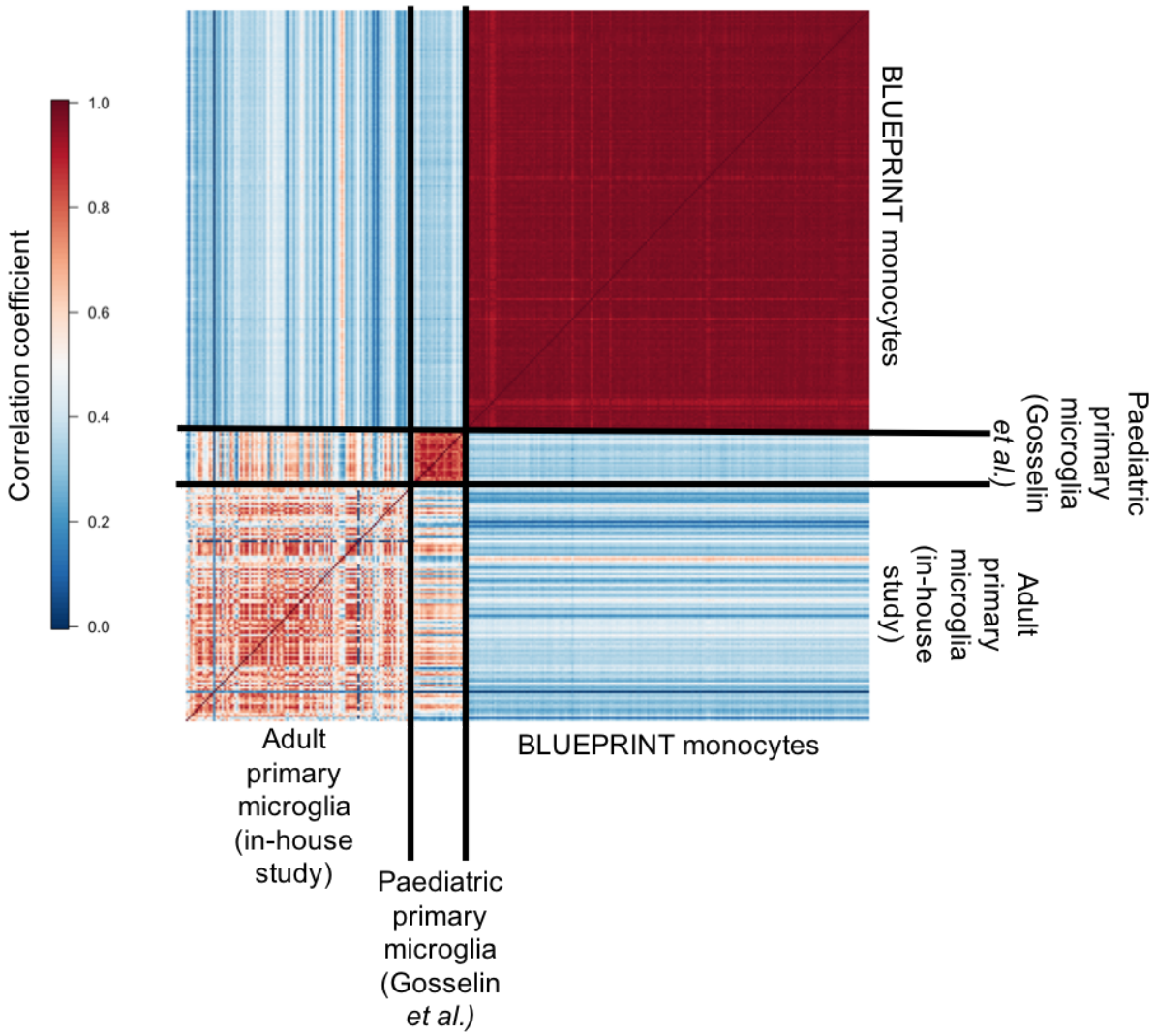


Figure 2.3 Heatmap of correlation of bulk RNA-seq gene expression between samples in primary microglia and BLUEPRINT monocytes

Average Spearman's rank correlations across all genes of gene expression for each sample in the in-house primary microglia dataset, fresh paediatric microglia samples from a published dataset[171] and BLUEPRINT monocyte dataset.

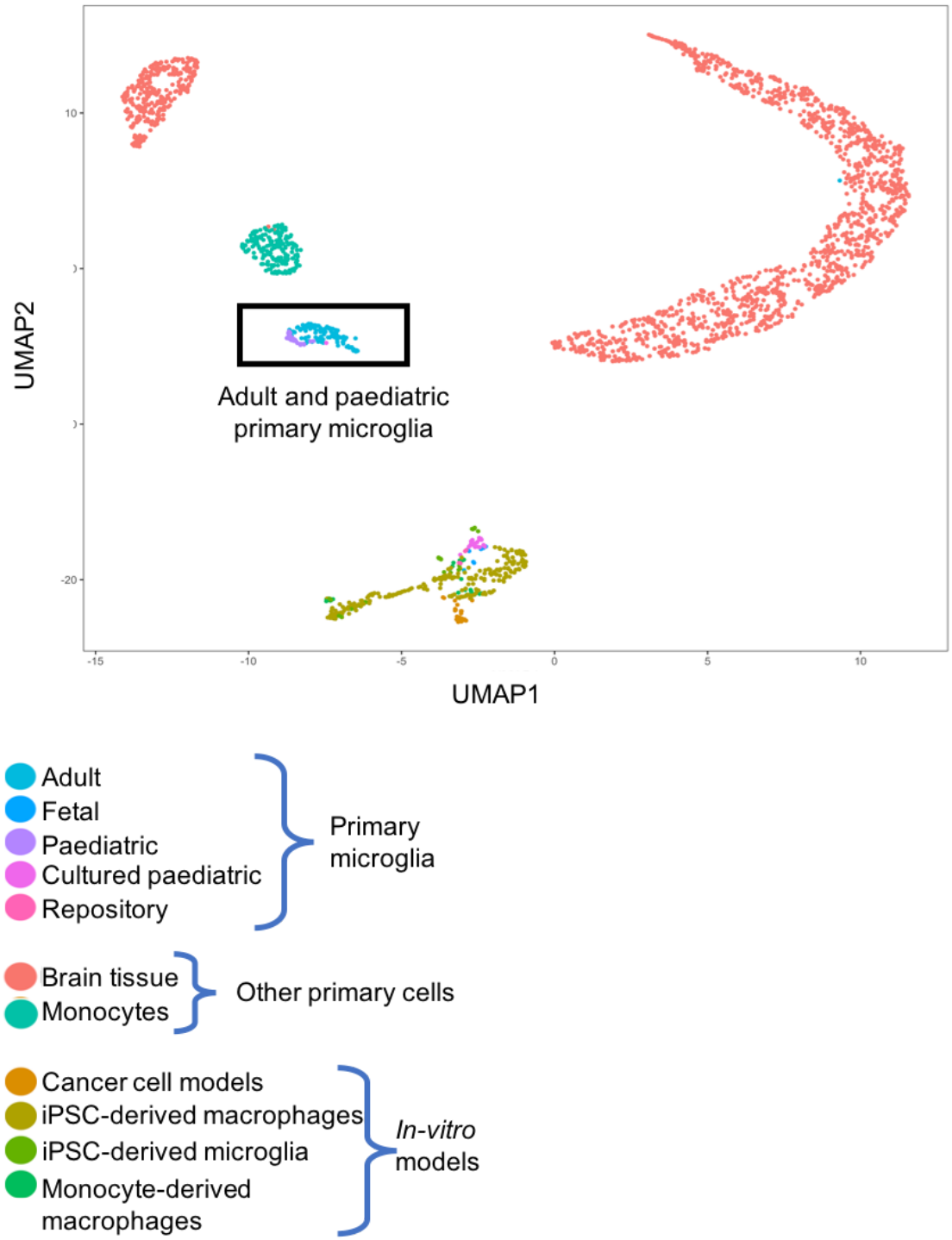


Figure 2.4 UMAP analysis of bulk primary microglia data and publicly available RNA-sequencing datasets

UMAP analysis from Seurat's RunUMAP function on a collection of publicly available datasets. Analysis run using the following parameters: PCs=15, n_neighbours = 30 and min_dist = 0.3. Samples highlighted as "Adult and paediatric primary microglia" included data from this study and publicly available datasets (section 3.2.1 for full details).

2.3.2 Metadata comparison

As much of the analysis completed in this chapter focuses on understanding the effect of clinical phenotypes on microglial transcriptomes, I initially wanted to ensure that there were no major confounding groups of clinical phenotypes. I, therefore, compared the number of patients across pairs of clinical phenotypes in both the single cell and bulk patient groups (Figure 2.5 and 2.6), all pairwise comparisons for the four meta group (age, sex, brain region and clinical pathology) are shown. Within both the bulk and single cell, patient groups clinical pathology and brain region were confounded because trauma patients were only found in one brain region.

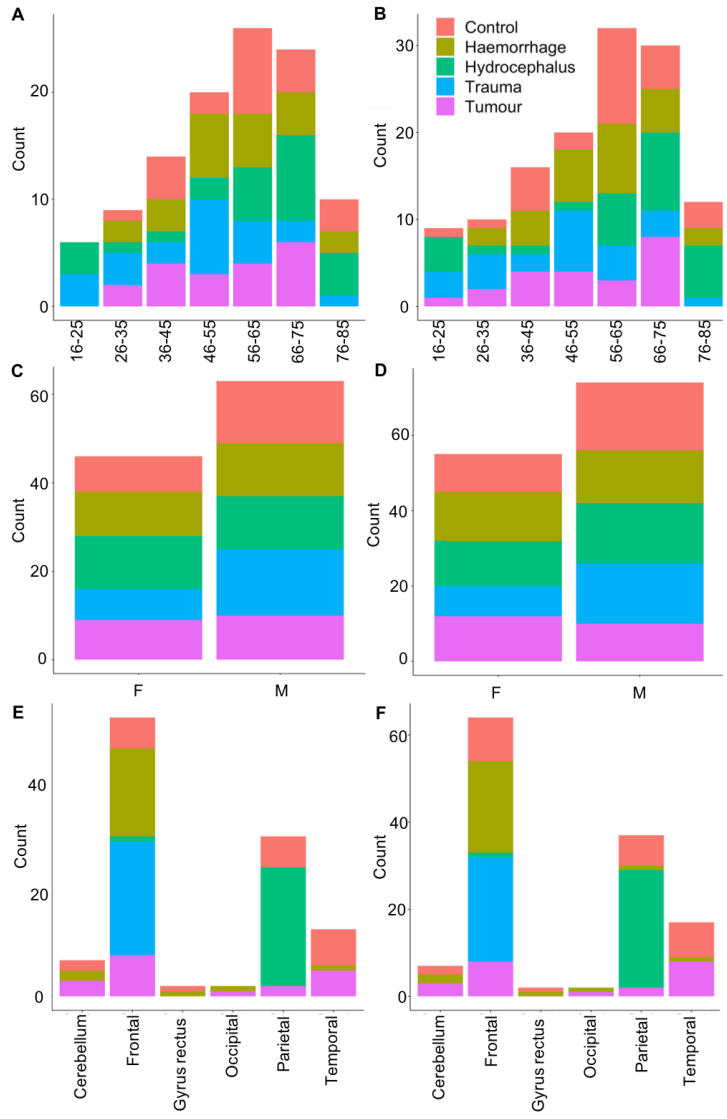


Figure 2.5 Frequency of patients from metadata groups within the bulk (A, C and E) and single cell (B, D and F) RNA-seq datasets

Numbers of patients in different age ranges (A and B), sexes (C and D) and brain regions (E and F) subdivided by clinical pathology (colour).

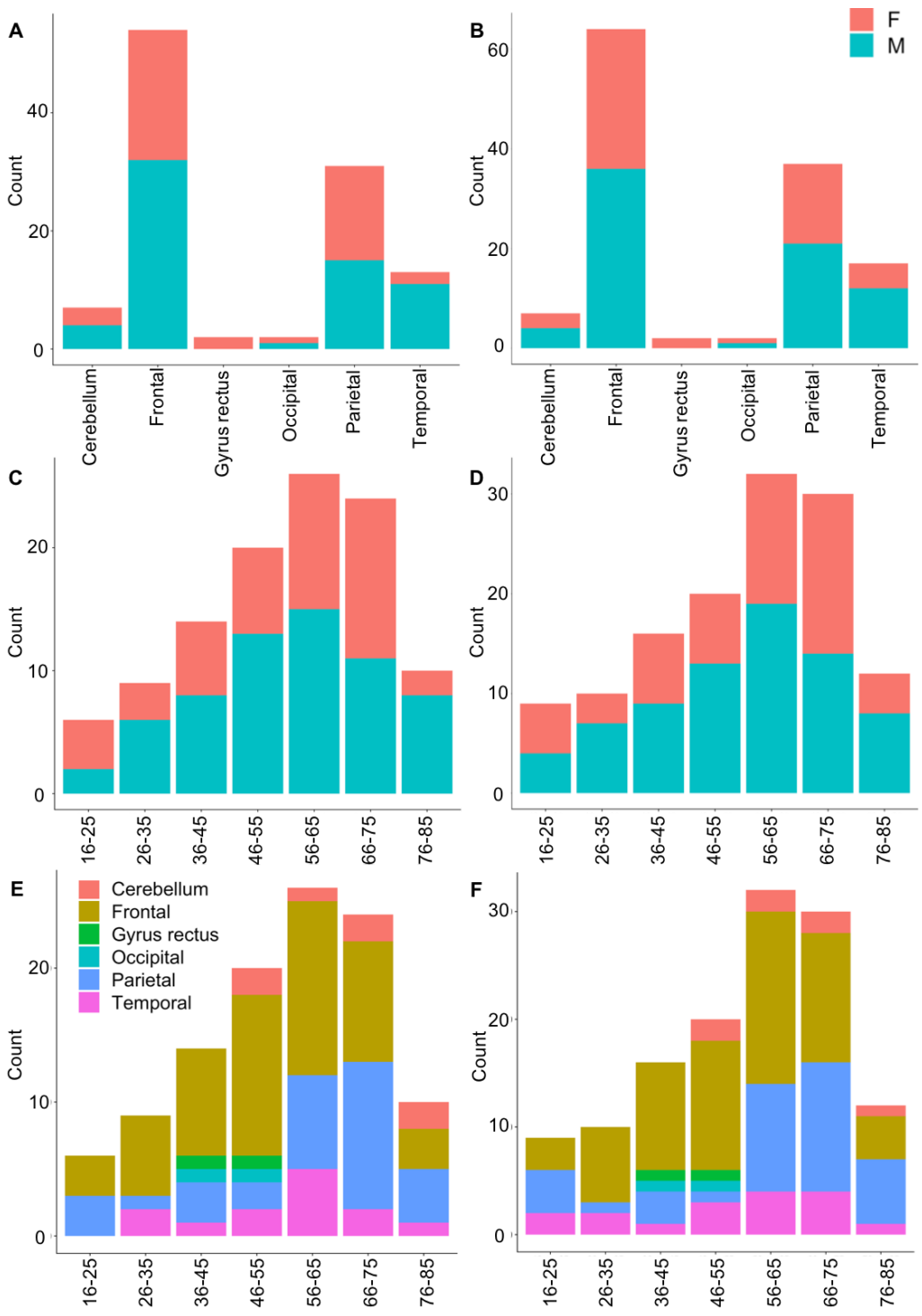


Figure 2.6 Frequency of patients from metadata groups within the bulk (A, C and E) and single cell (B, D and F) RNA-sequencing datasets

Numbers of patients with samples from different brain regions (A and B) and age ranges (C, D, E and F) subdivided by sex (A, B, C and D) and brain region (E and F).

2.4 Single cell clustering and identification of sub-populations

2.4.1 Comparison to publicly available single cell datasets

Initially we compared our microglia single cell data to two publicly available datasets, 68K peripheral blood mononuclear cells[224] (PBMCs) and 15K unsorted brain cells[225] (Figure 2.7). This allowed for the identification of infiltrating blood derived cells or contaminating neuronal cells while also providing a comparison of our sorted microglial cells to an unsorted dataset.

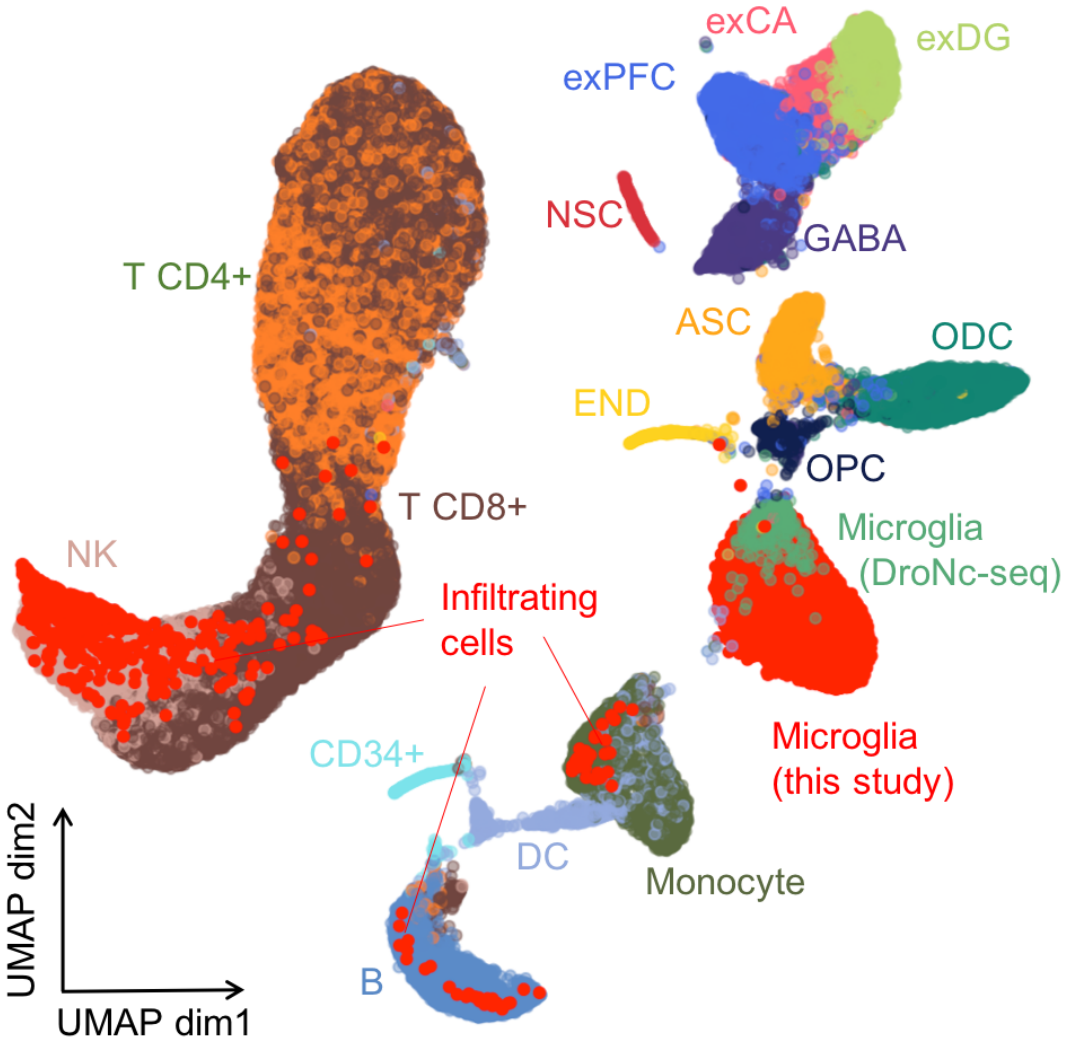


Figure 2.7 UMAP analysis of microglia single cell data and publicly available PBMC and whole brain tissue single cell datasets

Cells collected as part of this study coloured in red. Cell type annotations were obtained from original manuscripts: glutamatergic neurons from the PFC (exPFC);

pyramidal neurons from the hip CA region (exCA); GABAergic interneurons (GABA); granule neurons from the hip dentate gyrus region (exDG); astrocytes (ASC); oligodendrocytes (ODC); oligodendrocyte precursor cells (OPC); neuronal stem cells (NSC); endothelial cells (END); dendritic cell (DC); B cell (B); hematopoietic progenitor cell (CD34+); NK T cell (NK). Plot generated by Dr Natsuhiko Kumasaka.

A total 8,662 cells from our single cell dataset clustered with microglia identified within the unsorted brain cell dataset (see Table 2.2 for breakdown of identified cells in the dataset). Alongside the microglial cells identified a small fraction of the single cells collected as part of this study appeared transcriptionally similar to PBMC cells, specifically NKT cells, monocytes and B cells. These cells could represent either infiltrating cells that have entered the brain following disruption to the BBB or intravascular contamination of the tissue that occurred during the collection.

Cell Type	Number of cells	Number of patients
Microglia	8662	127
NKT cells	799	91
Monocyte	46	18
B cell	28	16

Table 2.2 Cell numbers and number of patients represented in each immune cell type collected.
Cell type classification determined by UMAP analysis and comparison to publicly available datasets that had been previously classified.

The cells identified as microglia also expressed known marker genes *P2RY12*, *CX3CR1* and *TMEM119* (Figure 2.8). These 8,622 cells were therefore taken forward for further analysis.

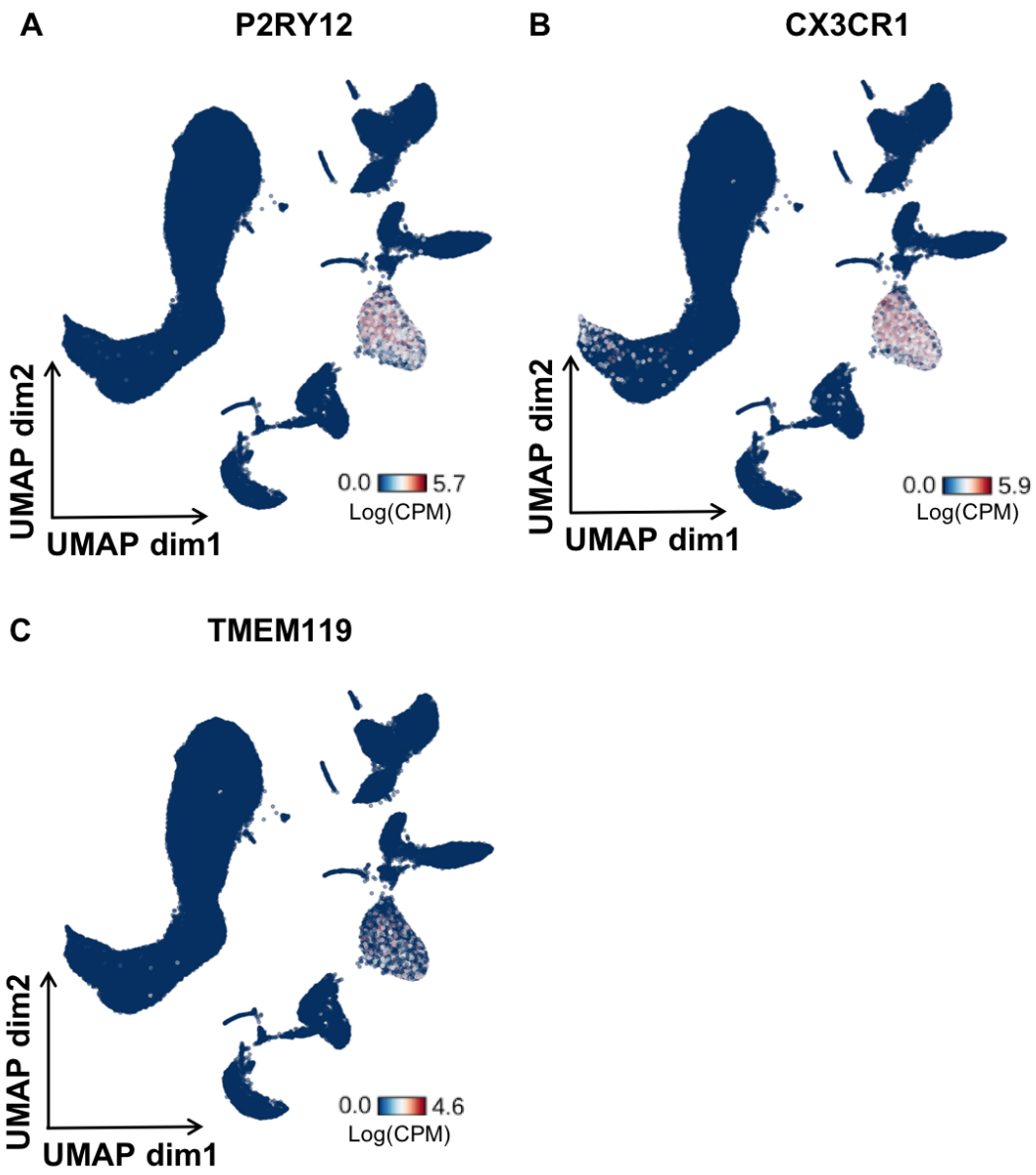


Figure 2.8 UMAP analysis of microglia single cell data and publicly available PBMC and whole brain tissue single cell datasets

Cells coloured by expression (CPM) of microglial marker genes *P2RY12* (A), *CX3CR1* (B) and *TMEM119* (C). Plot generated by Dr Natsuhiko Kumasaka.

2.4.2 Clustering of microglial cells and cluster maker analysis

Clustering of the microglia highlighted a relative homogeneity between cells although 4 transcriptionally distinct clusters were identified (Figure 2.9). A linear mixed model, with the cluster membership fitted as a random effect, was used to identify differentially expressed genes between cluster groups.

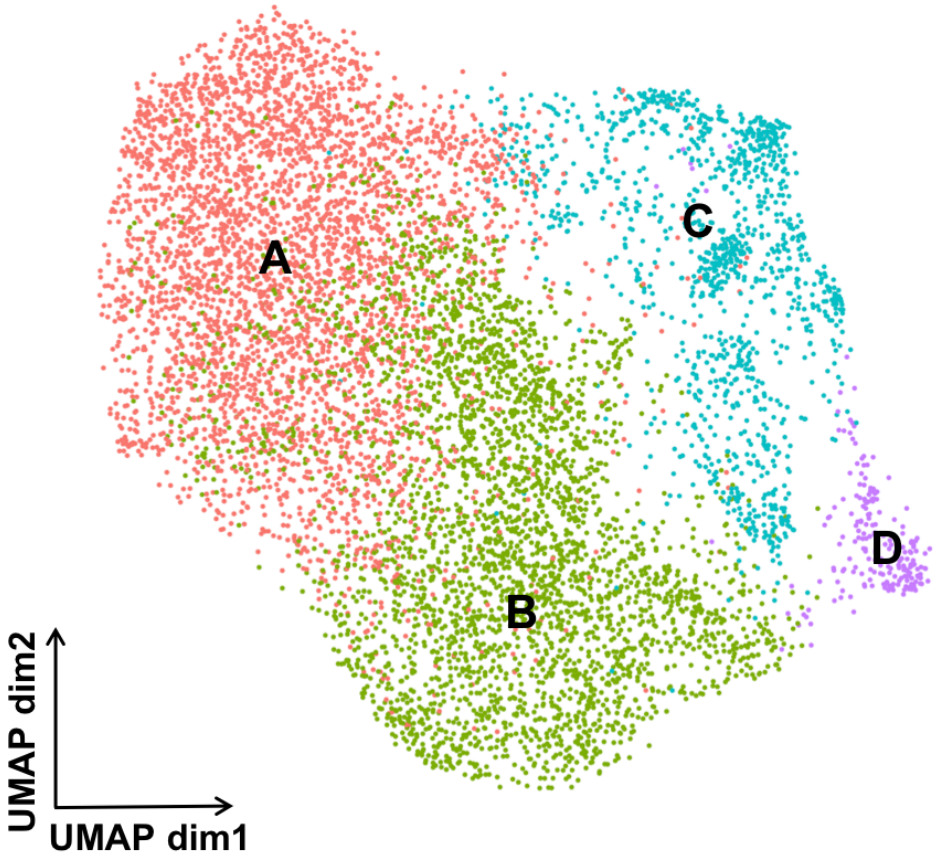


Figure 2.9 UMAP analysis of microglia cells from this study identified from previous analysis (Figure 2.7)

Cells coloured by cluster membership as determined by Louvain clustering (see section 2.2.8 for full clustering methodology).

Figure 2.10 highlights some of the cluster markers identified as part of this analysis and Table 2.3 shows the top 5 most enriched GO terms for cluster marker genes (identified as any gene with a LTSR value of >0.5 when comparing expression of cells in one cluster to all other cells).

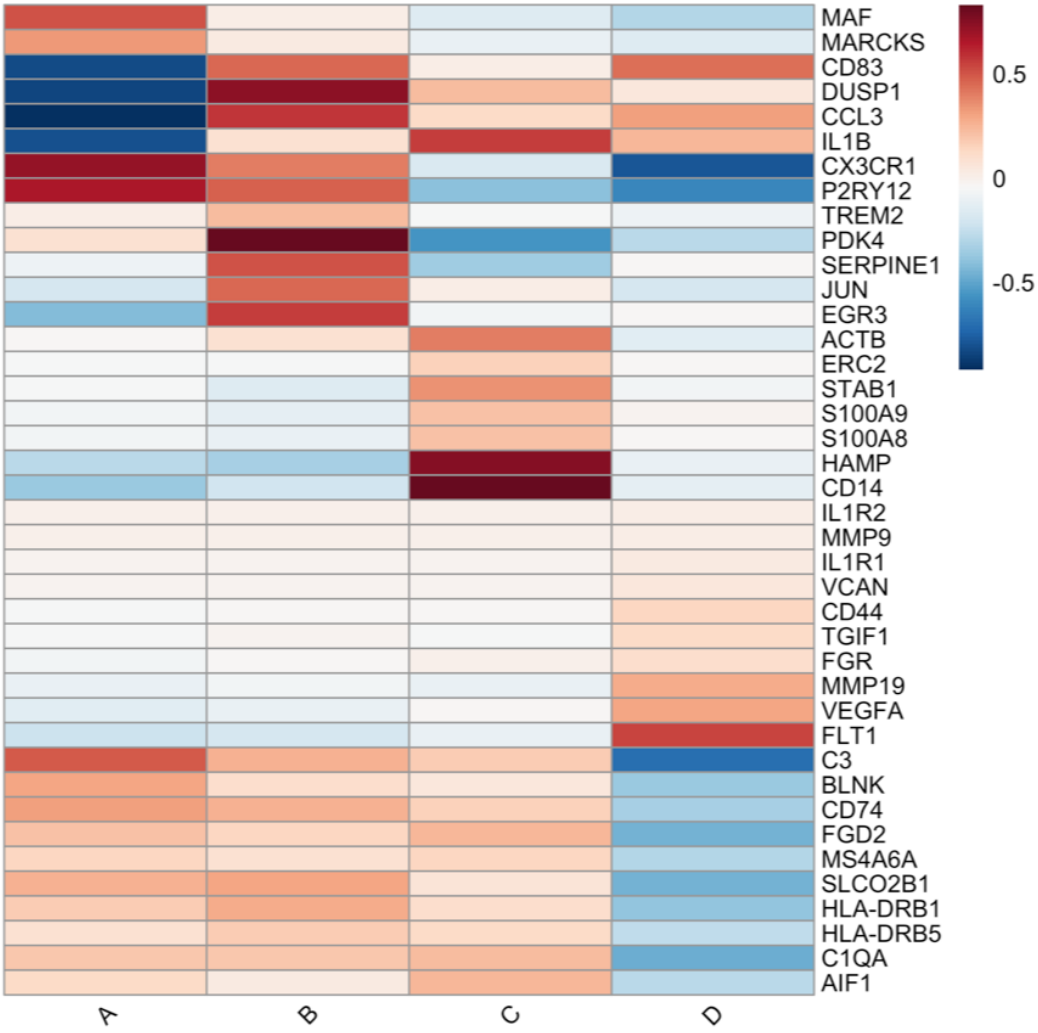


Figure 2.10 Cluster maker genes for microglia single cell data

Averaged, across cells in each cluster, normalised expression level (defined as the posterior mean of pathology random effect term, see section 2.2.8 for full details) of differentially expressed genes at the local true sign rate (*ltsr*) greater than 0.9.

As demonstrated in Figure 2.10 cells in clusters A and B had higher expression of microglial marker genes *P2RY12* and *CX3CR1* than cells in clusters C and D. Cells within cluster A also had significantly reduced expression of immune activation marker genes, like *IL1B* and *CCL3*, when compared to all other cells. GSEA of the genes differentially expressed within this cluster identified an enrichment of metabolic and translational processes. Cells in cluster A were therefore identified as homeostatic microglial cells with those in other clusters representing cells in differing activation states.

As well as increased expression of marker genes, cells associated with cluster B had increased expression of activation genes such as *JUN* and *EGR3*. These often represent early activation patterns of macrophage cells and therefore cluster B may represent a population of cells moving towards an activated state. Further investigation, using techniques such *in-situ* single cell transcriptomics, would be needed to confirm that these cells arise in the brain and are not artificially activated by the tissue processing used in this study.

Cells in cluster C had significantly increased expression of genes such as *CD14*, *ACTB* and *ERC2*. One of the other marker genes associated with cells in this cluster is *HAMP* which encodes for hepcidin protein, a key molecule in iron homeostasis. Iron homeostasis has been linked to multiple brain disorders including ischemia, cancer and Alzheimer's disease[227]. Enrichment analysis of marker genes associated with this cluster showed significant enrichment for terms such as immune response and immune system process, highlighting a clear activation pattern within these cells.

Like in cells associated with cluster C, those in cluster D were also enriched for terms such as immune system process. However, gene markers for cells in cluster D were also enriched for cell migratory and communication terms. Cluster D is also characterised by expression of *VEGF* and a receptor for the molecule, *FLT1*. FLT1 and VEGF have been shown to be important in angiogenesis in the brain particularly following traumatic brain injury[228,229]. Recent evidence has also suggested a potential role for VEGF response in microglial chemotaxis to amyloid beta, a key protein in AD [230].

Cluster	GO ID	Term name	Padj
A	GO:0016071	mRNA metabolic process	$6.22e^{-14}$
	GO:0006413	translational initiation	$6.22e^{-14}$
	GO:0006886	intracellular protein transport	$4.74e^{-13}$
	GO:0006613	cotranslational protein targeting to membrane	$4.74e^{-13}$
	GO:0070972	protein localization to endoplasmic reticulum	$5.16e^{-13}$
B	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	$1.66e^{-27}$
	GO:0006613	cotranslational protein targeting to membrane	$3.44e^{-27}$

	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	$1.06e^{-26}$
	GO:0045047	protein targeting to ER	$1.83e^{-26}$
	GO:0072599	establishment of protein localization to endoplasmic reticulum	$3.89e^{-26}$
C	GO:0006955	immune response	$3.34e^{-14}$
	GO:0002376	immune system process	$1.80e^{-13}$
	GO:0002252	immune effector process	$1.50e^{-08}$
	GO:0002682	regulation of immune system process	$1.50e^{-08}$
	GO:0043299	leukocyte degranulation	$2.74e^{-08}$
D	GO:0002376	immune system process	$2.48e^{-25}$
	GO:0048583	regulation of response to stimulus	$6.50e^{-22}$
	GO:0070887	cellular response to chemical stimulus	$5.78e^{-21}$
	GO:0007154	cell communication	$1.31e^{-20}$
	GO:0050896	response to stimulus	$1.79e^{-20}$

Table 2.3 Top enriched biological process terms for cluster marker genes

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Five most significantly enriched biological process terms for genes determined as cluster markers at the local true sign rate (*ltsr*) greater than 0.9 (section 2.2.8 for full details).

2.5 Clinical metadata and microglial transcriptome signatures

2.5.1 Variance components analysis

The large sample size of this study across a variety of patients also allowed us to study how a range of biological factors impact microglial gene expression. Variance components analysis highlights how much variability in gene expression can be explained by different biological and technological factors. Figure 2.11 shows that individual patients were the largest driver of variation within the dataset, this may represent the effect of genetic background on gene expression but could also be in part due to unknown factors that weren't collected as part of this study.

Of the non-technical factors, clinical pathology was the largest driver of variation contributing to more variation in gene expression than the other biological factors combined. The variance components analysis also highlighted how technical factors can impact gene expression and why they need to be accounted for in downstream analysis.

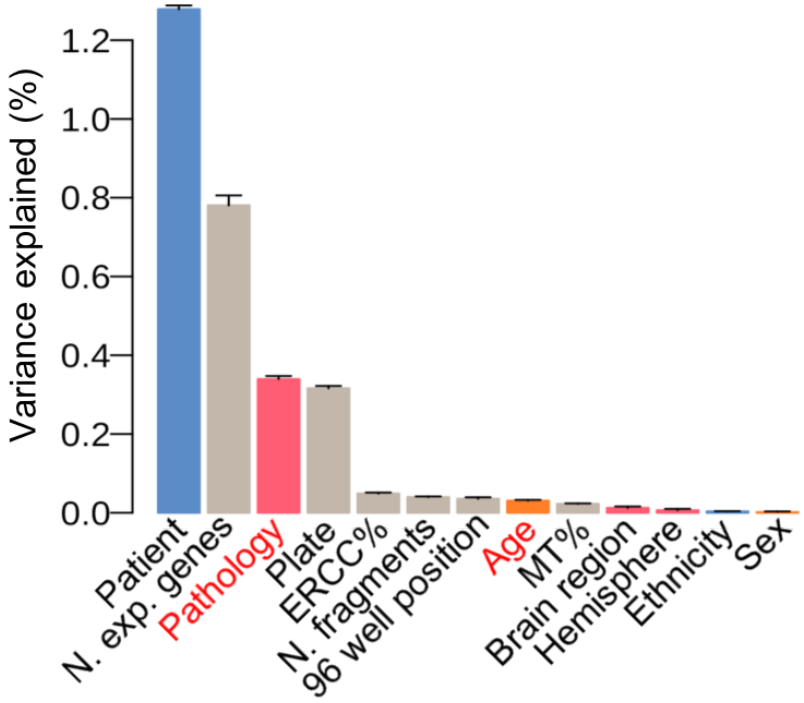


Figure 2.11 Variance components analysis

Proportion of variance explained by both biological and technical factors collected as part of this dataset. Plot generated by Dr Natsuhiko Kumasaka.

2.5.2 Gene expression linked to clinical metadata

Due to the size of the dataset collected as part of the study, we were able to determine genes whose expression is affected by clinical factors, while controlling not just for the other interlinked clinical factors but also technical factors that can influence gene expression.

The variance component analysis highlighted that pathology was the largest known clinical factor driving variation in this dataset. We therefore ran enrichment analysis to understand if cells part of different clusters were enriched for patients with certain

clinical pathologies. Figure 2.12 demonstrates the log odds ratio for enrichment of clinical pathologies in each cluster.

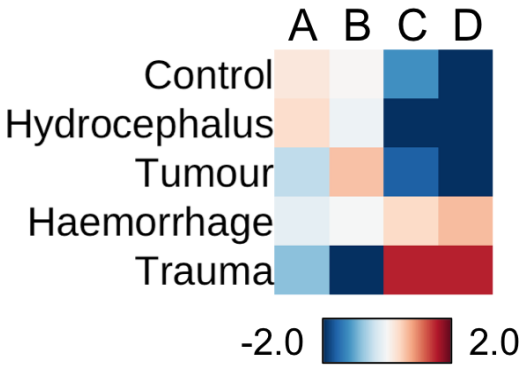


Figure 2.12 Odds ratios from Fisher's exact tests across clinical pathologies for each cluster.

The number of cells contributing to each cluster, from each pathology group were used to run two-tailed Fisher's exact tests. Results displayed show Odds Ratios for each test. Plot generated by Dr Natsuhiko Kumasaka.

Enrichment analysis showed that clusters C and D, those with distinct activation patterns, were significantly enriched for trauma patients, as well as haemorrhage patients, and cluster B was enriched for tumour patients (OR=4.9, $P=7.6x10^{-169}$).

While pathology was the largest clinical factor driving variation, other factors such as age, brain region and sex also contributed to variance within the dataset and therefore differentially expressed genes were calculated across clinical groups, controlling for other factors.

Table 2.4 summarizes the top 5 genes whose expression in microglia was positively or negatively correlated with age as well as the top 5 enriched GO terms for all correlated genes. Gene set enrichment analysis of the 156 genes whose expression was positively correlated, highlighted a significant enrichment in immune activation genes suggesting that microglia may take on a more active phenotype as we age.

There were 144 genes whose expression was negatively correlated with age, including microglia marker genes *P2RY12* and *CX3CR1*. Gene set enrichment

analysis highlighted an enrichment of genes involved in cell migration and regulation of locomotion ($p = 1.974 \times 10^{-5}$).

Genes and GO terms positively correlated with age				
Gene		**GO ID**	**Term name**	**Padj**
HLA-DRA		GO:0002376	immune system process	$6.60e^{-20}$
HLA-DRB1		GO:0006955	immune response	$7.57e^{-20}$
PADI2		GO:0001775	cell activation	$4.65e^{-18}$
MS4A6A		GO:0006952	defense response	$2.62e^{-17}$
HLA-DPA1		GO:0045321	leukocyte activation	$5.20e^{-17}$
Genes and GO terms negatively correlated with age				
Gene		**GO ID**	**Term name**	**Padj**
P2RY12		GO:0030334	regulation of cell migration	$1.92e^{-05}$
PDK4		GO:0070887	cellular response to chemical stimulus	$1.92e^{-05}$
CH25H		GO:0010033	response to organic substance	$1.92e^{-05}$
C3		GO:0051270	regulation of cellular component movement	$1.92e^{-05}$
CSF1R		GO:1901701	cellular response to oxygen-containing compound	$1.92e^{-05}$

Table 2.4 Top 5 genes and enriched biological process terms associated with age

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Five most significantly enriched biological process terms for genes with local true sign rate (*ltsr*) greater than 0.5.

Differential expression focussing on brain region, highlighted varying levels of heterogeneity across different areas of the brain. There were over 400 genes with higher expression in microglia originating from the occipital lobe, whereas only two genes were more highly expressed in microglia sourced from the frontal lobe. Pathway enrichment analysis showed genes more highly expressed in occipital microglia were enriched for immune activation pathways but also cell motility (GO:0048870) and migration (GO:0016477).

Region	Number of DE genes		GO ID	Term name	Padj
Occipital	441		GO:0006955	immune response	$4.15e^{-18}$
			GO:0002376	immune system process	$1.69e^{-15}$
			GO:0002252	immune effector process	$1.87e^{-14}$
			GO:0019221	cytokine-mediated signaling pathway	$3.05e^{-14}$
			GO:0034097	response to cytokine	$6.39e^{-14}$
Cerebellum	51		GO:2001242	regulation of intrinsic apoptotic signaling pathway	0.00170
			GO:0090288	negative regulation of cellular response to growth factor stimulus	0.00170
			GO:0048583	regulation of response to stimulus	0.00170
			GO:0051091	positive regulation of DNA-binding transcription factor activity	0.00170
			GO:0002376	immune system process	0.00260
Temporal	36		GO:0006614	SRP-dependent cotranslational protein targeting to membrane	$3.42e^{-20}$
			GO:0006613	cotranslational protein targeting to membrane	$3.44e^{-20}$
			GO:0045047	protein targeting to ER	$7.41e^{-20}$
			GO:0072599	establishment of protein localization to endoplasmic reticulum	$9.05e^{-20}$
			GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	$1.83e^{-19}$
Parietal	7		N/A		
Frontal	2				

Table 2.5 Top 5 genes and enriched biological process terms associated with brain region

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Five most significantly enriched biological process terms for genes with local true sign rate (*ltsr*) greater than 0.5.

There were fewer genes whose expression differed significantly based on sex, 55 with increased expression and 95 with increased expression in males. Table 2.6 shows the top genes with higher expression in males or females alongside the enrichment terms.

Genes and enriched GO terms in males				
Gene		**GO ID**	**Term name**	**Padj**
HLA-DQB1		GO:0006614	SRP-dependent cotranslational protein targeting to membrane	$4.25e^{-70}$
EEF1A1		GO:0006613	cotranslational protein targeting to membrane	$3.05e^{-69}$
HLA-DRA		GO:0045047	protein targeting to ER	$1.63e^{-67}$
RPL37		GO:0072599	establishment of protein localization to endoplasmic reticulum	$7.41e^{-67}$
RPS3A		GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	$1.74e^{-65}$
Genes and enriched GO terms in females				
Gene		**GO ID**	**Term name**	**Padj**
B2M		GO:0098542	defense response to other organism	$1.32e^{-09}$
H2BC8		GO:0006952	defense response	$2.09e^{-09}$
AC011586.2		GO:0051707	response to other organism	$5.36e^{-09}$
H4C5		GO:0045814	negative regulation of gene expression, epigenetic	$5.36e^{-09}$
H2BC3		GO:0009607	response to biotic stimulus	$5.36e^{-09}$

Table 2.6 Top 5 genes and enriched biological process terms associated with sex

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Five most significantly enriched biological process terms for genes with local true sign rate (*ltsr*) greater than 0.5.

2.6 Microglia and disease

2.6.1 Microglial gene expression and Alzheimer's disease (AD)

Next, I examined expression of known AD genes across the microglia dataset. I included familial AD genes (*APP, PSEN1* and *PSEN2*), and a selection of genes associated with late-onset AD. The late-onset AD genes included the large effect size gene and APOE rare missense variant genes (*TREM2*, *PLCG2* and *ABI3*). While these genes have been definitively linked to AD, many complex disease risk variants for late-onset AD identified by genome wide association studies (GWAS) lie in non-coding regions of the genome[134,136,137,231]. This presents a problem for expression analysis, because linking these signals to candidate genes is challenging. One approach to identifying the candidate causal genes is colocalization, which compares association signals between a GWAS and those from an expression quantitative trait loci (eQTL). I examined the expression of a set of genes identified as candidate causal AD risk genes identified as part of the same study described in this chapter (eQTL analysis carried out by Dr Natsuhiko Kumasaka). This gene set included: *BIN1*, *MEF2A*, *PTK2B*, *CASS4*, *CD33* and *EPHA1-AS1*.

Table 2.7 summaries whether these genes, and genes that have been identified as the "nearest gene" to an AD risk variant in more than one GWAS study (see Table 1.1), had increased expression within specific microglia clusters or between males and females. I also looked at whether the AD genes were positively or negatively correlated with age or whether expression was increased in a particular brain region. Only 4 of 30 the AD-linked genes studied here showed a significant correlation between expression level and age and the majority of the AD linked genes showed no differential expression across clusters. However the 6 genes whose expression was increased within specific clusters were within the "activated" populations while none were increased in the homeostatic population (cluster A).

Nearest Gene	Cluster marker?	Higher expression in male or females?	Higher expression in specific brain region?	Correlated with age?
APP	D			
PSEN1				
PSEN2				
APOE				Positively
TREM2	B	Male	Occipital	
PLCG2				
ABI3	C			
BIN1				Negatively
MEF2A			Occipital	
CASS4	B			Negatively
PTK2B				
CD33				
EPHA1-AS 1				
CR1				
CD2AP				
EPHA1			Occipital	
MS4A6A	D		Occipital	Positively
PICALM				
ABCA7				
SORL1				
SLC24A4				
DSG2				
INPP5D	D			
ZCWPW1				
FERMT2				
CLU				
ADAM10				
KAT8				
ACE				
ECHDC3				

Table 2.7 AD associated risk genes and microglia single cell expression.

AD associated genes cross-referenced against differentially expressed genes between clusters, sex, brain region and age.

2.7 Discussion

In this chapter I describe the collection and sequencing of the largest human primary microglia dataset to date. Dr Adam Young collected brain samples from 141 neurosurgical patients and sorted CD11b+ cells for bulk and single cell RNA-sequencing. From the 141 samples, 109 were included for bulk data analysis and 9,538 cells from 129 patients were analysed from smartseq single cell sequencing. This provides the largest RNA-sequencing resource of fresh primary human microglia to-date with patients in the study coming from a variety of clinical backgrounds. Due to the large scale of the dataset and the range of clinical backgrounds we have been able to run comparisons across pathologies, age ranges, sex and brain regions. The samples also cluster with other smaller datasets of fresh primary cells, despite larger amounts of between sample variability, confirming that our data matches well with high quality published datasets.

From single cell analysis,we have identified limited amounts of heterogeneity in primary microglia and suggest that the majority of the heterogeneity is driven not by distinct subpopulations of cells but of microglial populations that are in differing activation states. 3 of the 4 clusters identified within this dataset had increased expression of immune activation genes, although Cluster B may have represented pre-activated cells. The cells in clusters C and D were enriched for patients from specific pathological backgrounds, most significantly trauma patients. This suggests that the majority of microglia in the brain are in a homeostatic state that is only altered under trauma or disease.

I also demonstrated that selected genes had expression profiles that significantly correlated with age, with an increase in expression of inflammatory genes and a reduced expression of locomotion and motility genes with age. While there were small effects on gene expression linked with age in the primary microglia, there were almost no differentially expressed genes between male and female samples, which is similar to what has been suggested in large scale mouse studies[213]. It may be that in small sub-populations of cells there are more subtle sex or age effects, but as many

of the populations described here are made up of small numbers of cells the ability to detect this subtle differences is reduced.

As microglia have been suggested to be a pathogenic cell type in Alzheimer's disease (AD) and disease specific changes in microglial transcriptomes have previously been reported in AD patients[166,184], I also looked at specific changes in AD linked gene expression within our dataset. While many of the AD linked genes, both those identified in previous single cell studies and GWAS genes, were expressed within this dataset, there was no enrichment for increased gene expression within one specific microglia cluster. This further adds to the theory microglia react in a disease or pathology specific manner. Interestly, reactive microglia have been suggested to be a potential pathogenic cell type that links traumatic brain injury to an increased long-term risk of dementia. In this dataset there was no enrichment for AD linked genes within the trauma patients but this may be because samples were taken within a short time period of the trauma. It may be that as time progresses the cells take on a more AD specific phenotype.

Chapter 3: Comparison of *in-vitro* models of microglia

Collaboration note

Data collected for this chapter comes mainly from publicly available RNA-seq datasets. For details of these data sources please refer to the methods section of the chapter. However, a small number of samples were generated as part of other projects in the Gaffney Lab. The primary microglia are a subset of samples from the data described in Chapter 2, as part of REC 16/LO/2168. A number of the iPSC-derived macrophage samples are from the MacroMap project, involving Dr Andrew Knights, Dr Nikos Panousis and the CGaP core facility at the Wellcome Sanger Institute. Within the cancer cell line samples are a selection of samples generated by Carl Fishwick (GSK) as part of an Open Targets project.

3.1 Introduction

Although primary microglia are a critically important cell there are factors that limit the use of the primary cells in the laboratory. Primary human microglia are inaccessible, particularly as fresh rather than post-mortem samples, and recoverable cell numbers are relatively small. While it is possible to culture primary cells following isolation from the brain, previous data has shown that culturing primary microglia causes a significant change in gene expression and the cells have limited proliferation potential[171].

The limited ability for researchers to use primary cells for *in-vitro* studies, particularly large-scale genetics studies, means that there is a need to develop robust model systems for primary microglia, and to understand how well these models capture the biology of the primary cell. For primary microglia these model systems can range from established macrophage models to more specialised microglia systems. The models discussed in this chapter include: monocyte-derived macrophages (MDMs),

cancer-cell lines (such as THP-1 and U937 lines) and induced pluripotent stem cell (iPSC) models of both macrophages and microglia.

3.1.1 Monocyte-derived macrophages

Both monocyte-derived macrophages (MDMs) and primary microglia are part of the myeloid cell family and are both considered to be macrophages, with microglia representing a tissue-specific arm of the cell group. However, there are fundamental differences in the origin and developmental lineages of the two cell types. Primary microglia have been shown to develop from yolk-sac derived precursor cells that arise in early embryonic development[7,17,232]. Adult monocytes, on the other hand, are constantly replenished by bone-marrow derived cells. How these different lineages impact the cell function remains a controversial topic; particularly as it is known when the blood brain barrier (BBB) is disrupted, circulating monocytes can enter the central nervous system (CNS) and differentiate into brain macrophages[232].

While human MDMs are somewhat easier to derive than primary microglia, sampling primary human cells is still complex and comes with experimental limitations such as an inability to run repeated experiments and a lack system of manipulation. For instance introducing genetic modifications into MDMs can be inefficient and may impact function and expression in nonspecific ways[233,234].

3.1.2 Cancer cell lines

A large proportion of the *in-vitro* studies of macrophage function have been carried out in human myeloid leukemia lines, such as THP-1[235] and U937[236] cells. The patient derived cell lines are thought to represent cells similar to that of monocytes that can be pushed towards more macrophage like phenotypes through simulations with compounds such as phorbol-12-myristate-13-acetate (PMA)[237]. The differentiated cells appear morphologically similar to MDMs and have similar functional capabilities such as phagocytosis as the primary cells[237–239]. However, certain aspects of cancer cell line function have already been shown to differ from MDMs. For instance, THP-1 cell response to lipopolysaccharide (LPS) stimulation significantly differs when

compared to MDMs[240], showing a lack of IL-6 and IL-10 response and a reduction in IL-8 release compared to primary cells.

As the cell lines have been created from single patients, they provide a tool to repeatedly study cell effects on the same genetic background. However, the cells are derived from immortalised cancer cell lines and, therefore, their genetic background may not accurately represent that of healthy individuals. For instance, 119 genetically aberrant regions in the THP-1 genome have been detected[241], including deletions in the *PTEN* gene, a key tumour suppressor gene, and trisomy of chromosome 8.

3.1.3 iPSC derived macrophages

As mentioned in section 1.6, induced pluripotent stem cell (iPSC) based models provide an attractive option for studying human disease[191]. Like in the primary cell type (MDMs), iPSC-derived macrophage cells have been shown to express known myeloid cell marker genes such CD18 and CD68 as well as being functionally similar in their ability to phagocytose compounds[194,195]. Gene expression studies and cytokine profiling have also demonstrated a conserved pro-inflammatory response, such as that following LPS stimulation, in both iPSC and monocyte-derived macrophages[194,195], unlike that seen with cancer-cell lines. However, iPSC differentiated macrophages do not fully match the transcriptional phenotype seen in MDMs. For instance, MDMs have consistently shown an increased expression of the MHC-II cell surface marker[192,193] or genes that encode for the receptor[194,195]. Using differential expression analysis, it has also been noted that iPSC-derived macrophages often express selected genes at a higher level than their monocyte derived counterparts[194,195]. These genes are often enriched for extracellular matrix[194,195], cell adhesion[194] or fibroblast[195] processes.

Interestingly, through CRISPR knock-out of a variety of transcription factors the formation of the myeloid precursors cells generated by EB formation, as used in many of the studies above, has been shown to be *MYB* independent[242]. The formation of these precursors and downstream macrophage-like cell formation appeared to be dependent on the activation of *RUNX1* and *PU.1* and this specific

transcription factor pattern is also seen in yolk-sac myeloid progenitor development. It has, therefore, been suggested that the iPSC-derived macrophage differentiation protocols described above produce cells more closely related to tissue resident cells, such as microglia, as opposed to circulating monocytes[243], especially as the cells have been shown to have significantly increased expression of microglia-linked genes such as *TREM2* and *TMEM119* than monocytes.

3.1.4 iPSC derived microglia

As interest in microglia has increased, a number of research groups have focussed on pushing iPSC derived myeloid models closer to a specialised microglial phenotype as opposed to more generic macrophage-like cells[197–201]. The iPSC-derived microglia cells have consistently shown expression of known microglial genes such as *TMEM119*, *P2RY12*, *PU.1* and *CX3CR1*[197–201] and often have a ramified structure, with highly motile processes which are a unique feature seen in primary microglia.

As with iPSC-derived macrophage studies, many of the differentiation papers described here use transcriptional profiling through RNA-sequencing to determine how closely the in-vitro models match the primary cell type. The iPSC-derived microglia have been shown to have gene expression profiles more similar to fetal/cultured adult primary microglia than dendritic cells, monocytes[198,201], other neuronal cell types[197] and MDMs[199]. However all of these comparisons come with limitations: the number of primary samples studied are often small (< 10) and the comparison is also only run against one iPSC differentiation protocol. The largest published model comparison dataset includes RNA-sequencing data from over 50 primary microglia samples, from three independent studies, and compared it two iPSC-microglia differentiation protocols along with MDMs from one study[200]. In this dataset, iPSC-derived microglia appeared transcriptionally distinct from fresh adult primary microglia but were more similar to cultured microglial cells.

3.1.5 Limitations of current transcriptional comparisons across model systems

Many of the studies described above use transcriptional data to compare *in-vitro* models to primary cell types and in many cases this requires comparison of RNA-sequencing datasets from differing groups. However, comparisons across sequencing studies comes with caveats, particularly batch effects that can arise in these datasets[207–209]. These batch effects can arise from a range of biological and technical factors, particularly when data is processed by entirely different research groups.

The impact of batch effects can vary across studies. Unknown causes of variability can increase noise in samples and, therefore, reduce biological signals[207]. In extreme cases, when the unknown or technical batch effects are confounded with a condition of interest, they may even lead to incorrect biological conclusions. This is something to consider in many of the above studies, whereby often RNA-sequencing data is collected from different studies for differing cell types. It is, therefore, difficult to determine if the effects described are due to the differing cell types or differing experimental studies. However, it is not just technical batch effects that need to be controlled for. Processing pipelines post-sequencing can also significantly impact the quantification of gene expression[209]. Even when the same raw RNA-sequencing reads across the same samples were processed across independent analysis pipelines, abundance estimates of protein coding genes varied by more than four-fold. It is, therefore, key to not only try to reduce experimental and technical batch effects that arise during sample processing but also to ensure all data is processed through identical analysis pipelines.

As well as being aware of the potential batch effects that may have arisen within the studies described in this introduction, it is noted that none of the currently published work compares the transcriptional profile of all available *in-vitro* model systems for primary microglia. In particular, it would be interesting to compare iPSC-derived macrophages to the more specialised microglia differentiation protocols. In an ideal experiment all the samples would be collected from the same research group,

processed in an identical manner and matched for genetic background to try and reduce any batch effects that may arise. However, in a comparison of this scale, and particularly when collecting difficult to access primary cells, often it is not feasible to run these perfectly controlled experiments. In this chapter I have, therefore, collected a mixture of publicly available and in-house generated data across 5 cell types: primary microglia, MDMs, cancer cell lines (THP-1/U937) and iPSC-derived macrophages and microglia. While, in the study there must be comparisons across samples collected from different laboratories, to try and minimise the impact of study batch effects I ensured that data for each cell type came from multiple studies. As mentioned previously, processing pipelines can also impact quantification of gene expression[209] and so in order to counteract some of these potential issues, I collected raw sequencing data for each sample and processed all the data through an identical analysis pipeline. I have used gene expression analysis to understand how each of the model systems compared to primary microglia and gene network analysis to determine which pathways may need to be switched on to move model systems closer to the primary cell type.

3.2 Methods

3.2.1 Data collection and initial processing

Datasets for this study were identified from known large scale transcriptional comparison papers, in house datasets and through pubmed searches for data accession of the desired cell types. Other than in-house data (see collaboration note for the sources of these specific samples), all samples collected as part of this study were from publicly available sources (GEO, ENA, EGA and dbGAP). Table 3.1 summarises the 12 different studies (11 publicly available and in-house data) used within this dataset including accession codes and references for published work attached to the study. It should be noted that access to the samples from the Gosselin *et al.* study[171] are part of a managed access dataset for which use in this project was approved in October 2017.

Study authors	Accession code
Abud *et al.* (2017) [198]	GSE89189
Alasoo *et al*. (2015) [194]	EGAS00001000563
J. de Boer (GEO accession only)	GSE96544
Douvaras *et al*. (2017) [199]	GSE97744
Gosselin *et al.* (2017) [171]	dbGAP : phs001373.v1.p1
In-house	N/A
Gan *et al*. (2017) [244]	GSE97041
Muffat *et al*. (2016) [197]	GSE85839
Phanstiel *et al*. (2017) [245]	GSE96800
Yeung *et al.* (2017) [246]	ERP006216
Zhang *et al.* (2015) [195]	GSE55536
Zhang *et al.* (2016) [247]	GSE73721

Table 3.1 Sources of data collected

Accession codes and paper links to datasets used within this analysis project.

Table 3.2 shows a breakdown how samples from each study are separated by the cell types studied. During collection of these samples, I wanted to ensure that for each cell type I had samples from at least three independent studies. As well as dividing samples by cell type, metadata across the studies was collected. The available metadata varied across the studies and particularly for studies with only cell lines the metadata was limited. However, for all samples data was collected for a mixture of technical (sequencing type, sequencing depth) and experimental (sex, stimulation and culture status) effects. For primary microglia samples, the source of the samples was also identified. Samples collected as part of this dataset originated from 5 distinct sources: fresh adult microglia, fresh paediatric microglia, fetal microglia, cultured microglia and microglia purchased from repositories.

I downloaded raw sequencing files and converted all data into FASTQ file format. All data was then aligned to GRCh38 using the STAR alignment tool[221]. Following alignment, reads were quantified using featureCounts[222]. I used three different

normalisation methods following calculation of raw counts for comparison in this study: calculation of transcripts per million (TPM), variance stabilising transformation (VST) from the DESeq2 package[248] and quantile normalisation as described previously[249].

	Cell Type				
	Primary microglia (pmic)	**Monocyte-derived macrophage (MDM)**	**Cancer cell lines (THP-1/U937)**	**iPSC-derived macrophage**	**iPSC-derived microglia**
Abud[198]	6	-	-	-	9
Alasoo[194]	-	10	-	8	-
J. de Boer (accession only)	-	-	6	-	-
Douvaras[199]	4	8	-	-	10
Gosselin[171]	45	-	-	-	-
In-house	16	-	24	54	
Gan[244]	-	-	4	-	-
Muffat[197]	3	-	-	-	9
Phanstiel[245]	-	-	4	-	-
Yeung[246]	-	-	-	32	
Zhang[195]	-	9	-	18	
Zhang[247]	3	-	-	-	-
Total (studies)	**77 (6)**	**27 (3)**	**38 (4)**	**112 (4)**	**28 (3)**

Table 3.2 Data summary

Table with summary of number of samples for each broad cell type

3.2.2 Principal components and variance components analysis

Following normalisation, I used the prcomp function in R to to compute principal components (PCs) using either all genes in the dataset or across the top 500 most

variable genes. The most highly variable genes were identified using the rowVars function, to calculate variance for each gene row, as carried out in the DESeq2 plotPCA function[248]. Following principal components analysis (PCA), using the varimax function, I rotated calculated PCs to identify the most highly loaded genes for each PC.

As well as identification of individual genes that were driving PCs, I used variance components analysis to identify which metadata may be associated with variability in gene expression. Initially I filtered the dataset to include only protein coding and lincRNA genes that had at least a Log_2(TPM+1) of five across all samples. I used the lmer function of the lme4 package[250] to run a mixed effect linear model for individual genes, with each factor fitted as a random effect:

> lmer (expression ~ (1|study) + (1|cell) + (1|stimulated) + (1|sequence_type) + (1|cultured) + (1|sex))

As described in Chapter 2, I then used the VarCorr function of lmer to estimate the amount of variance attributed to each gene. Following this I calculated the proportion of variance each factor explained by dividing individual factor variance by the total amount of variance for each gene. I did this across all genes analysed as well as across two subsets of genes: microglia marker genes and AD linked genes (for list of genes see Table 3.3).

Microglia marker genes	**Alzheimer's disease genes**		
C1QA	*ABCA7*	*CR1L*	*NME8*
CX3CR1	*ACE*	*DSG2*	*NYAP1*
GAS6	*ADAM10*	*ECHDC3*	*PICALM*
GPR34	*ALPK2*	*EED*	*PILRA*
MERTK	*APH1B*	*EPHA1*	*PLCG2*
P2RY12	*APOC1*	*FBXO46*	*PTK2B*
PROS1	*APOE*	*FERMT2*	*SCIMP*
SALL1	*B4GALT3*	*HESX1*	*SLC24A4*

TMEM119	*BIN1*	*HLA-DQA1*	*SORL1*
	CASS4	*HLA-DRB1*	*TREM2*
	CCDC6	*INPP5D*	*TREML2*
	CD2AP	*KAT8*	*UNC5CL*
	CD33	*MEF2C*	*USP6NL*
	CELF1	*MS4A6A*	*ZCWPW1*
	CLU	*MYBPC3*	*ZNF652*

Table 3.3 Gene lists used in variance components analysis

Microglia marker genes identified from previously published studies[177,178,211,212] and Alzheimer's disease genes collated from Open Targets project OTAR037 (not yet published).

3.2.3 Differential expression and gene set enrichment analysis

I used the DESeq2 package[248] to run differential expression across the dataset. Before differential expression testing the dataset was filtered to only include genes with more than 5 reads in at least 3 samples in the data. The model was set to compare cell types while controlling for study effects where possible. Genes with an adjusted p-value of < 0.05 (with Benjamini & Hochberg multiple testing correction) and a $\log_2$ fold change (LFC) of > 1 were considered differentially expressed.

Gene lists, from differential expression or variance components analysis, were tested for specific gene set enrichment using the g:OSt function of the online gProfiler tool, version e94_eg41_p11_36d5c99[226]. The function uses a hypergeometric distribution model to run over representation analysis on given gene lists, to associate the gene sets with known biological pathways. Gene lists were provided to the tool as an ordered list and significant terms were identified as those with an adjusted p-value of < 0.05 (with Benjamini & Hochberg multiple testing correction).

3.3 Technical comparisons within the dataset

3.3.1 Normalisation comparison

It has been demonstrated that different processing pipelines can lead to significant differences in gene abundance estimates[209]. While a full comparison of how differing initial analysis pipelines (alignment and quantification) has not been carried out as part of this study, I was interested to look at how differing normalisation techniques could impact downstream results. I compared transcripts per million (Log_2(TPM+1)), quantile normalisation (QN) and the variance stabilising transformation (VST) described as part of the DESeq2 package[248].

Following normalisation of the data using each of these methods, I ranked genes by variance across all samples and compared the top 500 most variable genes for each normalised dataset. Figure 3.1 shows a venn diagram of the numbers of overlapping genes for each normalisation method. Only 236 of the top 500 genes for each normalisation method were shared between all three techniques, with QN normalisation having the most unique genes (165). Log_2(TPM+1) and VST normalizations had the greatest overlap across highly variable genes with 364 shared genes. This highlights that, even when initial alignment and quantification is identical across samples, differing normalization methods can still impact certain downstream analysis outcomes.

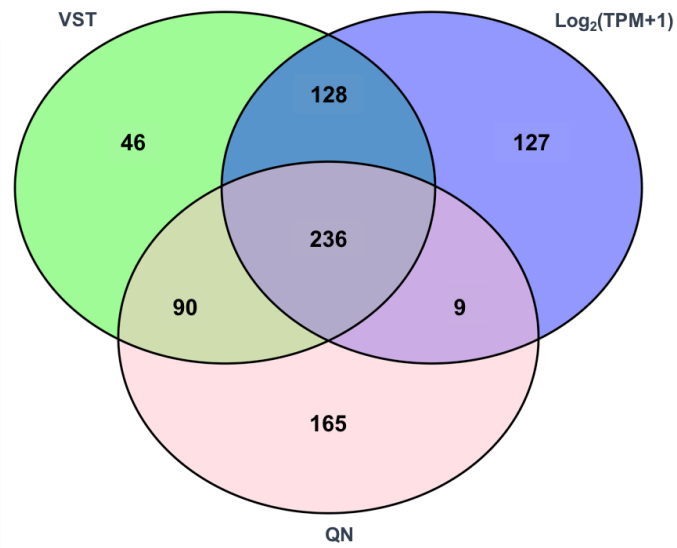


Figure 3.1 Venn diagram of overlapping most variable genes

Top 500 most variable genes were calculated following three independent normalisation methods: variance stabilising transformation (VST), quantile normalisation (QN) and transcript per million (Log_2(TPM+1)).

As well as identifying specific differences in the most variable genes across normalisation methods, I also wanted to understand how these differences may impact downstream PCA and the biological conclusions that could be drawn from it. I took the top 500 genes calculated above for each normalisation and used those genes to run PCA. I plotted samples (Figure 3.2) based on their PC scores for the first two principal components and coloured samples by cell type to compare the pattern of sample distribution across the normalisation methods.

Broadly the patterns of sample clustering were the same across all three normalisation methods. PC1 captured the variation in iPSC based models (both macrophages and microglia). Across all three normalisation methods PC2 captured a similar spread of cell types with the cancer cell models at one end, MDM/iPSC macrophages/iPSC microglia in the middle band and a group of primary microglia at the opposite end. This suggests that even though the specific genes driving the PCs may differ slightly between normalisation methods, the biological conclusions that can be drawn from initial PCA was similar.

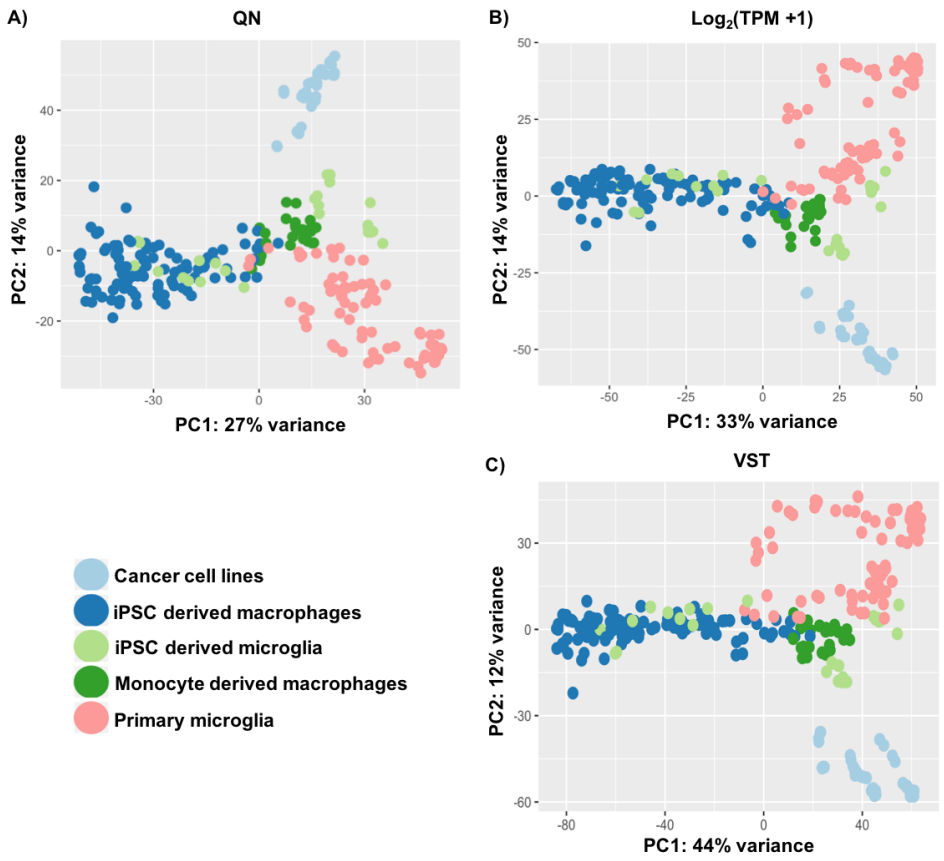


Figure 3.2 PC1 vs PC2 for three normalisation methods

Principal component analysis of RNA-sequencing samples, using the top 500 most variable genes following 3 normalisation methods: A) quantile normalisation (QN), B) transcripts per million (log_2(TPM+1)) and C) variance stabilising normalisation (VST).

3.3.2 Variance components analysis

In order to futher understand which biological and technical factors may be driving variation within the dataset, I used variance components analysis to calculate the proportion of variation explained across individual genes for six factors: study, cell type, cultured/non-cultured cells, naive/stimulated cells, single/paired end sequencing and sex. I used Log_2(TPM+1) normalised data to calculate this proportion first across all genes, as well as specifically in AD genes and microglia marker genes. Figure 3.3 highlights the spread of the proportion of variance for each of the factors subdivided by the gene groups. When looking at variation across all genes, study explained the largest proportion of variation. However, when looking at only microglia marker genes cell type and the culturing status of cells became more important. Sex and stimulation status had little effect on variation within all three gene groups and, while

on average sequence type only explained a very small proportion of variability, the variability across all genes was relatively high with over 50% of variability explained by sequence type in a small number of genes.

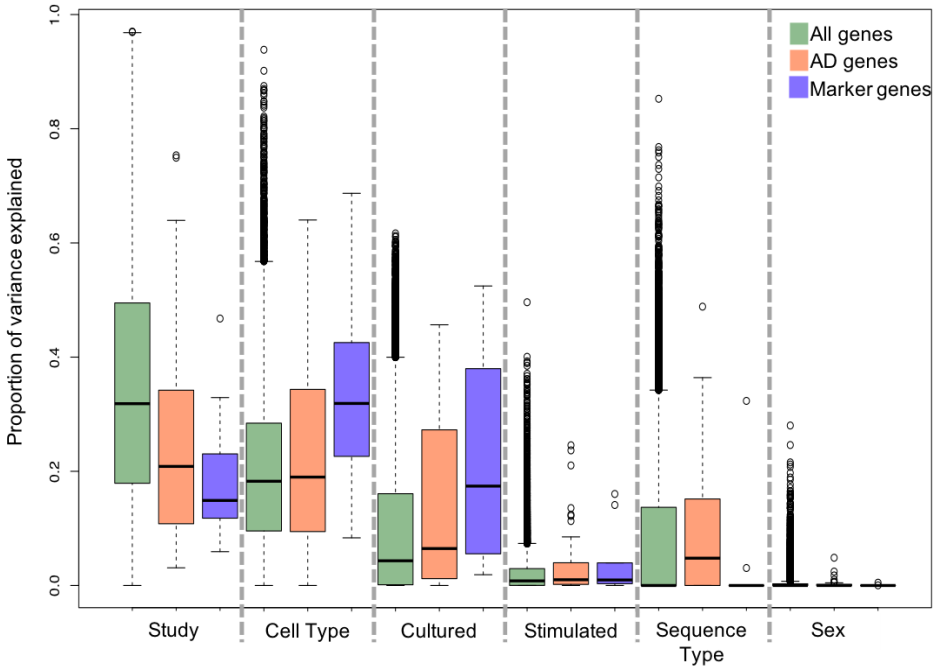


Figure 3.3 Variance components analysis

Proportion of variance explained by metadata groups - across all genes (green), Alzheimer's disease (AD) linked genes (orange) and microglia marker genes (purple).

3.3.3 Effects of differing gene set inputs on principal components analysis

The variance components analysis described above showed that across all genes in this dataset study explains on average the largest proportion of variation in gene expression, however this changed as the genes were subsetted. I wanted to understand if changing the number of genes included in PCA would impact the outcome and interpretation of the analysis. I used all genes and the 500 top most variable genes, as suggested in the standard DESeq2 pipeline, to run PCA and compared sample distribution across PC1 and PC2 (Figure 3.4). When looking at grouping of different cell types across the first two PCs, both gene inputs appeared to capture some similar biological patterns, with PC2 appearing to separate the cancer cell models from the other cell types included here. However, when all genes were used as an input (Figure 3.4 A), PC1 appears to capture variability in primary

microglia. The same PC when using the top 500 most variable genes (Figure 3.4 B), appears to capture variability in the iPSC based systems. Colouring samples by study shows that there may be less integration of different studies when all genes are used (Figure 3.4 C) compared to the top 500 (Figure 3.4 D). Although this is only true outside of the cancer cell line samples, where in both gene inputs, the cell type differences appear to be a larger driver of variation than study to study effects. Based on these results, in all downstream analysis of computed principal components using top 500 most variable genes (Figure 3.4 B) in order to minimise any study based effects.

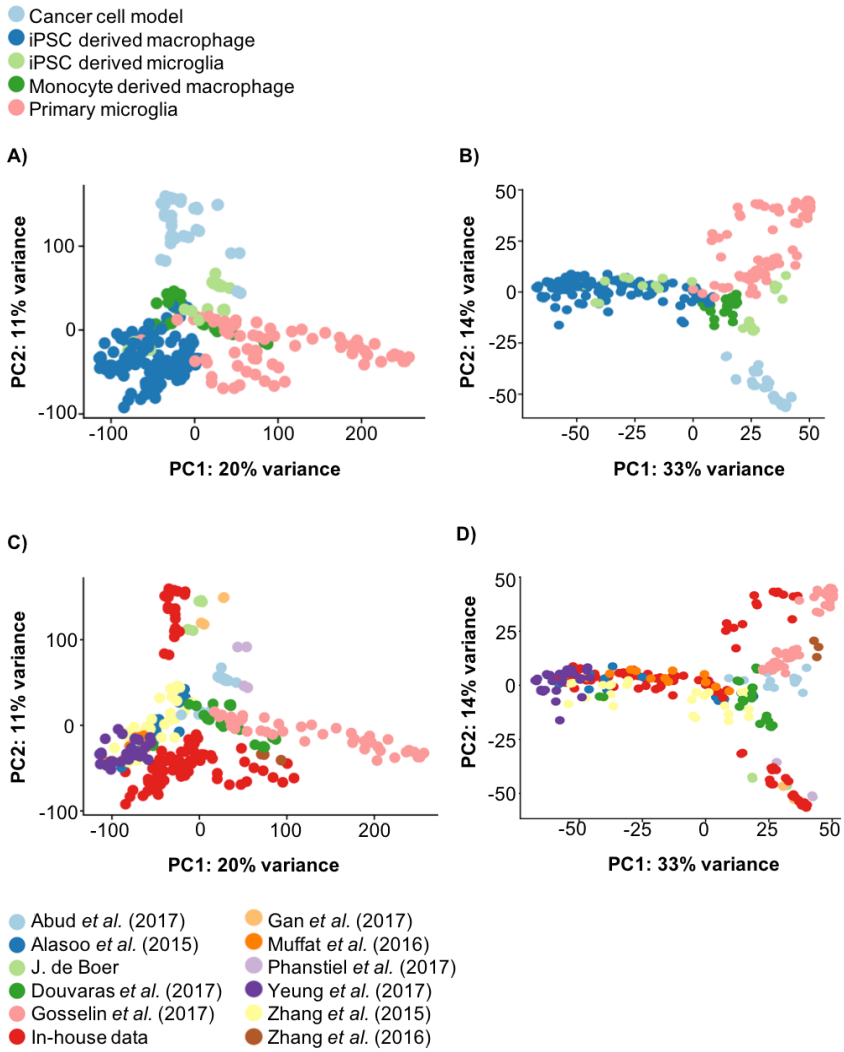


Figure 3.4 PC1 vs PC2 for all genes and top 500 genes

Samples plotted following calculation of principal components with: all genes (A and C) and the 500 most variable genes (B and D). All samples are coloured by cell type (A and B) or study (C and D).

3.4 Utilising principal component analysis to identify sources of variation

3.4.1 Defining principal components

Following the assessment of how technical factors could influence PCA described above, I then wanted to understand whether PCA could be used to understand drivers of variation within this dataset. First I focused on the spread of samples across PC1 and PC2 as shown in Figure 3.5. The largest amount of variation in the top 500 most variable genes (33%) appeared to capture variation within the iPSC derived macrophages and microglia, while PC2 (14% of variation) appeared to separate samples by cell type (Figure 3.5 A). The cancer cell models had the lowest PC2 scores, with a band of MDMs and iPSC-derived cells falling in the middle range of scores and the primary microglia with the highest PC2 scores. The primary microglia separated into two almost distinct groups, with some samples sitting much closer to the iPSC model/MDM band in the central part of the PC. In order to understand what might have been driving this variation along PC2, particularly amongst the primary microglia samples, I looked at the culture status of each sample (Figure 3.5 B). This showed that samples that had been cultured had lower PC2 score than the fresh primary microglia and suggested that cultured primary microglia cells looked more like iPSC-derived samples. It is also worth noting that fetal microglia (Figure 3.5 C), even when sequenced without culturing, also had PC2 scores more similar to that of iPSC-derived cells.

Next I tried to characterise the variation in expression captured by additional PCs. Figure 3.6 shows samples projected on PC3 vs PC4 coloured by available metadata groups. PC3 was associated with stimulation status ($p = 5.11e^{-14}$ following Welch Two Sample t-test between PC3 score and stimulation status), while the factors driving PC4 remained unclear.

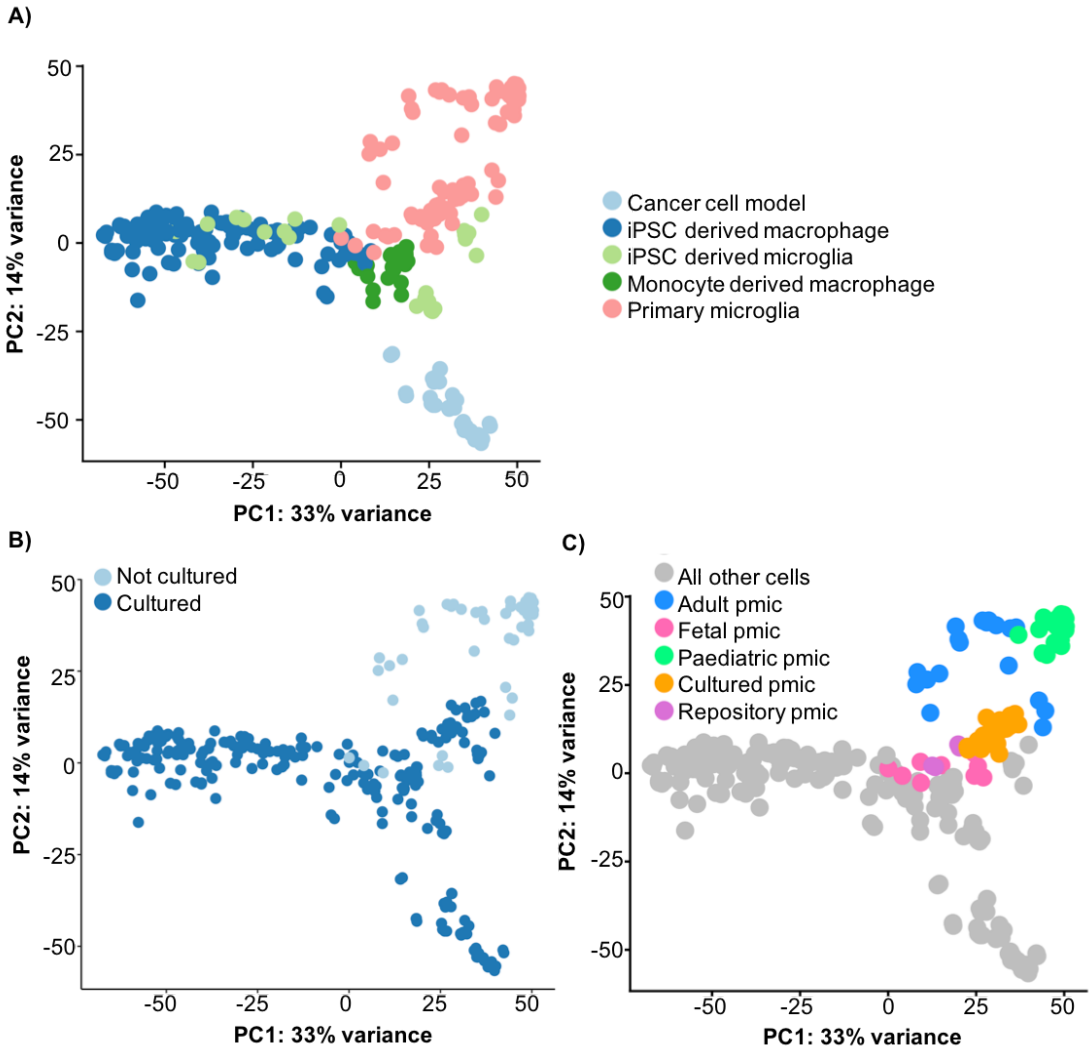


Figure 3.5 PC1 vs PC2 calculated using the top 500 genes

Samples plotted following calculation of principal components with top 500 most variable genes. Coloured by cell source (left panel) and cultured status (right panel).

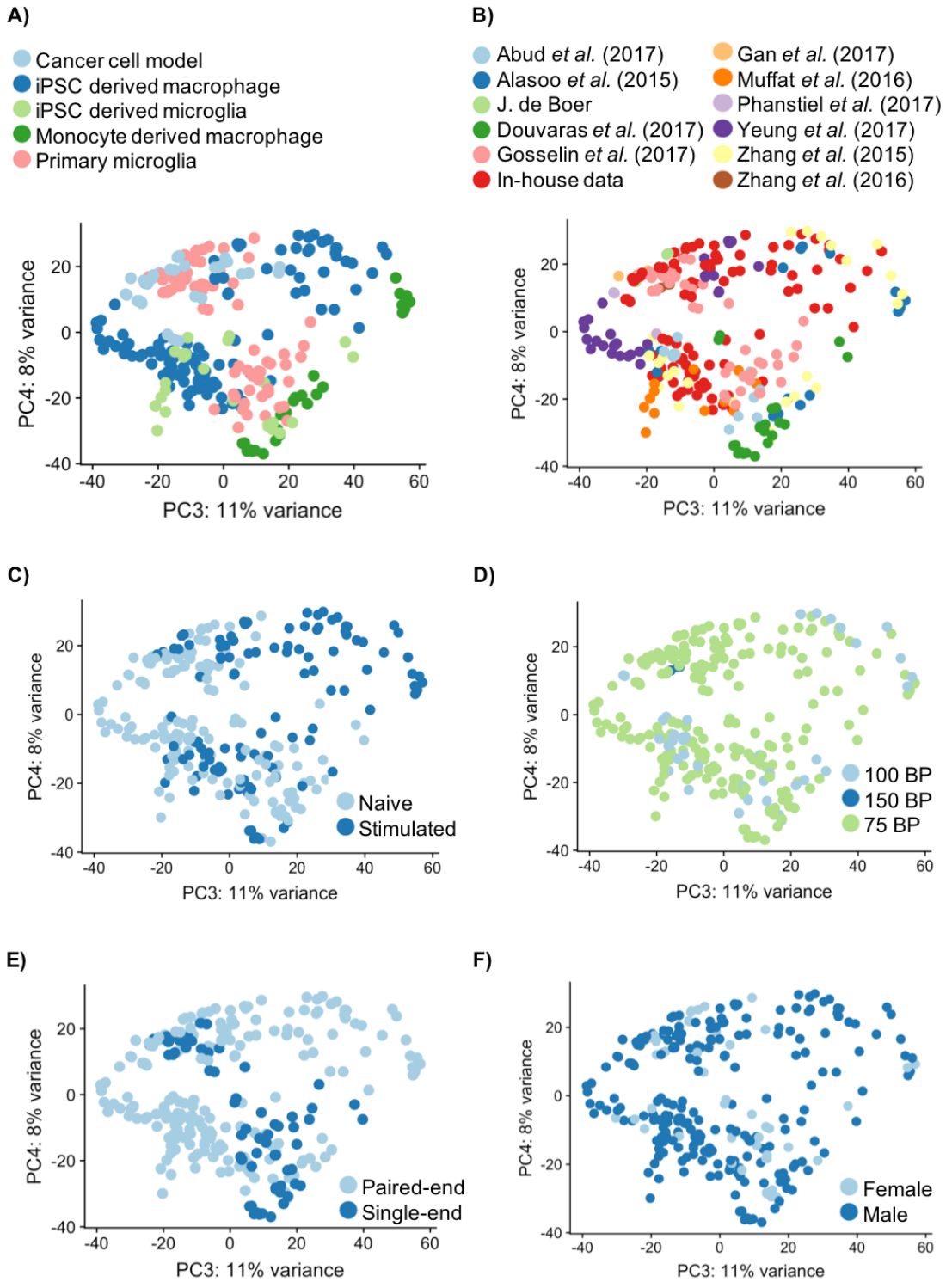


Figure 3.6 PC3 vs PC4 calculated using the top 500 genes

Samples plotted following calculation of principal components with top 500 most variable genes. Coloured by: A) cell type, B) study, C) stimulation, D) sequencing read length, E) sequence type and F) sex.

3.4.2 Varimax analysis of principal components

While PCA provides a tool for understanding drivers of variation with a gene expression dataset, as shown above this often relies on associating principal components with known metadata which is not always possible. Therefore, techniques have been developed to increase the interpretability of PCA. Varimax is an orthogonal rotation technique that allows the identification of specific variables that heavily load principle components. In the case of gene expression data, it links the expression of specific genes with each PC. I, therefore, used the varimax function in R to rotate the first 5 PCs in order to further understand what may have been driving the major sources of variation within the dataset. Table 3.4 highlights the most heavily loaded genes for each component. The genes most negatively loaded on PC1 included collagen genes as well as genes linked to the extracellular matrix and cell adhesion. Previous work comparing iPSC-derived macrophages to MDMs, showed that similar gene sets were more highly expressed in the iPSC-derived cells[194]. It may be that the variability in expression of these genes across the iPSC based model systems, represents variation in the completeness of differentiation as many of the genes are also highly expressed in undifferentiated cells.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Top 5 loaded genes (-ve)	*COL3A1*	*CCL13*	*GPR34*	*RNASE1*	*RN7SL2*
	COL1A1	*MMP9*	*ADORA3*	*C1QC*	*CHIT1*
	IGFBP5	*ANXA2*	*PALD1*	*STAB1*	*RN7SL3*
	POSTN	*S100A4*	*DDIT4L*	*C1QB*	*HIST1H1E*
	CTGF	*CD36*	*PDK4*	*C1QA*	*SCARNA7*
Top 5 loaded genes (+ve)	*CAT*	*FOSB*	*CXCL10*	*ELANE*	*RNASE2*
	MMP9	*CH25H*	*IDO1*	*CTSG*	*CD93*
	SPN	*P2RY12*	*ACOD1*	*AZU1*	*MT-TN*
	CHI3L1	*CX3CR1*	*TNFAIP6*	*PRTN3*	*MT-ATP8*
	CSTA	*EGR3*	*CCL8*	*CES1*	*MT-TL1*

Table 3.4 Top 5 loaded genes for each principal component

Varimax analysis of the first 5 principal components from the top 500 most variable genes. Top 5 most negatively and positively loaded genes for each component.

When looking at the genes that were driving PC2, those most positively loaded included many known microglia marker genes such as *P2RY12* and *CX3CR1* as well as transcription factors such as *SALL1*. Figure 3.9 highlights expression (log_2(TPM+1)) of *P2RY12* and *SALL1* across the first two PCs.

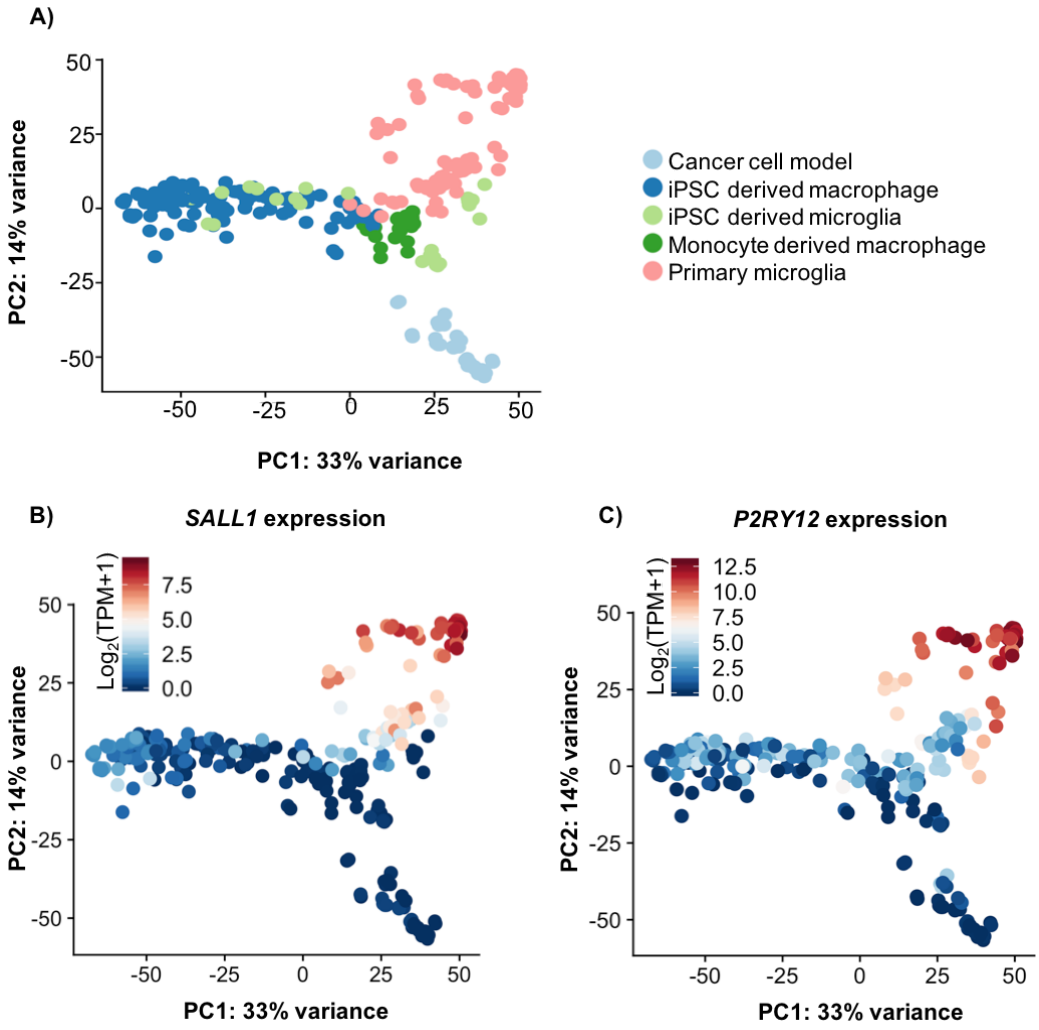


Figure 3.7 PC1 vs PC2 coloured by expression of genes heavily loading PC2

Samples plotted following calculation of principal components using the top 500 most variable genes. Samples coloured by: A) cell type and B) & C) expression (Log_2(TPM+1)) of microglia marker genes SALL1 and P2RY12 respectively.

Genes most negatively loading on the third PC were linked to inflammatory pathways in immune cells (such as CXCL10 and CCL8). This further supports the hypothesis that PC3 may capture stimulation effects. The genes most negatively loading on PC4 included many of the C1Q complex and gene set enrichment analysis highlighted

terms such as defense response (GO:0006952). The genes most positively loaded on PC4 included immune activation linked genes. Genes that were found to drive PC5 included mitochondrial genes and apoptosis-linked genes such as *CD93*. This suggested that PC5 may have been capturing sample quality. As much of the data collected for this analysis was from publicly available sources it is difficult to obtain information regarding the quality of the cells that are used in the analysis prior to sequencing (i.e. ratio of live/dead cells prior to sequencing, RIN value of RNA) and therefore accurately determining what may have been driving PC5 was difficult.

3.5 Differential expression between cell types

3.5.1 Primary microglia vs all models

Initially I used differential expression (DE) analysis, using the DESeq2 package, to compare primary microglia to all the *in-vitro* model systems in order to understand which regulatory mechanisms and programmes were not well captured by all existing models. Figure 3.8 shows the MA plot following DE analysis comparing primary microglia to all other model systems. I used this analysis to curate a list of 7297 genes which had a significantly (p_{adj} < 0.05 and a LFC > 1) higher expression in primary microglia than any of the *in-vitro* model systems. I shall refer to this gene set as the primary microglia marker (PMM) gene set throughout the remainder of this thesis. The PMM gene set included many known microglia marker genes including: *P2RY12* (p_{adj} = 5.73e^{-41} and LFC = 7.4), *CX3CR1* (p_{adj} = 4.23e-^{27} and LFC = 6.4) and *TMEM119* (p_{adj} = 9.05e^{-80} and LFC = 7.0). As well as including microglial cell surface markers, the list of genes also included transcription factors such as *SALL1* that may need to be switched on in order for model systems to move closer to the primary phenotype.

As well as identifying individual genes of interest in the PMM gene set, I also ran gene set enrichment analysis (GSEA) on the PPM genes to identify molecular pathways that were not switched on in the model systems. Table 3.5 highlights the top 10 enriched terms within the PMM gene set. Many of the enriched terms were

linked to neuronal signalling, including nervous system development and synaptic signalling. This suggests that many of the signalling processes missing from the *in-vitro* model systems studied here are related to the CNS microenvironment that microglia are normally found in.

There were also 2686 genes with a significantly (p_{adj} < 0.05 and a LFC > 1) higher expression in the *in-vitro* model systems compared to primary microglia (Figure 3.8), including genes such as *POSTN* and *TTR*. GSEA of the genes highlighted an enrichment for extracellular matrix terms like extracellular matrix organization (GO:0030198, $p_{adj} = 3.5e^{-27}$) and extracellular structure organization (GO:0043062, $p_{adj} = 2.52e^{-25}$).

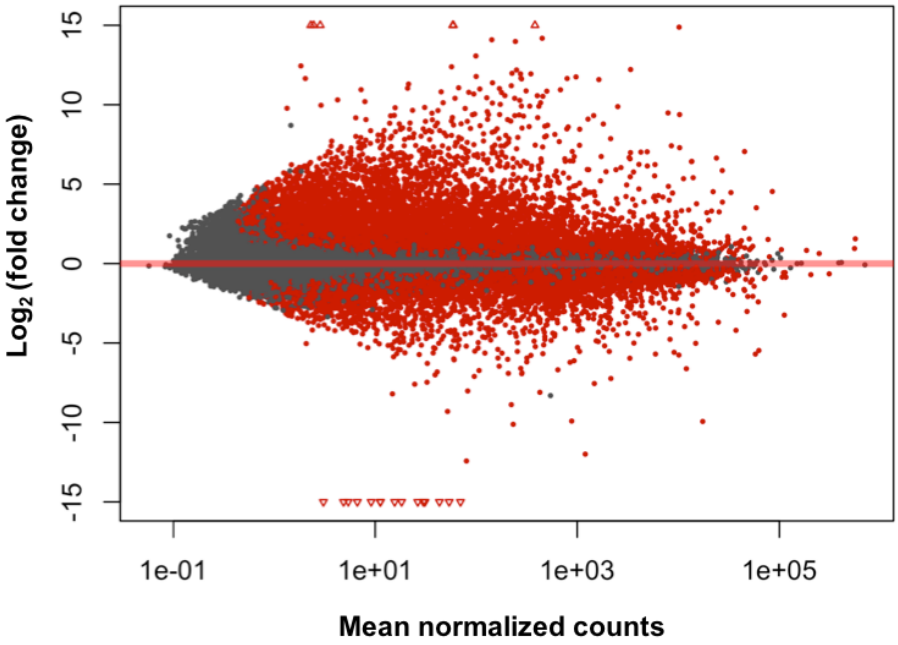


Figure 3.8 MA plot following differential expression analysis comparing primary microglia to all other cell types

Average normalised counts of individual genes plotted against Log_2(fold change) in expression when comparing primary microglia to all other cell types. Points coloured in red represent genes reaching a p_{adj} threshold of < 0.05 and triangular points are genes were the Log_2(fold change) falls outside the limits of the graph.

Term name	Term ID	P_{adj}
nervous system development	GO:0007399	$8.18e^{-29}$
ion transport	GO:0006811	$8.80e^{-28}$
trans-synaptic signaling	GO:0099537	$2.89e^{-26}$
cell adhesion	GO:0007155	$7.66e^{-26}$
anterograde trans-synaptic signaling	GO:0098916	$7.66e^{-26}$
chemical synaptic transmission	GO:0007268	$7.66e^{-26}$
biological adhesion	GO:0022610	$8.76e^{-26}$
synaptic signaling	GO:0099536	$4.02e^{-25}$
cell development	GO:0048468	$1.57e^{-24}$
cation transport	GO:0006812	$2.04e^{-23}$

Table 3.5 Top enriched biological process terms in the PMM gene set

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Ten most significantly enriched biological process terms for genes with higher expression in primary microglia compared to all model systems.

3.5.2 Primary microglia vs individual model systems

PCA analysis of the dataset (section 3.4.1) identified cell type as a potential factor driving PC2 with iPSC derived cells sitting as an intermediate along the PC between primary microglia and cancer models. This suggested that iPSC-derived cells may represent a closer cell type to primary microglia than cancer cell models. To confirm this theory, I ran DE comparing primary microglia to cancer cell models and iPSC-derived cells individually (Figure 3.9). There were more genes with significantly higher expression ($p_{adj} < 0.05$ and a LFC > 1) when primary microglia were compared to cancer cell models than when compared to iPSC-derived cells (13996 and 6963 respectively). As well as having more DE genes in total, the average Log_2(fold change) across the primary/cancer cell model comparison was also higher than the primary/iPSC-derived comparison (3.9 and 2.7 respectively).

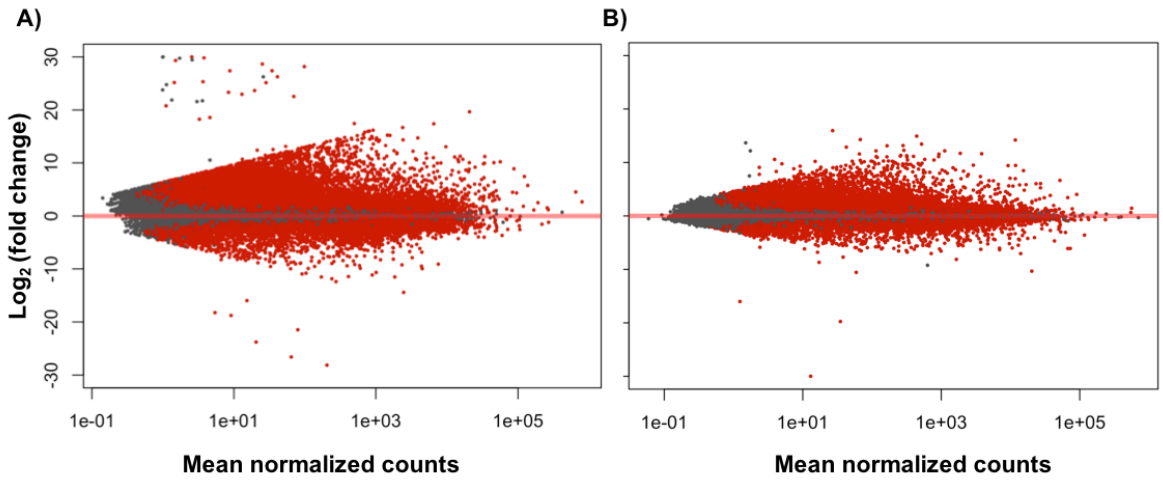


Figure 3.9 MA plots comparing primary microglia to cancer cell lines and iPSC-derived cells

Average normalised counts of individual genes plotted against Log_2(fold change) in expression when comparing primary microglia to cancer cell models (A) or iPSC-derived cells (B). Points coloured in red represent genes reaching a p_{adj} threshold of < 0.05 (FDR).

I also ran GSEA on both gene lists and table 3.6 highlights the top enriched terms on genes more highly expressed in primary microglia when compared to cancer cell models and iPSC-derived cells individually. While each gene list identified unique terms, such as cell adhesion and ion transport, neuronally linked terms were also present in both GSEA.

Top GO:BP terms for primary microglia vs cancer cell models			**Top GO:BP terms for primary microglia vs iPSC-derived cells**		
Term name	**Term ID**	**P_{adj}**	**Term name**	**Term ID**	**P_{adj}**
cell adhesion	GO:0007155	$1.17e^{-41}$	nervous system development	GO:0007399	$6.03e^{-36}$
biological adhesion	GO:0022610	$1.17e^{-41}$	trans-synaptic signaling	GO:0099537	$2.74e^{-28}$
cell communication	GO:0007154	$1.50e^{-29}$	neurogenesis	GO:0022008	$2.74e^{-28}$
signaling	GO:0023052	$2.92e^{-29}$	ion transport	GO:0006811	$5.21e^{-28}$
regulation of multicellular organismal process	GO:0051239	$3.34e^{-29}$	chemical synaptic transmission	GO:0007268	$5.21e^{-28}$

system development	GO:0048731	$4.76e^{-28}$	anterograde trans-synaptic signaling	GO:0098916	$5.21e^{-28}$
nervous system development	GO:0007399	$4.76e^{-28}$	synaptic signaling	GO:0099536	$5.89e^{-28}$
anatomical structure development	GO:0048856	$3.40e^{-26}$	generation of neurons	GO:0048699	$3.21e^{-26}$
regulation of signaling	GO:0023051	$2.53e^{-25}$	cell development	GO:0048468	$7.51e^{-26}$
multicellular organismal process	GO:0032501	$9.23e^{-25}$	multicellular organismal process	GO:0032501	$2.62e^{-25}$

Table 3.6 Significantly enriched biological process terms for genes with significantly higher expression in primary microglia compared to individual model systems.

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Ten most significantly enriched biological process terms for genes with higher expression in primary microglia compared to cancer cell models and iPSC-derived cells individually.

The output of these individual DE analyses suggested that, when looking at gene expression, iPSC-derived cells were transcriptionally more similar to primary microglia than cancer cell models but both systems still lacked the CNS microenvironment stimulus identified by GSEA on the PMM gene set.

3.5.3 iPSC macrophages vs iPSC microglia

Within the iPSC-derived data collected for this study, some of the protocols were developed to push myeloid progenitor cells towards macrophages whereas others were more specifically developed to move the progenitor cells closer towards primary microglia. Next I compared iPSC-derived macrophages and iPSC-derived microglia to understand whether more complex microglia differentiation protocols produce markedly different cells to standard macrophage differentiation protocols. It should be noted that for this differential expression analysis, study could not be fitted in the differential expression model (unlike all previous analysis), because, for this comparison, study was confounded with cell type.

I found 4975 genes with significantly higher expression in iPSC-derived microglia and 5461 genes that had higher expression in iPSC-derived macrophages (p_{adj} <0.05 and LFC > 1). Genes with significantly increased expression in iPSC-derived microglia were enriched for ion transport terms whereas those with significantly increased expression in iPSC-derived macrophages were enriched for developmental terms (Table 3.7). As I wanted to understand whether specific microglia differentiation protocols pushed the cell model systems closer to the primary cell type, I compared the list of genes more highly expressed in iPSC microglia to the PMM gene set described in section 3.5.1. There were 2,164 genes that overlapped between the two lists, approximately 30% of the total genes in the PMM gene set. This suggested that there were some PMM genes that were also enriched in iPSC-derived microglia compared to their macrophage counterparts, potentially highlighting a shift closer to the primary phenotype. These genes included some known microglia marker genes such as *P2RY12* and *CX3CR1*.

Top GO:BP terms for genes with increased expression in iPSC-derived macrophages			**Top GO:BP terms for genes with increased expression in iPSC-derived microglia**		
Term name	**Term ID**	**P_{adj}**	**Term name**	**Term ID**	**P_{adj}**
system development	GO:0048731	$7.76e^{-57}$	ion transport	GO:0006811	$1.32e^{-18}$
multicellular organism development	GO:0007275	1.43^{e-52}	cation transport	GO:0006812	$1.50e^{-16}$
anatomical structure development	GO:0048856	$3.86e^{-52}$	transmembrane transport	GO:0055085	$4.51e^{-15}$
anatomical structure morphogenesis	GO:0009653	$2.00e^{-50}$	regulation of ion transport	GO:0043269	$2.87e^{-14}$
developmental process	GO:0032502	$1.63e^{-48}$	ion transmembrane transport	GO:0034220	$3.80e^{-14}$
multicellular organismal process	GO:0032501	$6.86e^{-43}$	cation transmembrane transport	GO:0098655	$6.01e^{-14}$
cell adhesion	GO:0007155	$1.59e^{-39}$	metal ion transport	GO:0030001	$3.56e^{-13}$
biological adhesion	GO:0022610	$1.65e^{-39}$	inorganic ion transmembrane transport	GO:0098660	$1.33e^{-11}$
animal organ development	GO:0048513	$7.37e^{-38}$	regulation of biological quality	GO:0065008	$1.71e^{-11}$

regulation of multicellular organismal process	GO:0051239	$1.97e^{-35}$	chemical homeostasis	GO:0048878	$5.13e^{-11}$

Table 3.7 Significantly enriched biological process terms for genes with significantly higher expression in iPSC-derived macrophages or microglia when compared to each other

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Ten most significantly enriched biological process terms for genes with higher expression in primary microglia compared to cancer cell models and iPSC-derived cells individually.

3.6 Expression of Alzheimer's disease genes across model systems

One common use of the scalable *in-vitro* cell model systems is to study the mechanism of action of individual genes and how perturbation of gene expression may impact cell function. This is particularly useful when trying to understand how disease risk linked genes identified by genome wide association studies (GWAS) may impact cell function in disease. As microglia have been suggested to be a pathological cell type in Alzheimer's disease (AD)[1,31], I examined the level of conservation of expression of known or suspected AD risk genes between primary microglia and the different cellular model systems.

3.6.1 Expression of known Alzheimer's disease genes

I first looked at the expression of three genes associated with familial AD: *APP*, *PSEN1* and *PSEN2*. Figure 3.10 shows expression (DESeq2 normalised) of each of the three genes for each sample. Expression of each of the three genes was not significantly increased in primary microglia compared to *in-vitro* cell models.

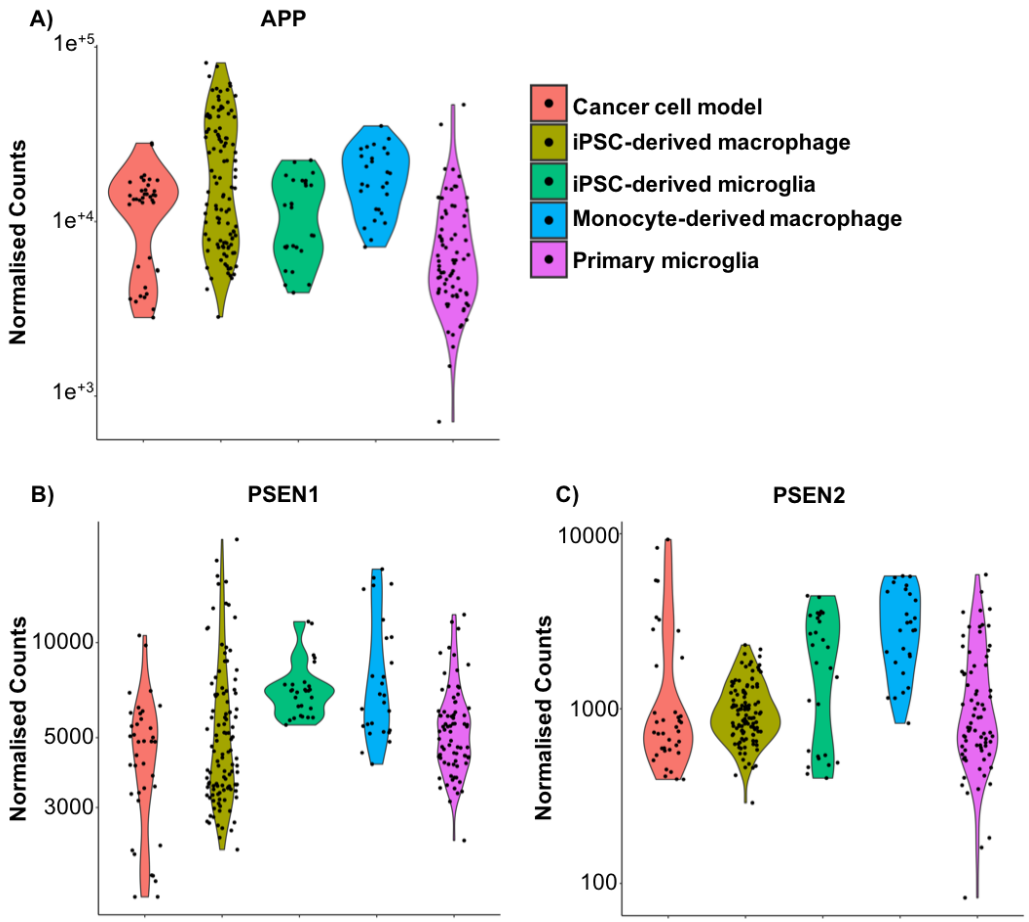


Figure 3.10 Expression of familial AD genes by cell type

DESeq2 normalised expression data of familial AD disease genes, samples separated by broad cell type.

Next I examined the expression of genes associated with late-onset AD. The strongest signal of gene association with AD risk is the APOE region, with APOEε4 associated with the largest risk increase[123]. *APOE* was significantly more highly expressed in primary microglia when compared to all other model systems (p_{adj} = 1.41e^{-10}, LFC = 2.24) Figure 3.11 A, and particularly comparing primary microglia to cancer cell lines (p_{adj} = 1.96e^{-15}, LFC = 4.42). *APOE* was also significantly (p_{adj} = 3.03e^{-10}, LFC = 2.1) more highly expressed in iPSC-derived microglia than in iPSC-derived macrophages, suggesting that, for studying *APOE* function, microglia rather than macrophage differentiation protocols may be preferable.

Rare missense variants in *TREM2*[251,252], *ABI3* and *PLCG2*[130] have all been associated with increased AD risk, and have suggested immune functions . There

was no significant difference in expression of *PLCG2* (Figure 3.14 B) across any of the cell types. Expression of *TREM2* and *ABI3* (Figure 3.14 C and D respectively) were significantly reduced in cancer cell lines compared to primary microglia (p_{adj} = $2.7e^{-8}$, LFC = 3.1 and p_{adj} = $2.87e^{-128}$, LFC = 7 respectively). However, expression in iPSC-derived cells was not significantly different to that seen in primary microglia and, therefore, iPSC based systems could be used as *in-vitro* models for studying the effect of these genes.

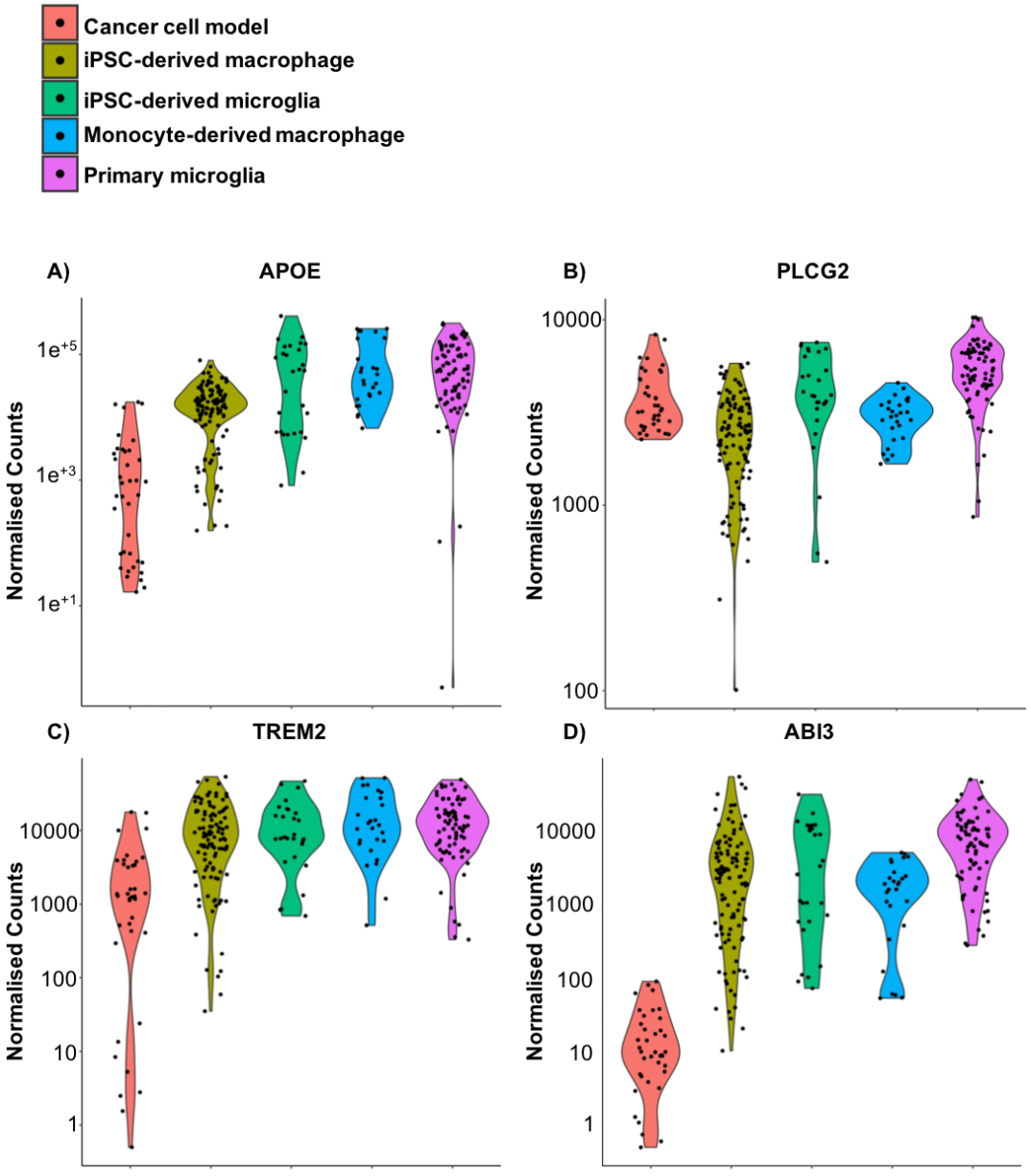


Figure 3.11 Expression of late onset AD rare and high effect size genes by cell type

DESeq2 normalised expression data of late onset AD disease genes, samples separated by broad cell type.

3.6.2 Expression of late onset Alzheimer's disease linked genes

As described in section 2.6.2 the study described in Chapter 2 of this thesis was part of a large collaborative project that also included an expression quantitative trait loci (eQTL) map of adult primary human microglia(Young *et al.* - paper in preparation). The identified eQTLs were then co-localised with variants identified from AD genome wide association studies (GWAS) to identify candidate causal AD risk genes and variants.

One of the strongest signals of colocalisation we identified was found at the *BIN1* locus that appeared to be driven by the rs6733839 SNP which in turn perturbed a binding site for the transcription factor MEF2A. *BIN1* had significantly increased expression in primary microglia when compared to all model systems (p_{adj} = 8.03e^{-33} and LFC = 3.18), (Figure 3.12 A). While the expression of *MEF2A* (Figure 3.12 B) was not significantly different when primary microglia were compared to the model systems collectively, expression of the gene was significantly reduced when primary microglia were compared to cancer cell models individually (p_{adj} = 2.09e^{-13} and LFC = 2.14).

As well as developing our understanding of the *BIN1* risk loci, the eQTL/GWAS co-localisation also identified other potential SNP-gene links at AD risk loci including: *PTK2B*, *CASS4*, *CD33* and *EPHA1-AS1* (Figure 3.12 C-F). There was no significant difference in expression of *CD33*, *PTK2B* or *EPHA1-AS1* when comparing primary microglia and the model systems but expression of *CASS4* was significantly increased in primary cells compared to all other model systems(p_{adj} = 3.57e^{-14} and LFC = 2.61).

Table 3.8 summarises the DE between primary microglia and cancer cell models or iPSC-derived cells for all of the genes described in this section (3.6) as well as other genes that have been identified as the "nearest gene" to an AD risk variant in more than one GWAS study (see Table 1.1 for full list and matching subset in Table 2.11). Of the 30 AD genes identified, 70 % had a statistically similar expression in at least one model system compared to primary microglia. However, for 9 individual AD

genes neither cancer cell models or iPSC-derived cells accurately captured the expression profile of primary microglia (p_{adj} < 0.05 and LFC > 1).

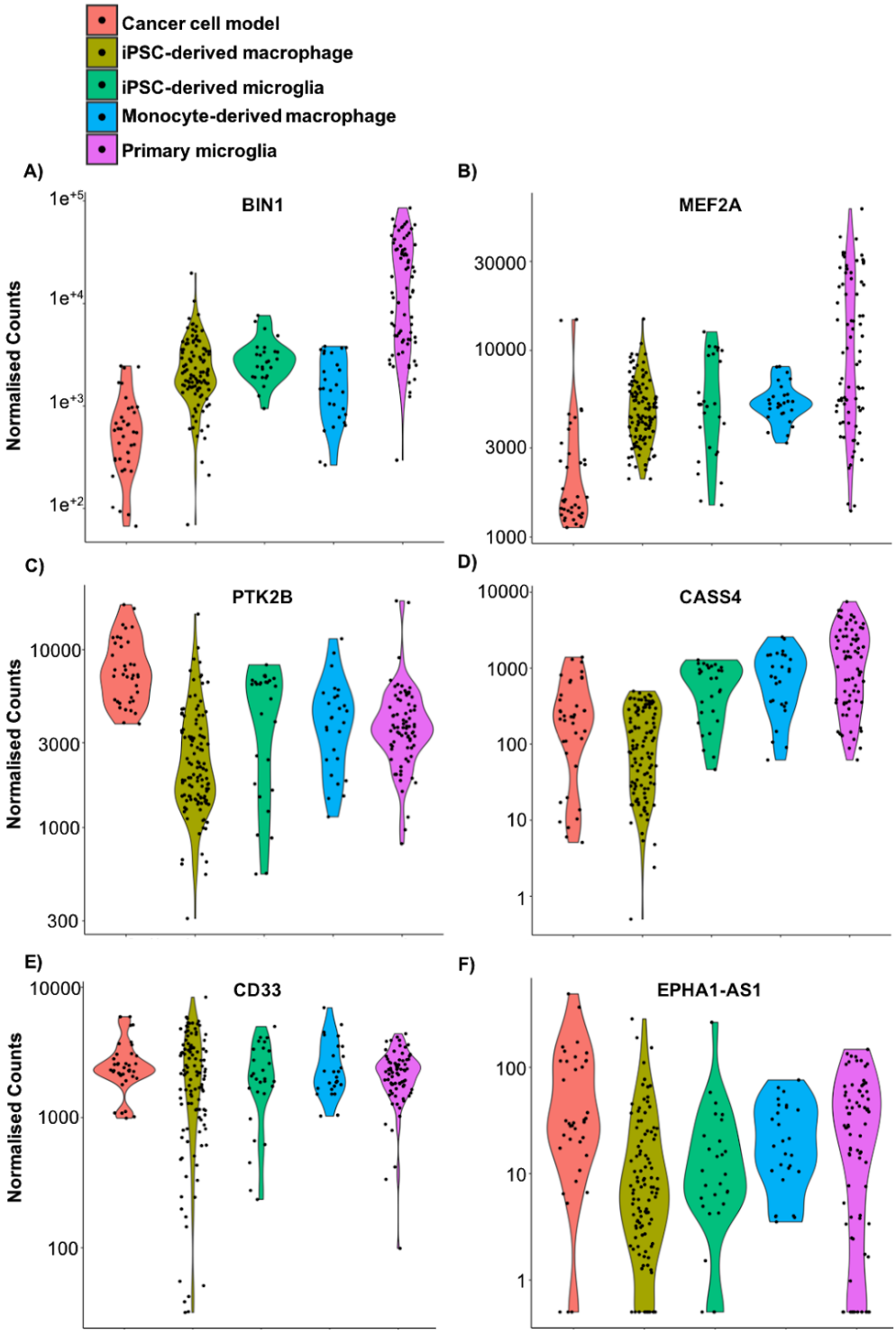


Figure 3.12 Expression of late onset AD risk genes

DESeq2 normalised expression data of late onset AD disease genes, samples separated by broad cell type

Gene name	Is expression statistically similar in primary microglia and	
	cancer cell models?	iPSC-derived cells?
APP	Yes	Yes
PSEN1	Yes	Yes
PSEN2	Yes	Yes
*APOE**	No	No
TREM2	No	Yes
PLCG2	Yes	Yes
ABI3	No	Yes
*BIN1**	No	No
MEF2A	No	Yes
*CASS4**	No	No
PTK2B	Yes	Yes
CD33	Yes	Yes
EPHA1-AS1	Yes	Yes
*CR1**	No	No
CD2AP	Yes	Yes
EPHA1	Yes	Yes
MS4A6A	No	Yes
PICALM	No	Yes
ABCA7	Yes	Yes
*SORL1**	No	No
*SLC24A4**	No	No
DSG2	Yes	Yes
*INPP5D**	No	No
ZCWPW1	No	Yes
FERMT2	Yes	Yes
*CLU**	No	No
ADAM10	Yes	Yes
KAT8	Yes	Yes
ACE	Yes	Yes
ECHDC3	No	No

Table 3.8 Comparison of AD gene expression in primary microglia and model systems

Summary of differential expression of AD genes in primary microglia when compared to cancer cell models and iPSC-derived cells. Statistical differences determined by DESeq2 analysis and genes with an $p_{adj} < 0.05$ and LFC > 1. * next to a gene name highlights genes not captured by either of the model systems studied here.

3.7 Discussion

In this chapter I used publicly available RNA-sequencing datasets to compare the transcriptome of primary human microglia to a variety of *in-vitro* cell models. I obtained raw read level data from multiple independent studies and processed them using a uniform analysis pipeline. I showed that even with the uniform alignment and quantification pipeline, downstream analysis can still be impacted by normalisation techniques. The normalisation methods studied here, Log_2(TPM+1), QN and VST, had relatively low levels of overlap when identifying the top 500 most variable genes within the dataset, with less than 250 genes matching across all three methods. However, PCA using the top 500 most variable genes resulted in broadly similar sample distribution when PC1 vs PC2 scores for each sample were plotted. Variance components analysis revealed that, when expression at all genes was considered, study was the major driver of gene expression variation illustrating the importance of collecting data from the same cell type across multiple experiments.

Using PCA I was able to capture interpretable biological signals including the completeness of iPSC differentiation across PC1 and the differing cell types along PC2. Interestingly, PC2 also captured a separation in primary microglia samples with cultured primary microglia and fetal samples having lower PC2 scores than fresh adult/pediatric primary cells. It appeared that along this PC, these cells became more transcriptionally similar to iPSC-derived cells. Linking PCs with biological factors often requires prior knowledge of sample metadata to identify drivers of variation or technical batch effects. However, as the data collected for this study was mainly sourced from publicly available sources, I could only collect metadata provided alongside the samples. The amount of information about samples varied from source to source meaning there may have been technical batch effects within the dataset

that could not be identified and so the driver behind each PC could not be established.

When comparing primary microglia to all the model systems studied here many of the enriched gene sets were linked to neuronal processes. Previous work in primary human microglia, has shown that even culturing primary cells for 6 hours following dissociation of brain tissue can reduce the expression of specific gene patterns in primary cells[171]. Many of the genes that were identified as part of the environmentally linked signature described in primary cells including *TMEM119*, *CX3CR1* and *P2RY12*, were also identified as having significantly lower expression in the model systems when compared to primary microglia. This environmental signalling may also explain the separation of primary microglia samples along PC2, with cultured and fetal samples lacking the cues and stimuli from the developed CNS fully capture the microglia specific transcriptional signature.

Comparison of iPSC-derived macrophages to iPSC-microglia suggested that more specific differentiation protocols pushed differentiated cells closer towards the primary phenotype with significantly increased expression of genes such as *P2RY12* and *CX3CR1*. However, the iPSC-microglia still did not fully reflect the transcriptional signature of primary cells, and expression of microglial-linked TFs such as *SALL1* was lower in iPSC-derived cells. All of the iPSC-derived microglia samples used here represent monoculture systems, with only the chemical components of the differentiation media being used to push the cells towards the microglial phenotype. However, more complex differentiation protocols that involve culturing microglia alongside neurons have also been developed[198,200,202–206]. These culturing systems should more closely represent the brain environment, as they provide both the chemical stimuli and contact with neurons microglia may require for complete differentiation. This concept is explored further in Chapter 4 of this thesis, where I have used bulk and single cell RNA-sequencing of co-culture and organoid derived microglia, from a previously published protocols[200], to look at how neurons influence microglial gene expression.

As microglia are thought to be pathogenic cells in Alzheimer's disease[31], I also used this dataset to compare expression of disease risk genes across the model systems.

This builds on extended analysis carried out on the primary microglia dataset described in Chapter 2, in which it has been shown that iPSC-derived macrophages share a similar genetic architecture to primary microglia(Young *et al.* - paper in preparation). In the analysis carried out by Dr Natsuhiko Kumasaka, eQTL/GWAS co-localisations identified in primary microglia were replicated in iPSC-derived macrophages. However, as demonstrated this does not always translate to similar expression levels across cell types, genes such as *BIN1*, *APOE* and *CASS4* all had significantly higher expression in primary microglia compared to the iPSC model systems.

Chapter 4: Complex *in-vitro* model systems

Collaboration note

The samples collected as part of this chapter were processed as part of a collaboration with the Livsey Lab, based at the time at the Gurdon Institute and now at UCL. Stem cell differentiations were carried out by Dr Phil Brownjohn and Dr Moritz Haneklaus as well as 10X sample processing, along with Dr Julie Jerber. Bulk sample processing was completed by Dr Andrew Knights. All sequencing was completed at the Wellcome Sanger Institute, and initial analysis (alignment and quantification) of sequencing data was done by core informatics facilities at the institute.

4.1 Introduction

Work carried out in Chapter 3 of this thesis compared primary microglia to a variety of *in-vitro* model systems and highlighted that, while induced pluripotent stem cell (iPSC) based model systems provide a closer model system than cancer-cell lines, they still lack expression of many genes associated with primary microglia. Many of the genes with higher expression in primary microglia can be linked to neuronal and central nervous system (CNS) pathways. This suggests that the unique microglial transcriptomic signatures are driven by environmental stimuli in the brain that are not well captured by monoculture based *in-vitro* models. Consistent with this, freshly sequenced primary microglial samples have an environment dependent gene expression signature that is not observed in cultured primary cells[171].

While culturing primary human microglia has been shown to cause a reduction in expression of specific CNS-linked genes, it has also been demonstrated that culturing cells with factors that mimic the neuronal environment can rescue some of that expression[171]. Therefore, some of the monoculture iPSC microglia models use small compounds, such as C3CL1 and CD200, within the media of their cultures in order to better mimic the environment of the central nervous system (CNS)[198,201]. However, microglia are in constant contact with neurons[4] and it may be that it is a

mixture of both soluble factors in the CNS and physical contact with neurons that provides the signals needed for specific microglia gene expression.

4.1.1 Co-culture and organoid model systems

In order to better mimic the CNS environment of primary microglia in a dish, there have been methods developed to culture *in-vitro* microglia in the presence of neurons in order to push them closer towards the primary cell type. The most straightforward method is to co-culture single layers of both cell types together. Co-culturing iPSC-derived microglia with rat hippocampal neurons has been shown to cause a significant upregulation of 156 genes (adjusted $p < 0.01$), including *SIGLEC11*, *MITF* and *SLC2A5*, when compared to their monoculture iPSC-derived cells[198]. However as iPSC-derived neuronal differentiation protocols exist, it is also possible to culture iPSC-derived microglia alongside iPSC-derived neurons[202]. The media used in these co-culture systems often requires supplementation with compounds such as IL-34 and GM-CSF in order to maintain microglial survival and the distinctive ramified morphology of the cells. When compared to monocultured iPSC-derived macrophages, co-cultured microglial cells have been shown to have higher levels of expression of genes linked to chemotaxis/migration and regulation of cell adhesion[202].

While co-culture systems provide the most simple way to closer mimic the CNS environment, 3D organoid systems can provide an even more realistic method of modeling the brain environment in a dish. These culture systems use microfluidic culture platforms with different chambers for unique cell types[205] or spinning bioreactors[200,203,204,206] in order to maintain the 3D architecture of the organoids. It has been suggested that microglia will spontaneously form within certain neuronal organoids that are developed through embryoid body formation[204]. However, while the cells detected in these organoids are IBA1 positive and express *RUNX1* at comparable levels to primary microglia, expression of microglia marker genes such as *TMEM119*, *P2RY12* and *CX3CR1* were significantly lower. Expression of these genes increased as culture time increased, suggesting there was some maturation of the cells within the culture but never to a comparable level to primary cells.

Although it may be possible to allow microglia to spontaneously develop within neuronal organoids, iPSC microglia-like cells can also be differentiated externally and then added to already formed organoids. Brownjohn *et al.*[200] generated myeloid precursors through established iPSC differentiation protocols[192,193] and matured the precursors with IL-34 and GM-CSF to create a monoculture of microglia-like cells. The cells were then added to neuronal 3D organoids to understand how the microglia would interact with neuronal cultures. The iPSC-derived microglia were shown to rapidly migrate from the surface to deep within the organoid structure and assume a highly ramified morphology. The authors also noted that the microglia cells survived in the organoid culture using only the standard organoid culture media, they required no supplementation, suggesting of all the required signals for microglial survival were supplied by the neuronal culture system, unlike when using co-cultured models.

While some efforts have been made to compare these complex models to primary microglia and monoculture systems, no comprehensive analysis comparing all three has been carried out. This means it is not entirely clear whether culturing iPSC-microglia alongside iPSC-derived neurons moves them along a trajectory towards primary microglia.

4.1.2 Single cell sequencing and developmental trajectory inference

Bulk-RNA sequencing of iPSC-derived differentiated cultures can provide a method to look at how well the transcriptional profile of model systems captures the profile of the primary cell type being studied. However, as single cell RNA-sequencing technology has developed our ability to understand two key points of iPSC-differentiation has significantly increased. First, it provides researchers with the power to better understand the heterogeneity of cells within a differentiated population[253–255], which means rare populations can be identified that may be missed with bulk RNA-sequencing. Secondly, single cell sequencing allows researchers to track individual cells along a developmental or differentiation trajectory[256,257].

Computationally these dynamic processes within individual cells can be studied using trajectory inference methods, sometimes referred to as pseudotime analysis, in which cells are ordered along a process based on gene expression. There are a large

number of analysis tools available to run pseudotime analysis. Each of the tools has a unique algorithm for determining cell trajectories but they can broadly be split into two groups depending on whether they are built around free or fixed trajectory[258]. Monocle3 is one example of a free, unbiased algorithm that builds a tree based trajectory of cells along a differentiation pathway[259]. The package works by projecting cells onto a Uniform Manifold Approximation and Projection (UMAP) plot[260], clustering cells through a Louvain algorithm. The algorithm not only divides cells into clusters but also larger "partitions" of cells. When determining the trajectory pathways in a dataset, Monocle 3 can recognise the movement of cells within different partitions as distinct trajectories. The authors argue this removes the assumption from their model that every cell derives from a common ancestor cell.

The first part of this chapter focuses on this question by combining bulk RNA-sequencing data, generated in collaboration with the Livesey lab, from monoculture, co-cultured and organoid derived microglia with the large comparative dataset analysed in Chapter 3. I have then used single cell analysis and trajectory inference analysis to further understand how differing stem cell derived models of microglia may fit along a developmental trajectory. Using the tools available in the Monocle3 package, I have identified genes differentially expressed across the developmental trajectories in order to understand which cellular pathways are key to pushing *in-vitro* models of microglia towards the primary cell type.

4.2 Methods

4.2.1 Cell culture, dissociation and sorting

Monoculture stem cell derived microglia were derived using a previously developed protocol from within the Livesey lab[200]. Cultures were created using the H9 embryonic stem cell line and the KOLF_2 iPSC line, from the HiPSC database. For bulk sequencing samples, the two lines were cultured individually whereas the lines were combined for single cell sequencing. Stem cell derived neurons were cultured using an established protocol[261] and combined with fully differentiated stem cell-derived

microglia cells. Organoid cultures were also differentiated as previously described[200], although the number of days organoids were kept in cultured varied (between 12 and 15 days).

Sample dissociation was carried out using the Papain Dissociation System purchased from Worthington Biochemical Corporation. Cells were initially washed with PBS before being transferred into a 1.5 mL tube containing 200 µl of dissociation mix (Table 4.1) and incubated for 20-40 minutes. During the incubation cell solutions were agitated regularly or incubated directly on a heated shaking block. Following incubation, samples were then titrated to further break down clumps of cells before using centrifugation to pellet the cells. The cell pellet was resuspended in 175 µl of the inhibitor mix (Table 4.1) and then a further 90 µl of Ovomucoid and 90 µl of EBSS were added to the resuspended cell pellet. The cells were then centrifuged again and the resulting liquid was removed leaving the dissociated cell pellet. Dissociated cells were then used in the next stage of the processing pipeline, detailed in section 4.2.2 and 4.2.3. For samples that required cell sorting, pellets were resuspended in FACS buffer and sorted using CD45 FACS staining.

Dissociation mix	Inhibitor mix	FACS buffer
145 µl Papain	148.25 µl EBSS	18.6 ml PBS
10 µl Dnase I	8.75 µl Dnase I	1.33 ml BSA (7.5 %)
45 µl EBSS	17.5 µl Ovomucoid	80 µl EDTA (0.5 M)

Table 4.1 Buffer compositions for cell dissociation and sorting

4.2.2 Bulk sequencing preparation

As the numbers of isolated microglia cells from the complex model systems were relatively low the samples were processed by a slightly modified version of the low-input pipeline developed in-house by Dr Andrew Knights and described in section 2.2.3 of this thesis. Isolated cells were lysed directly in 50 µL of the lysis binding buffer described in Table 2.1, for monoculture cells this was following dissociation and for the complex models, this was after CD45 FACS sorting to isolate myeloid cells. The lysed samples were then directly added to oligo-DT beads without the need for a kit-based RNA extraction. The RNA-sequencing libraries were then

prepared exactly as described for the primary microglia samples in section 2.2.3. All samples used in this study went through a 14 cycle amplification PCR (Figure 2.2). Samples varied in cell number across the culture systems, with those isolated from the organoid systems falling in the lower range (Table 4.2).

Cell line	Culture system	Cell numbers
H9GFP	Co-culture	50k
KOLF2	Co-culture	50k
H9GFP	Co-culture	35k
KOLF2	Co-culture	32k
H9GFP	Co-culture	27k
KOLF2	Co-culture	50k
H9GFP	Organoid	12k
H9GFP	Organoid	7k
KOLF2	Organoid	7k
H9GFP	Organoid	6.5k
KOLF2	Organoid	13k
KOLF2	Organoid	23k
H9GFP	Monoculture	30k
KOLF2	Monoculture	30k
H9GFP	Monoculture	50k
KOLF2	Monoculture	50k
H9GFP	Monoculture	50k
KOLF2	Monoculture	50k
KOLF2	Monoculture	25k

Table 4.2 Sample summary for bulk RNA-sequencing

4.2.3 Single cell sequencing preparation

Samples generated for 10X single cell sequencing were a mixture of sorted and unsorted samples, summarised in Table 4.3. Single cell suspensions were processed by the Chromium Controller (10x Genomics) using single Cell 3' Reagent Kit v2 (PN-120237). All the steps were performed according to the manufacturer's specifications. Barcoded libraries were sequenced using HiSeq4000 (Illumina, one lane per 10x chip position) with 75bp paired end reads. Information regarding the

number of cells loaded into each inlet as well as the number of returned cells and resulting reads/cell can also be found in Table 4.3.

Culture system	FACS	Days in culture	Number of cells loaded	Number of cells sequenced	Mean reads per cell
monoculture	unsorted	NA	16670	8675	40800
co-culture	unsorted	NA	21537	8353	41685
organoid	unsorted	12	25826	9045	36152
organoid	sorted	12	9835	4215	73349
organoid	sorted	15	8450	3223	98736
organoid	unsorted	15	17765	8862	35045

Table 4.3 Sample summary for 10X single cell sequencing

4.2.4 Bulk RNA-sequencing data processing and analysis

In order to ensure continuity with the data analysed in Chapter 3 of this thesis, raw bulk RNA-sequencing data generated as part of this data was processed through the same pipeline: STAR followed by featureCounts quantification. Following Log_2(TPM+1) normalisation, I again used the prcomp function in R to carry out principal components analysis (PCA), principal components (PCs) were calculated using the top 500 most variable genes or genes identified as having significantly higher expression in primary microglia when compared to all monocultured models (see section 3.5.1). I also used the varimax function to rotate calculated PCs to identify the highly loaded genes for each PC. I extended my dimensionality reduction analysis to also compute PCs from the residuals following linear regression study effects, to control for the known batch effects that can arise when comparing across sequencing studies. Residuals were calculated for each sample across each gene using either of the following linear model:

lm (expression ~ study)

Differential expression analysis was carried out using the DESeq2 package[248] with sequence preparation, (normal or low-input library preparation) used as a variable in the analysis. Gene lists were run through gene set enrichment analysis using g:OSt

function of the online gProfiler tool[226]. For full description of the analysis pipelines see section 3.2.3.

4.2.5 Single cell RNA-sequencing data processing and quality control

10X single cell samples were aligned and quantified using cellranger version 3.0.2 and GRCh38, the final combined dataset contained 42317 cells. Following Seurat's standard preprocessing pipeline, I calculated the percentage of mitochondrial genes across samples and filtered out cells with > 10% mitochondrial genes to remove dying cells. I also removed cells with less than 100 or greater than 3000 features to remove poor quality cells and potential doublets. Following these quality control steps, 31259 cells remained for further analysis. Data was then normalised and scaled, before PCA was run on the 3000 most variable genes. I then ran clustering and UMAP analysis using 15 PCs and a 0.5 resolution. I used known myeloid marker gene (CD45 and AIF1) expression to identify and subset the microglia-like cells from the dataset, identifying 8928 myeloid cells for downstream analysis.

4.2.6 Cluster identification, differential expression analysis and trajectory analysis

Filtered and subsetted raw count data for the identified myeloid cells was then processed using the Monocle3 package[259]. Raw count data was normalised and preprocessed using the first 100 PCs. Normalisation was carried out by the estimation of size factors for each cell and dispersions across genes before $\log_{10}$ normalisation. UMAP analysis was used to visualise the cells and the cluster_cells function within Monocle3 was used, with a resolution of $1x10^{-4}$, to group cells. The initial clustering of cells by Monocle3 used "community detection" as a method of classifying cells[262] which was first used as part of the phenoGraph package[263]. As well as grouping cells into "clusters" the cluster_cells function also split cells into "partitions" using the PAGA algorithm[264], which are considered more "well separated" cells than those seen in clusters. Partition markers were identified using the "top_markers" function, across all genes, and significant markers were identified as those with a q value (FDR corrected p value) of < 0.05.

The initial trajectory graph was identified using the “learn_graph” function of Monocle3 before cells were ordered along a pseudotime using the “order_cells” function. The function requires the selection of a “start node”, i.e. the group of cells thought to represent the earliest point in the developmental pathway. For this analysis the start node was selected by identifying the earliest branch node from the trajectory analysis. Genes whose expression was significantly linked to a position within the pseudotime were identified using the “graph_test” function. This runs a spatial autocorrelation analysis, known as Moran’s I, which identifies correlations of gene expression in cells considered in nearby space to each other[259], which in this case means cells in close space within the pseudotime trajectory. Again significant genes were identified as those with a q value of < 0.05.

4.3 Bulk RNA-sequencing comparison of complex and simple model systems

4.3.1 Dimensionality reduction

Following initial processing of data I combined the newly generated samples with the gene counts matrix used in Chapter 3 and then calculated $Log_2(TPM+1)$ normalised counts for all samples. I ran PCA across the dataset, using the top 500 most variable genes and plotted the samples based on their PC scores. Figure 4.1 shows samples, plotted based on PC1 vs PC2 and coloured by cell type with the new samples included. While the distribution of samples with new samples was broadly similar to the original dataset (Figure 3.5 A) there are two important points to note. Firstly the iPSC-derived and ES-derived (red data points in Figure 4.1) monoculture samples clustered close to the other monoculture samples, despite being from different studies. Secondly, the co-cultured and organoid derived microglia moved slightly further up PC2 closer to the primary microglia than the monoculture models. This suggested that for genes heavily loading PC2, the complex model microglia had an expression profile more similar to that of primary microglia than their monoculture counterparts.

To get a clearer picture of the drivers of variation within the updated dataset I also continued to plot the samples further down the PCs. Figure 4.2 shows samples plotted on the PC3 vs PC4 axis coloured by cell type (A) and stimulation (B). Figure 4.2 shows samples plotted on the PC4 vs PC5 axis coloured by study (C) and sequencing preparation method (D). Although simply looking at the PC plots does not provide comprehensive proof of what may have been driving variation in the dataset, PC3 appeared to capture a stimulation effect while PC5 may have represented a mixture of study and sequence preparation effects.

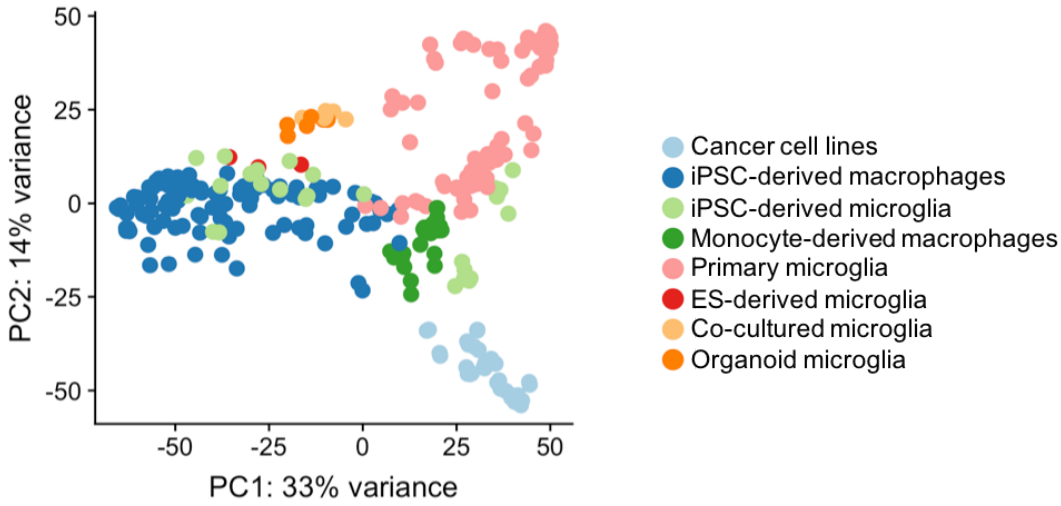


Figure 4.1 PC1 vs PC2 of model comparison dataset

Principal components analysis (PCA) across the top 500 most variable genes, plotted as PC1 vs PC2 scores and coloured by cell type. The original dataset (A), described in Chapter 4, is included for comparison to the complete dataset described in this chapter (B).

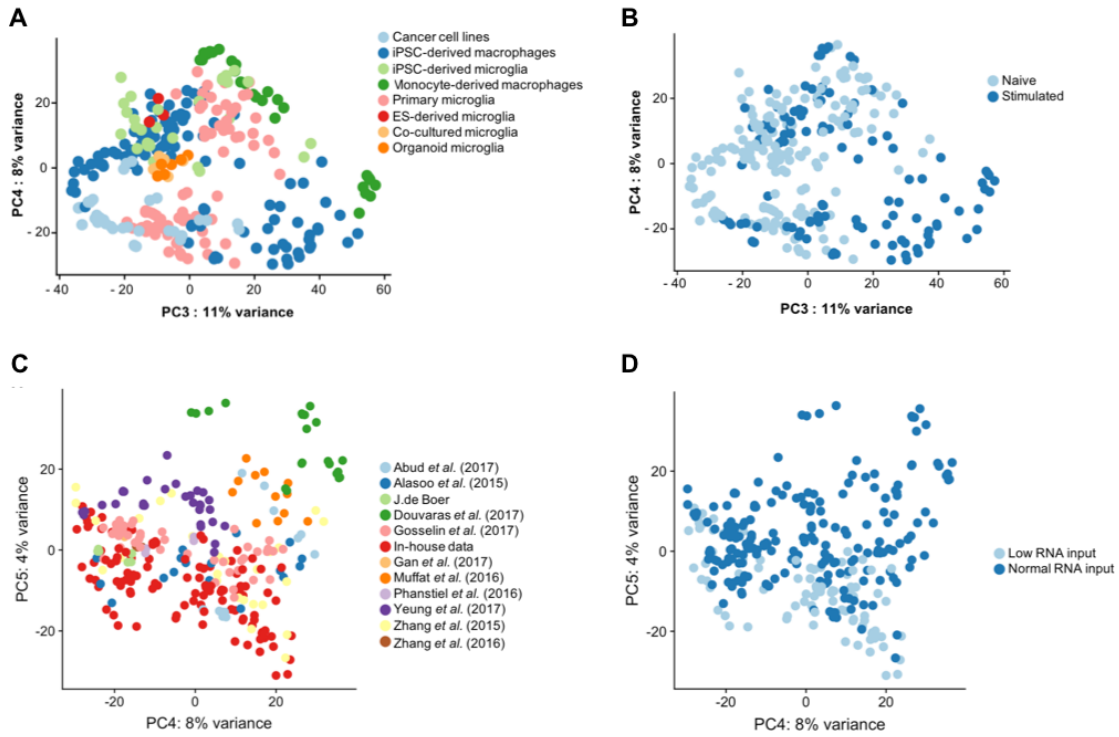


Figure 4.2 PC3 vs PC4 and PC4 vs PC5 of model comparison dataset

Principal components analysis (PCA) across the top 500 most variable genes, plotted as PC3 vs PC4 scores and coloured by cell type (A) and stimulation (B) and PC4 vs PC5 scores and coloured by study (C) and sequencing preparation method (D).

As well as looking at the visual representation of the PCA, I used varimax analysis to determine which of the most variable genes used in the PCA was driving each component. Table 4.4 shows the top 5 most heavily loaded genes for each PC, which were compared to the genes identified using the same analysis for the original dataset, see table 3.5 in section 3.3.4. The majority of genes identified in the varimax analysis matched those seen in the original dataset and the PCs had a similar sample spread.

PC1		PC2		PC3	
+ve	**-ve**	**+ve**	**-ve**	**+ve**	**-ve**
CAT	*COL3A1*	*FOSB*	*CCL13*	*CXCL10*	*GPR34*
MMP9	*COL1A1*	*CH25H*	*S100A4*	*IDO1*	*ADORA3*
CCL22	*IGFBP5*	*P2RY12*	*ANXA2*	*ACOD1*	*SLC40A1*
CHI3L1	*POSTN*	*CX3CR1*	*CD36*	*TNFAIP6*	*PALD1*
CSTA	*CCN2*	*EGR3*	*MMP9*	*CCL8*	*PDK4*

MARCO	*CCN1*	*EGR1*	*IGFBP4*	*CXCL11*	*MAF*
CD48	*COL1A2*	*SALL1*	*ANPEP*	*RSAD2*	*DDIT4L*
CD52	*LUM*	*SIGLEC8*	*DDIT4L*	*CCL5*	*P2RY12*
S100A4	*LOX*	*DUSP1*	*MT-TN*	*SLAMF7*	*HPGDS*
AC245128.3	*SERPINE1*	*LINC01736*	*CYP1B1*	*CXCL9*	*GPR82*
PC4		**PC5**			
+ve	**-ve**	**+ve**	**-ve**		
RNASE1	*ELANE*	*RN7SL2*	*RNASE2*		
C1QC	*CTSG*	*RN7SL3*	*MT-TA*		
STAB1	*AZU1*	*CHIT1*	*F13A1*		
C1QA	*PRTN3*	*SCARNA7*	*MT-TL1*		
C1QB	*CES1*	*HIST1H1E*	*IL1B*		
CCL13	*CITED4*	*CYP27A1*	*RNA5SP151*		
VSIG4	*SLPI*	*RN7SL471P*	*MT-TN*		
GPR34	*CD70*	*FBP1*	*MT-ATP8*		
MRC1	*ASS1*	*C015660.2*	*RPL41P1*		
SPP1	*COL9A2*	*SCARNA21*	*AC090498.1*		

Table 4.4 Varimax analysis of the first 5 PCs

Varimax analysis of the first 5 principal components from the top 500 most variable genes. Top 5 most negatively and positively loaded genes for each component.

While the principal components analysis described above, suggested that the complex models may move closer to the primary phenotype, it did not control for known study based batch effects. Variance components analysis on the original dataset (Figure 3.3) identified study as the largest driver of variation across all genes in the dataset and it is therefore important to take this potential batch effect into account when comparing samples. I used linear regression to calculate the residuals for each gene across all samples when fitting study as a random effect. I then used the residuals as input for PCA, using both all genes (Figure 4.3) and the top 500 most variable genes (Figure 4.4). While the regression of study based effects allows for the control of potential study based effects, as this analysis compares cell types across different studies, the effects may have been confounded. This is highlighted in Figures 4.3 B and 4.4 B whereby samples from cancer cell lines are clustered with

primary microglia, despite differential expression analysis (section 3.5.2) highlighting large transcriptional differences between the cell types. This suggested that using a linear model to regress out study based effects, may have also removed some of the biology that is confounded by the study.

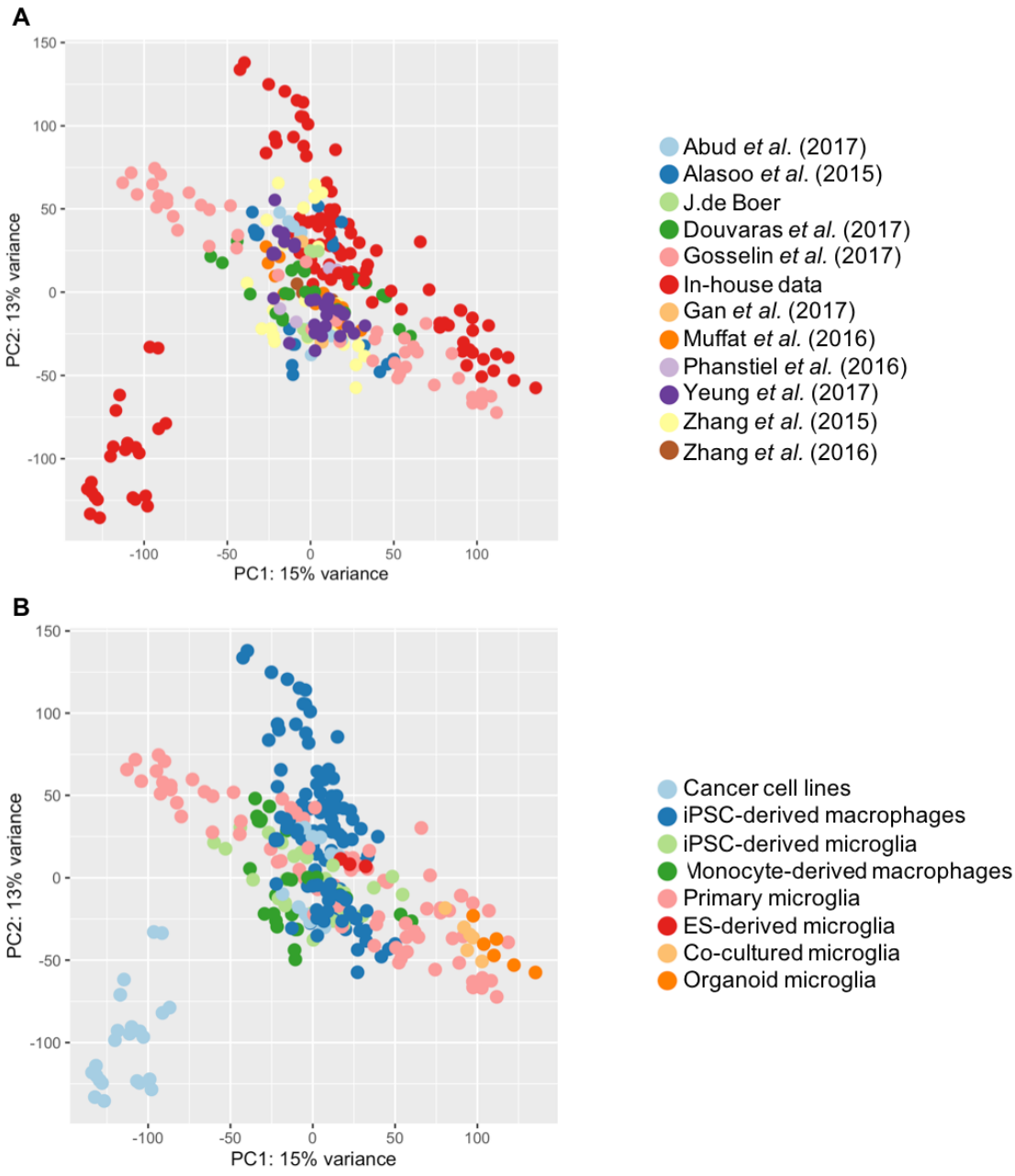


Figure 4.3 PC1 vs PC2 of residual values across all genes following removal of study based effects

Principal components analysis (PCA) calculated, using residuals from a linear regression of study effects, across all genes. Samples are plotted by PC1 vs PC2 scores and are coloured by study (A) and cell type (B).

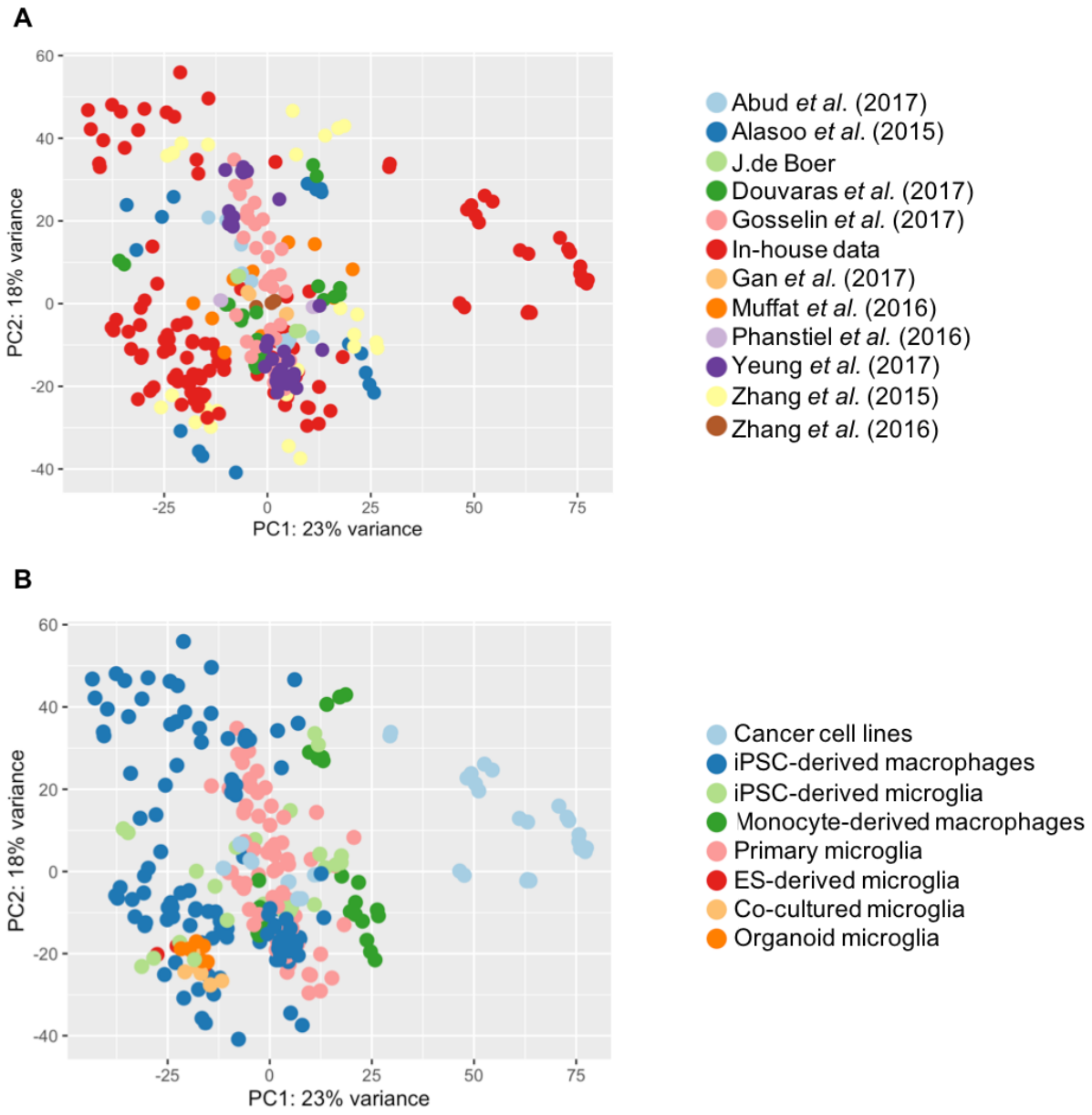


Figure 4.4 PC1 vs PC2 of residual values from the top 500 most variable genes following removal of study based effects

Principal components analysis (PCA) calculated, using residuals from a linear regression of study effects, across the top 500 most variable genes. Samples are plotted by PC1 vs PC2 scores and are coloured by study (A) and cell type (B).

As well as using linear models to regress out study based effects for input into PCA, I also ran the analysis using Log_2(TPM+1) normalised values for the 7297 genes identified as part of the PMM dataset (section 3.5.1) as shown in Figure 4.5. The PMM gene set was identified as genes with a significantly higher expression in primary microglia than all the monocultured based models studied in Chapter 3 of

this thesis. Importantly the analysis used to identify this gene set controlled for study based batch effects.

Figure 4.5 shows that when using these genes as input for PCA, PC1 captured variability in cell type with primary microglia most positively loading the PC. The primary microglia were again separated along the first PC, with cultured and fetal microglia sitting closer to the monocultured *in-vitro* models (Figure 4.5B). Using the PMM gene set as input for PCA also showed the complex *in-vitro* models were closer on PC1 to fresh primary microglia.

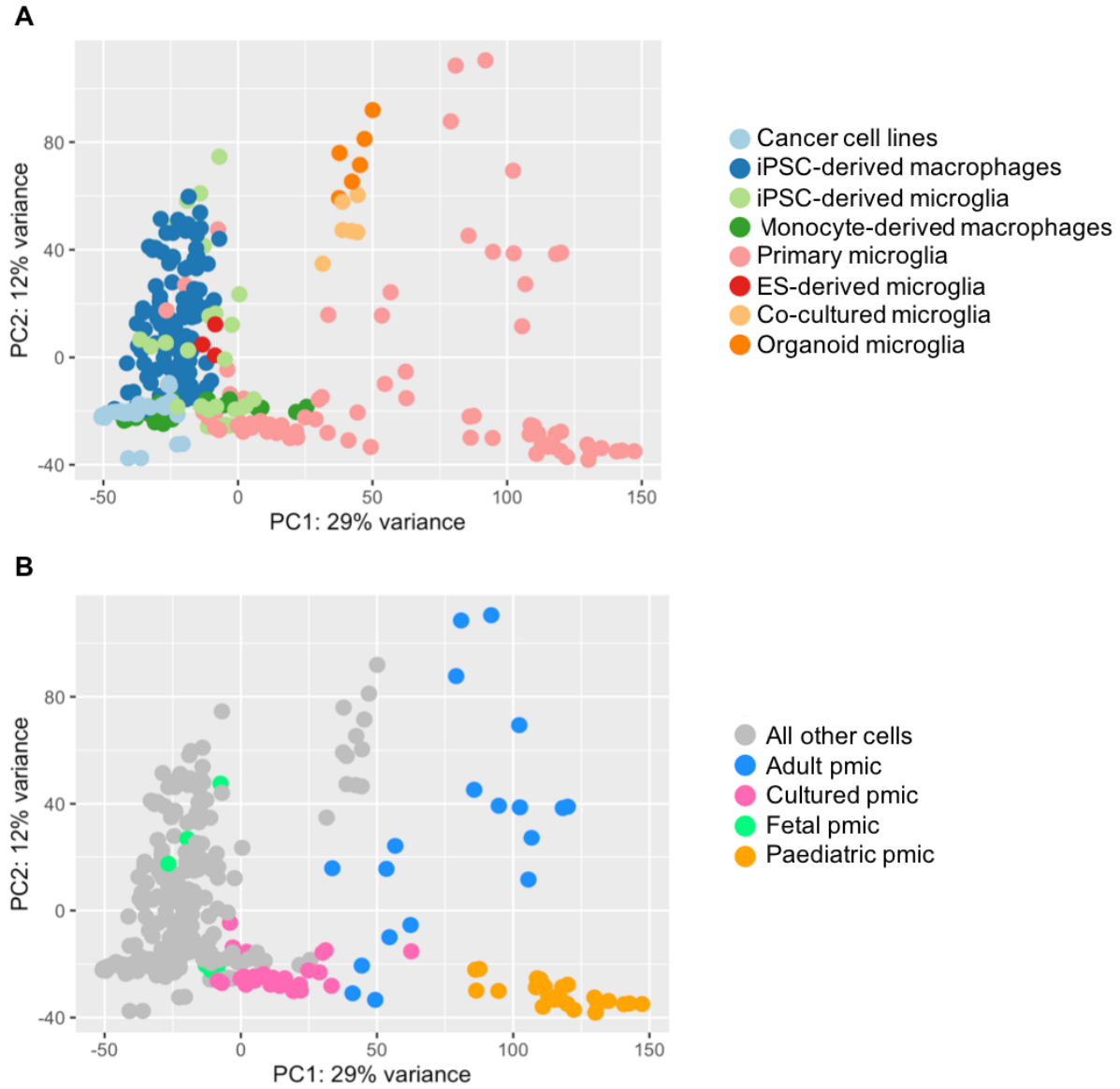


Figure 4.5 PC1 vs PC2 of all samples using the PMM input gene list

Principal components analysis (PCA) calculated using the 7297 genes identified in the PMM gene set (section 3.5.1). Samples are plotted by PC1 vs PC2 scores and are coloured by cell type (A) and primary microglia source (B).

4.3.2 Differential expression analysis

The dimensionality reduction techniques described in the section above provide useful tools for understanding global patterns of gene expression across the model systems. However, I was also interested in specific differences in gene expression when comparing the complex model systems to both their monoculture counterparts and primary microglia. As the number of samples collected for the model systems in this bulk analysis was relatively small, differential expression (DE) was run with these samples as one “complex models” group of samples.

Initially I compared monocultured iPSC-derived microglia to the stem cell derived complex models and found that there were only 760 genes expressed at a significantly higher level in the monoculture model systems whereas 4783 genes were more highly expressed in the complex models ($p_{adjust} < 0.05$ and +/- 1 $\log_2$ fold-change (LFC)). The majority of gene expression changes between monoculture and complex models involved higher gene expression in the complex models (as highlighted by the MA plot in Figure 4.6) at the lower end of expression which suggested that genes were mainly “switched on” in the presence of neurons.

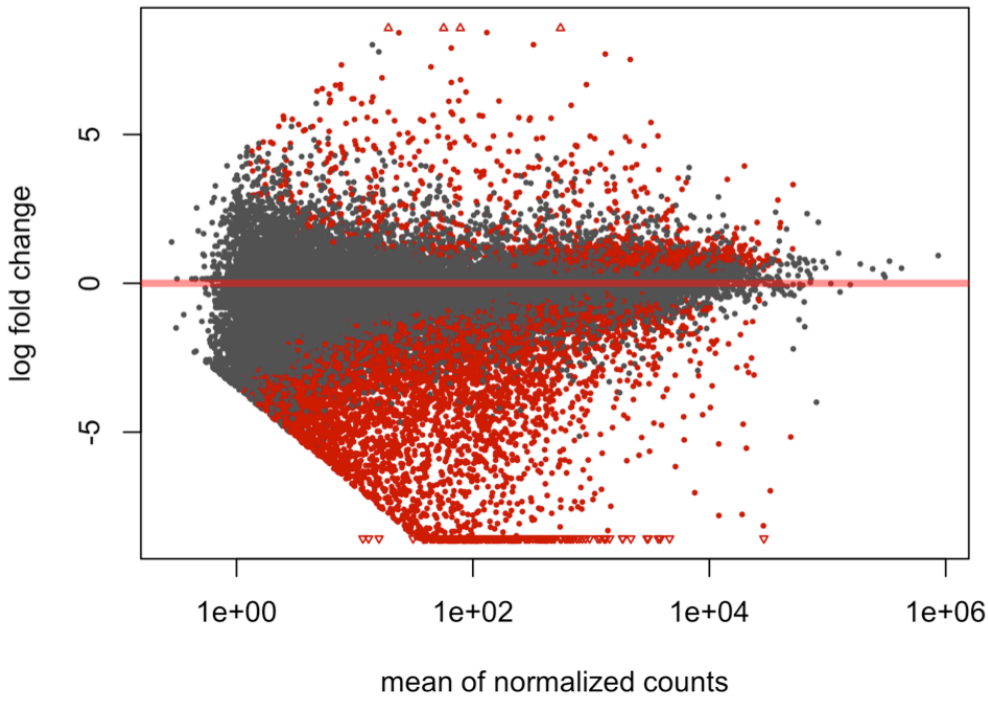


Figure 4.6 MA plot of differentially expressed genes comparing monoculture vs complex stem cell derived microglia

Log_2 fold change (LFC) plotted against the mean of normalised counts for each gene tested when comparing monoculture iPSC-derived microglia to iPSC-derived microglia from complex model systems. Points coloured in red are those reaching significance (following 5% FDR correction) and triangular points represent genes that have a LFC outside the limits of the graph.

Using the online gProfiler tool I ran gene-set enrichment analysis (GSEA) within the differential expressed genes. The small number of genes with higher expression in the monoculture systems were linked to extracellular matrix pathways and pattern specification process, which have been linked to cell differentiation, suggesting that monocultured stem cell derived microglia may represent a less mature cell or less complete differentiation. GSEA of the genes more highly expressed in complex models showed an enrichment for nervous system development and neuronal differentiation (Table 4.5). This suggested that culturing stem cell derived microglia alongside neurons may help to capture some of the CNS-linked transcriptional signature seen in primary microglia.

Term name	Term ID	P_{adj}
nervous system development	GO:0007399	$8.99e^{-67}$
neuron differentiation	GO:0030182	$2.17e^{-48}$
neurogenesis	GO:0022008	$6.15e^{-48}$
generation of neurons	GO:0048699	$8.22e^{-48}$
chemical synaptic transmission	GO:0007268	$1.87e^{-45}$
anterograde trans-synaptic signaling	GO:0098916	$1.87e^{-45}$
trans-synaptic signaling	GO:0099537	$1.98e^{-45}$
cell projection organization	GO:0030030	$3.93e^{-45}$
synaptic signaling	GO:0099536	$1.06e^{-44}$
plasma membrane bounded cell projection organization	GO:0120036	$8.95e^{-43}$

Table 4.5 GSEA on genes with higher expression in CD45+ from complex models when compared to monoculture cells.

Statistical enrichment analysis through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Top ten GO: biological process terms

I also ran DE to compare the complex model samples to the primary microglia and found 4622 genes with significantly higher expression in primary cells, including known microglia marker genes such as P2RY12, CX3CR1 and TMEM119. GSEA (Table 4.6, left hand column) for these genes showed an enrichment for cell activation terms. There were also 5536 genes with a significantly higher expression in the complex model systems, including the CSF2RA gene, which is involved in macrophage differentiation. Within the genes more highly expressed in the model systems there was a significant enrichment for genes linked to the axoneme and cilium assembly (Table 4.6) which could be linked to the formation of the ramified morphology seen in microglial cells. Interestingly, both gene lists showed enrichment for CNS linked terms. Genes with higher expression in primary microglia were enriched for terms such as oligodendrocyte differentiation (GO:0048709, $p_{adj} = 1.51e^{-7}$) and central nervous system myelination (GO:0022010, $p_{adj} = 4.7e^{-7}$) while genes with higher expression in the complex models were enriched for terms like

forebrain development (GO:0030900, p_{adj} = 0.003) and brain morphogenesis (GO:0048854, p_{adj} = 0.005).

Primary microglia			Complex models		
Term name	**Term ID**	**Padj**	**Term name**	**Term ID**	**Padj**
leukocyte activation	GO:0045321	$4.67e^{-16}$	cilium assembly	GO:0060271	$4.92e^{-13}$
cell activation	GO:0001775	$4.67e^{-16}$	cilium organization	GO:0044782	$6.23e^{-13}$
immune response	GO:0006955	$1.24e^{-15}$	microtubule-based movement	GO:0007018	$2.96e^{-12}$
immune system process	GO:0002376	$3.01e^{-14}$	cilium movement	GO:0003341	$1.29e^{-11}$
interferon-gamma-mediated signaling pathway	GO:0060333	$1.51e^{-13}$	microtubule-based process	GO:0007017	$4.51e^{-11}$

Table 4.6 GSEA on DE genes comparing primary microglia to complex models

Statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Five most significantly enriched biological process terms for both genes with higher expression in primary cells and complex models when compared to each other.

4.4 Identification and clustering of myeloid cells within the single cell dataset

To extend the analysis carried out with the bulk sequencing data, I wanted to understand how the three *in-vitro* model systems varied at the single cell level and whether culturing stem cell derived microglia with neurons moved the cells further along a developmental trajectory.

4.4.1 Clustering analysis to identify myeloid cells within the full population

The single cell dataset generated for this study was from a mixture of sorted and unsorted samples from the complex model systems and therefore contained a

mixture of myeloid and non-myeloid cells. Following removal of poor quality cells, (high mitochondrial gene percentage and too many or too few captured genes), I normalised and scaled the 31259 cell dataset. Following PCA, I used the top 15 PCs to run UMAP analysis (Figure 4.7).

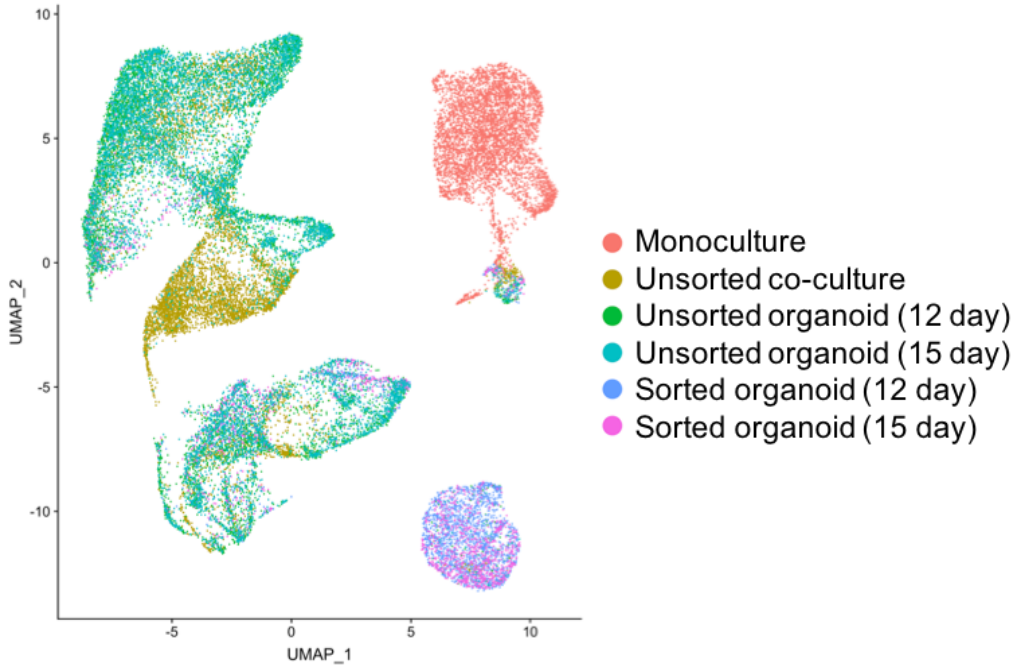


Figure 4.7 UMAP of full dataset

UMAP analysis following Seurat filtering, normalisation and scaling. UMAP run using the RunUMAP function of Seurat, using the first 15 principal components. Cells coloured by model system

Following initial UMAP analysis, I ran clustering analysis using Seurat's graph based clustering algorithm with the first 15 principal components and a resolution of 0.5 (Figure 4.8 A) and also looked at expression of known myeloid cell marker genes, *CD45* and *AIF1* (Figure 4.8 B and C). Expression of myeloid marker genes was only seen in clusters 1, 4, 11 and 12 and therefore these cells were subsetted from the original dataset for downstream analysis.

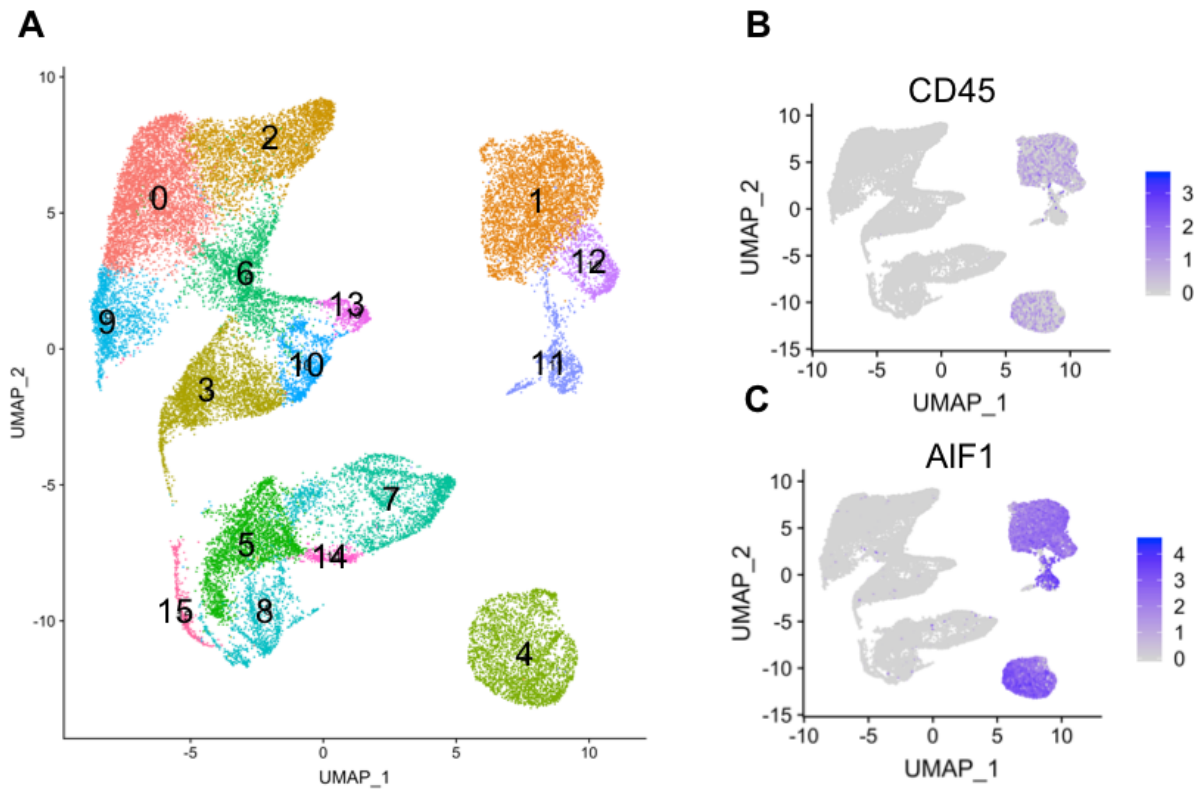


Figure 4.8 Identification of myeloid cells

UMAP analysis following Seurat filtering, normalisation and scaling. UMAP run using the RunUMAP function of Seurat, using the first 15 principal components. Clustering carried out using Seurat's clustering algorithm using 15 principal components and a 0.5 resolution. Cells coloured by: cluster (A) and expression of myeloid marker genes CD45 (B) and AIF1 (C).

4.4.2 Partition and cluster analysis using Monocle3

Following quality control filtering and identification/separation of the myeloid cells from within the single cell dataset, I used the raw data and processed the new myeloid only dataset, through the standard Monocle3 processing pipeline. Initially, I used UMAP analysis to visualise the cells and Figure 4.9 shows each cell coloured by the sample it originated from. The UMAP plot was split into three major groups of cells, one made up of entirely cells from the monoculture system and a second made up of cells originating from all the model systems studied. The final large group of cells, was dominated by CD45 sorted myeloid cells from organoid culture systems. However, there were also cells present in this cluster that were from the unsorted organoid and unsorted co-culture model systems. The fraction of these cells within

the larger cluster was small but this may be due to a smaller number of cells arising from these samples in total (2817 cells from sorted organoid sample versus 206 and 299 from the unsorted co-culture and organoids respectively).

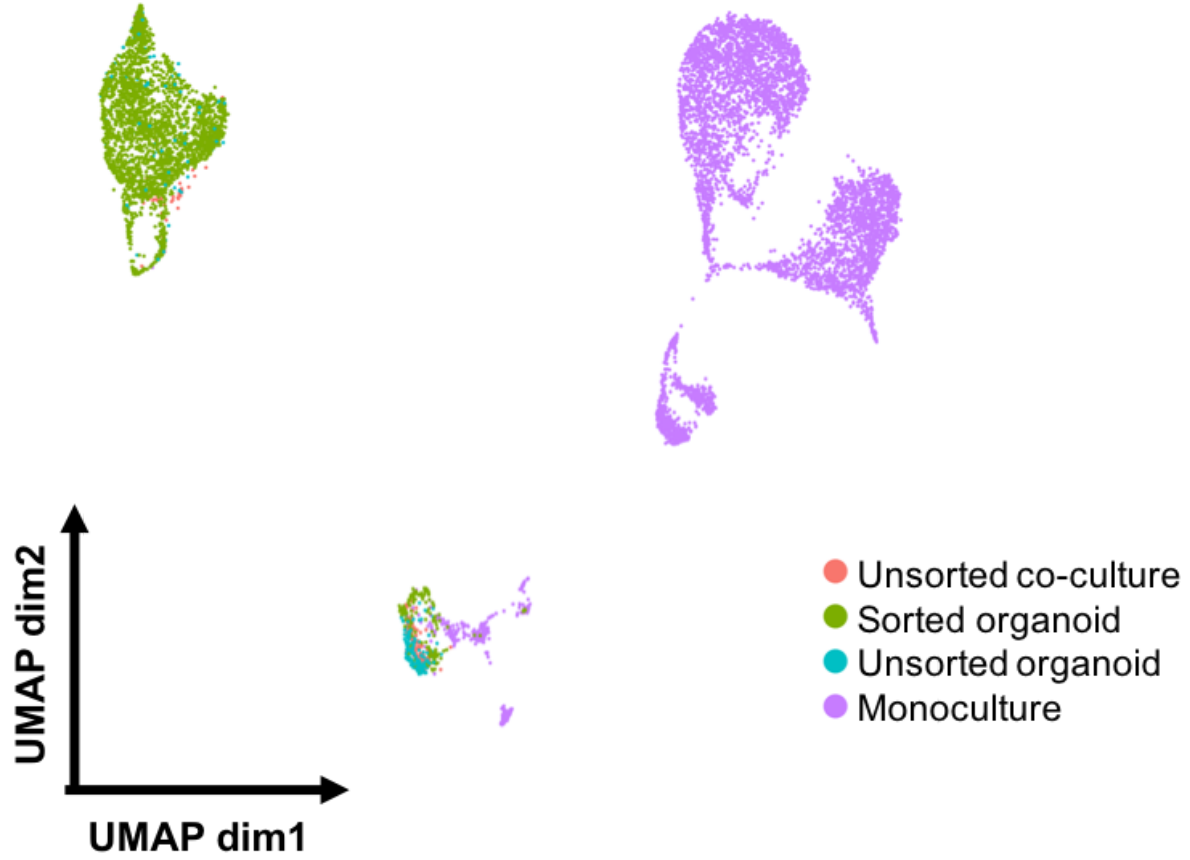


Figure 4.9 UMAP of myeloid cells in Monocle3

UMAP analysis following Monocle3 preprocessing. UMAP run using the reduce_dimension function of Monocle3. Cells coloured by model system.

After running UMAP analysis to visualise the cells, I used the “cluster_cells” function to formally group cells. Figure 4.10 shows the UMAP plot of cells coloured by both partitions (A) and clusters (B) and Figure 4.11 summarises the number of cells within each partition attributed to the different culture systems (A) and the partition assigned to the cells from each culture system (B).

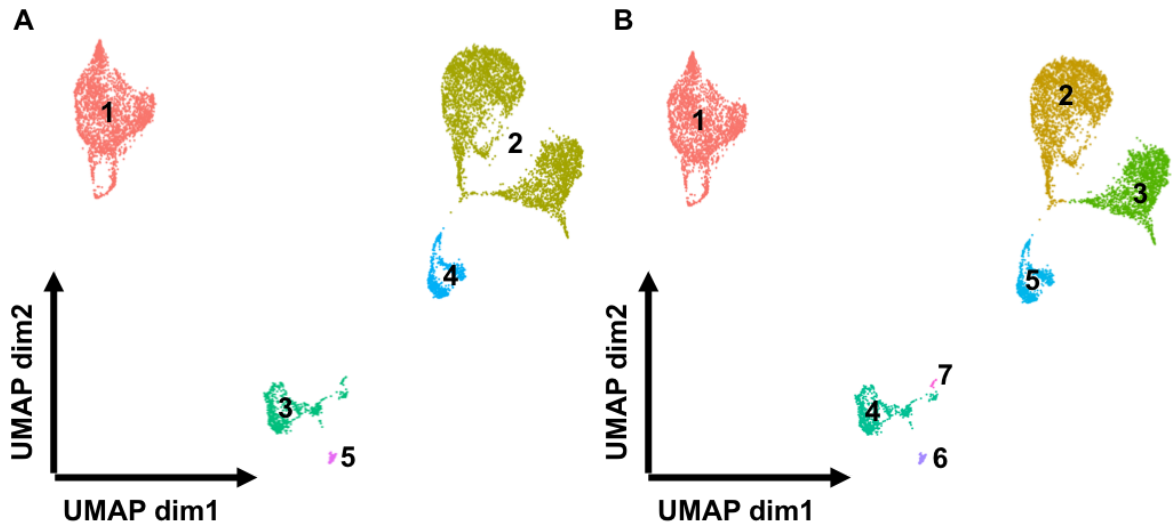


Figure 4.10 UMAP of myeloid cells in Monocle3

UMAP analysis following Monocle3 preprocessing. UMAP run using the "reduce_dimension" function of Monocle3. Cells coloured by partition (A) and cluster (B) determined by the "cluster_cells" function.

Interestingly, three partitions (2, 4 and 5) only contained cells from within the monoculture system whereas partitions 1 and 3 were made up of cells from each model system studied here, although the contribution of monoculture based cells to partition 1 was minimal (2 cells). This suggests that monoculture differentiations generate a more heterogeneous population of cells than complex models. As suggested above, partition 1 was dominated by cells from the sorted organoid sample, 2639 cells out of 2800 total, but 35% of cells from the unsorted organoid and 26% of cells from the co-culture system were also present in this partition just at lower absolute numbers.

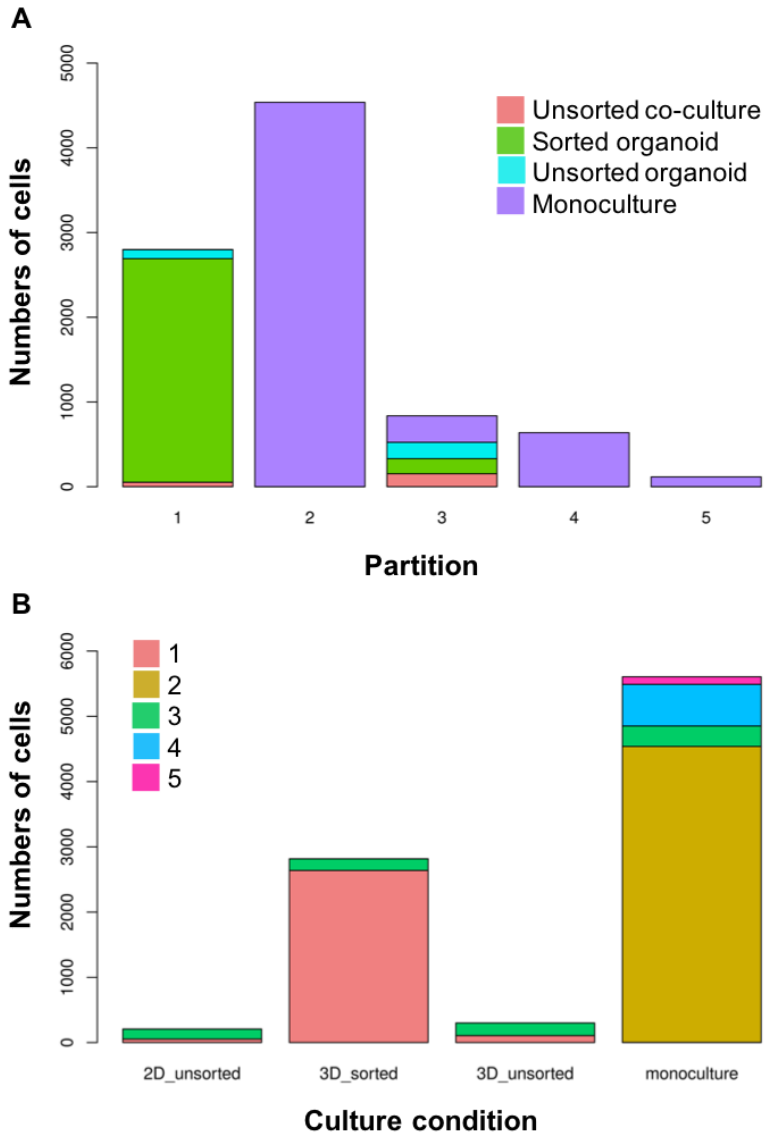


Figure 4.11 Number of cells in each partition

Number of cells in each partition, using monocle3 “cluster_cells” function, coloured by the culture system the cells originated from (A). Number of cells in the culture system coloured by the partition, using monocle3 “cluster_cells” function, the cells were assigned to (B).

4.4.3 Partition marker genes

First, I wanted to identify differentially expressed genes within each partition, using the “top_marker” function, to understand what transcriptional changes may have been impacting the partitioning of the cells. Figure 4.12 highlights specific marker genes for each partition (labelled 1-5)

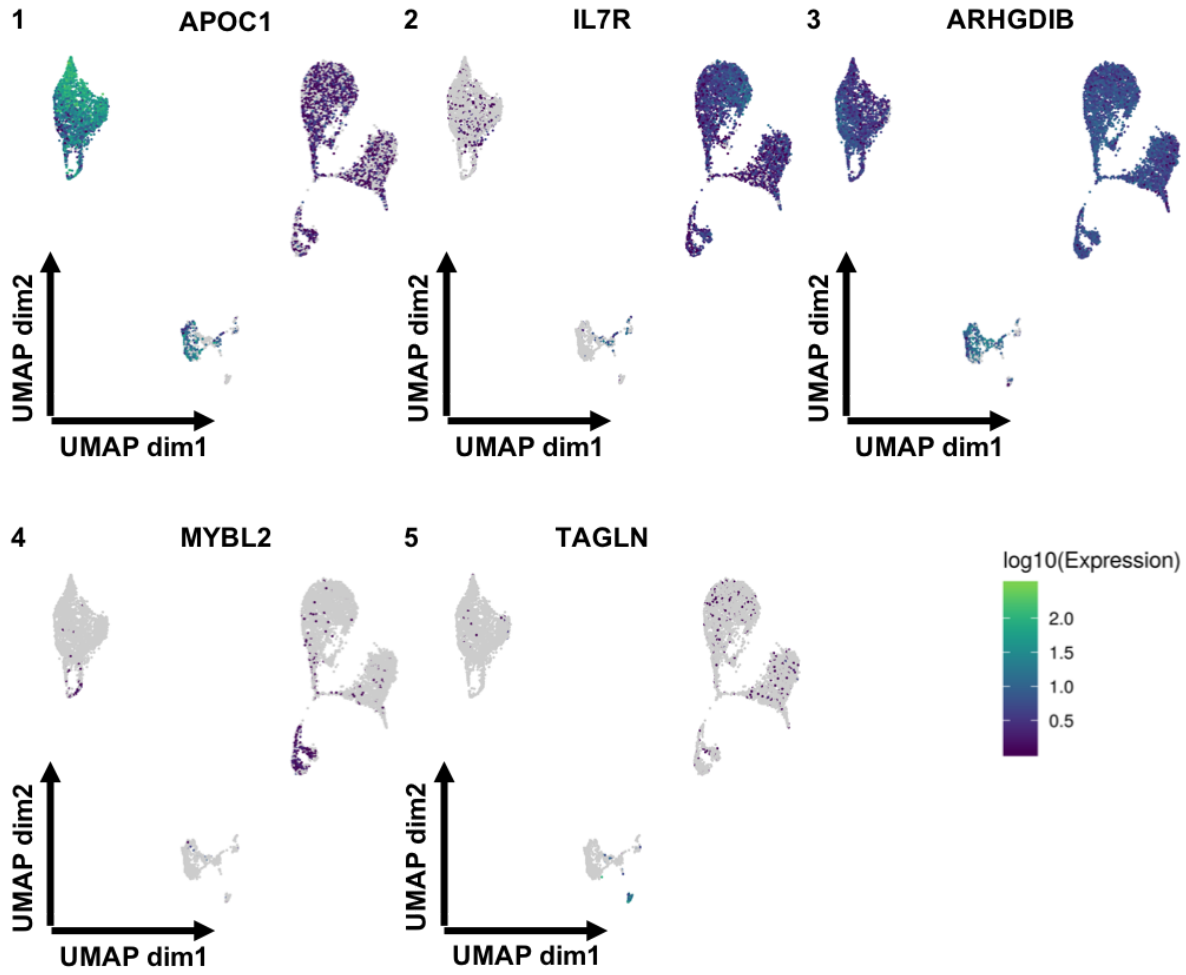


Figure 4.12 UMAP of myeloid cells in Monocle3 coloured by marker gene expression

UMAP analysis following Monocle3 preprocessing. UMAP run using the "reduce_dimension" function of Monocle3. Cells coloured by expression of marker genes for each partition (1-5) determined by "top_marker" function of Monocle3.

Table 4.7 highlights the top gene markers for each partition (based on the marker score) and the top enriched biological process terms for the 50 marker genes identified for each partition. The partitions only associated with only monoculture cells (2, 4 and 5) were all enriched for distinct gene sets, which suggested they represented different subpopulations of cells within the same culture system. Partition 2 for instance, appeared to represent a more activated population of cells while partition 5 cells were linked to cytoskeleton terms. Partition 3 cells were enriched for endoplasmic reticulum and protein targeting terms.

Of the top 50 partition 1 marker genes, 28 were also identified within the PMM gene set, described in section 3.5.1 in this thesis, which included genes with higher expression in primary microglia compared to the simple *in-vitro* model systems. This was compared to between 1 and 4 overlapping genes in the other partitions. This suggested that partition 1 cells may represent a population closer to that of primary microglia, with increased expression of genes such as *APOC1*, *CCL3L1* and *PDK4*. GSEA of partition 1 markers highlighted an enrichment in cell migration genes as well as genes associated with organic substance response which would support this theory. As the cells in partition 1 were mainly associated with organoid samples, they would be expected to be more active than those in a monoculture system as they would be constantly responding to and interacting with neurons.

		GSEA		
Partition	**Marker genes**	**Term name**	**Term ID**	**padj**
1	*CCL4L2*	response to organic substance	GO:0010033	$3.38e^{-07}$
	APOC1	ERK1 and ERK2 cascade	GO:0070371	$3.38e^{-07}$
	RNASET2	response to stress	GO:0006950	3.38^{e-07}
	CCL3L1	response to external stimulus	GO:0009605	$6.37e^{-07}$
	ABCA1	mononuclear cell migration	GO:0071674	$7.34e^{-07}$
2	*IL7R*	leukocyte activation	GO:0045321	$1.22e^{-14}$
	FTH1	neutrophil degranulation	GO:0043312	$1.77e^{-14}$
	CCL13	cell activation involved in immune response	GO:0002263	$1.77e^{-14}$
	BRI3	leukocyte activation involved in immune response	GO:0002366	$1.77e^{-14}$
	S100B	neutrophil activation involved in immune response	GO:0002283	$1.77e^{-14}$
3	*ACTB*	SRP-dependent cotranslational protein targeting to membrane	GO:0006614	$3.29e^{-39}$
	GAPDH	cotranslational protein targeting to membrane	GO:0006613	$6.54e^{-39}$
	EEF1A1	protein targeting to ER	GO:0045047	$3.40e^{-38}$
	ARHGDIB	establishment of protein localization to endoplasmic reticulum	GO:0072599	$6.69e^{-38}$
	AIF1	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	GO:0000184	$4.10e^{-37}$
4	*PCLAF*	electron transport chain	GO:0022900	$3.24e^{-05}$

	TOP2A	oxidation-reduction process	GO:0055114	$3.24e^{-05}$
	DEK	oxidative phosphorylation	GO:0006119	$6.76e^{-05}$
	HIST1H4C	leukocyte activation	GO:0045321	$7.24e^{-05}$
	MYBL2	mitochondrial ATP synthesis coupled electron transport	GO:0042775	$8.17e^{-05}$
5	*TAGLN*	actin filament-based process	GO:0030029	$2.39e^{-09}$
	TPM2	actin cytoskeleton organization	GO:0030036	$2.89e^{-08}$
	TPM1	symbiotic process	GO:0044403	$5.93e^{-08}$
	KRT18	cytoskeleton organization	GO:0007010	$5.93e^{-08}$
	KRT8	SRP-dependent cotranslational protein targeting to membrane	GO:0006614	$9.06e^{-08}$

Table 4.7 Partition marker genes and GSEA on top 50 partition markers

Partition markers determined using the “top_marker” function of monocle3. Top 5 markers (determined by marker score) displayed for each partition. Top 50 markers for each partition then used for statistical enrichment analysis using an ordered list through the g:GOSt programme of g:Profiler with significance determined at a 5% FDR. Five most significantly enriched biological process terms displayed.

As marker gene expression had suggested cells in partition 1 represented cells potentially closer to primary microglia I also wanted to see if expression of Alzheimer’s disease (AD) linked genes increased within that specific cluster. I took the list of 9 AD genes, identified in Table 3.7, whose expression was not well captured by any of the monoculture based systems studied in Chapter 3 and compared expression across partitions (Figure 4.13). Many of the genes were not well expressed across any of the cell partitions and may represent AD genes with functions linked to very specific microglial pathways that are still not captured by these model systems. *APOE* was identified as a marker gene for cells within partition 1 and, while not significant, *CLU* also appeared to have increased expression within the same population of cells. Both of these genes are involved in lipid processing pathways and suggests this may be an AD linked pathway that is only possible to study in more complex model systems.

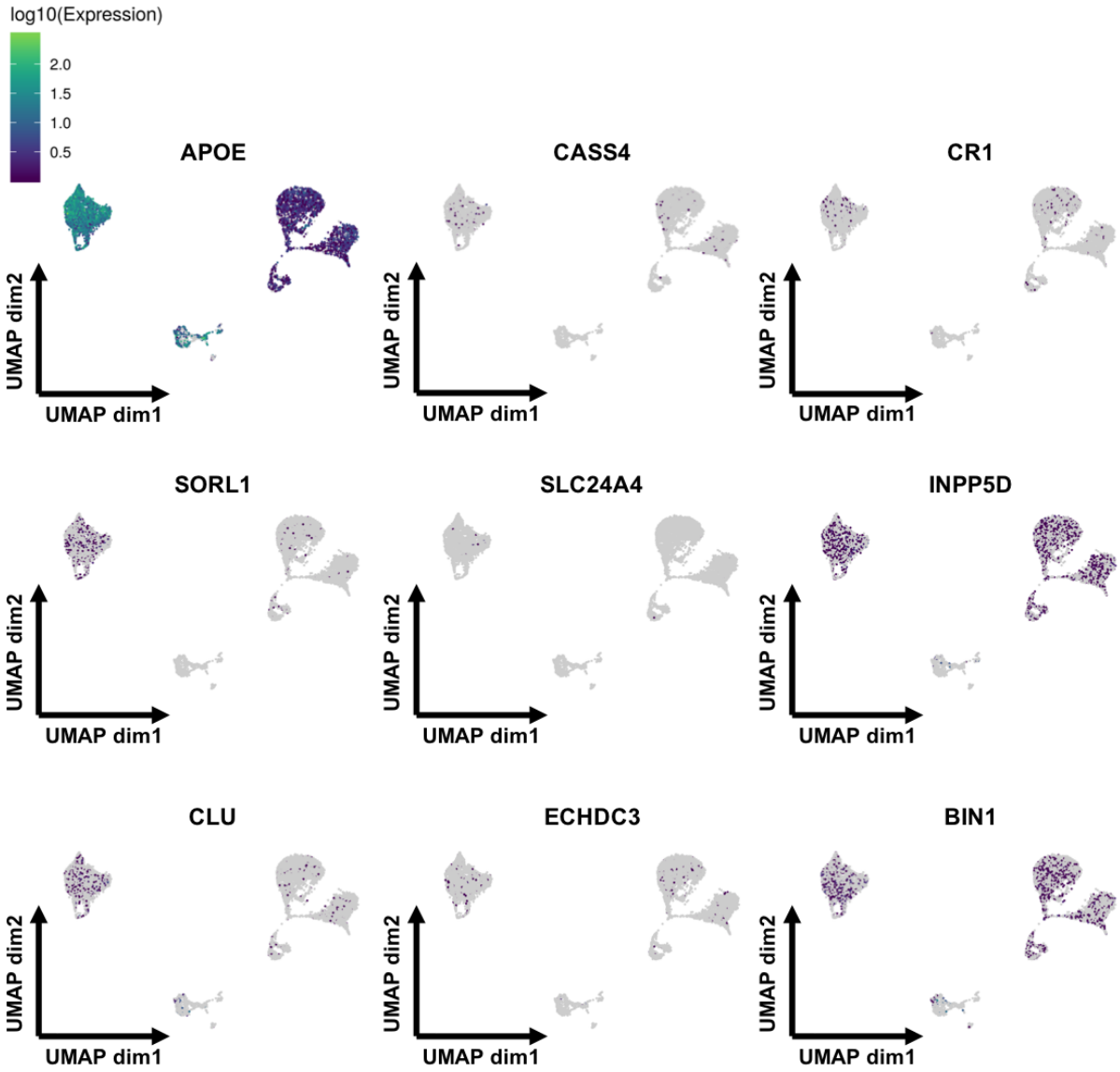


Figure 4.13 UMAP of myeloid cells in Monocle3 coloured by AD gene expression

UMAP analysis following Monocle3 preprocessing. UMAP run using the "reduce_dimension" function of Monocle3. Cells coloured by expression of AD genes not well captured by monoculture model systems, identified in Table 3.7.

4.5 Cell trajectory analysis across model systems

4.5.1 Creation of the trajectory graph

Following identification of partitions and marker genes, I then used the trajectory tool within Monocle3 to determine a cell trajectory graph and order cells along the

pseudotime established from that trajectory (Figure 4.14). Broadly the pseudotime analysis showed cells moving from the monoculture system, through an intermediate step in partition 3 (which includes cells from all culture systems) along to the cells in partition 1 which are predominantly from organoid systems. This further supports the theory that cells from the complex model systems may move along a developmental pathway.

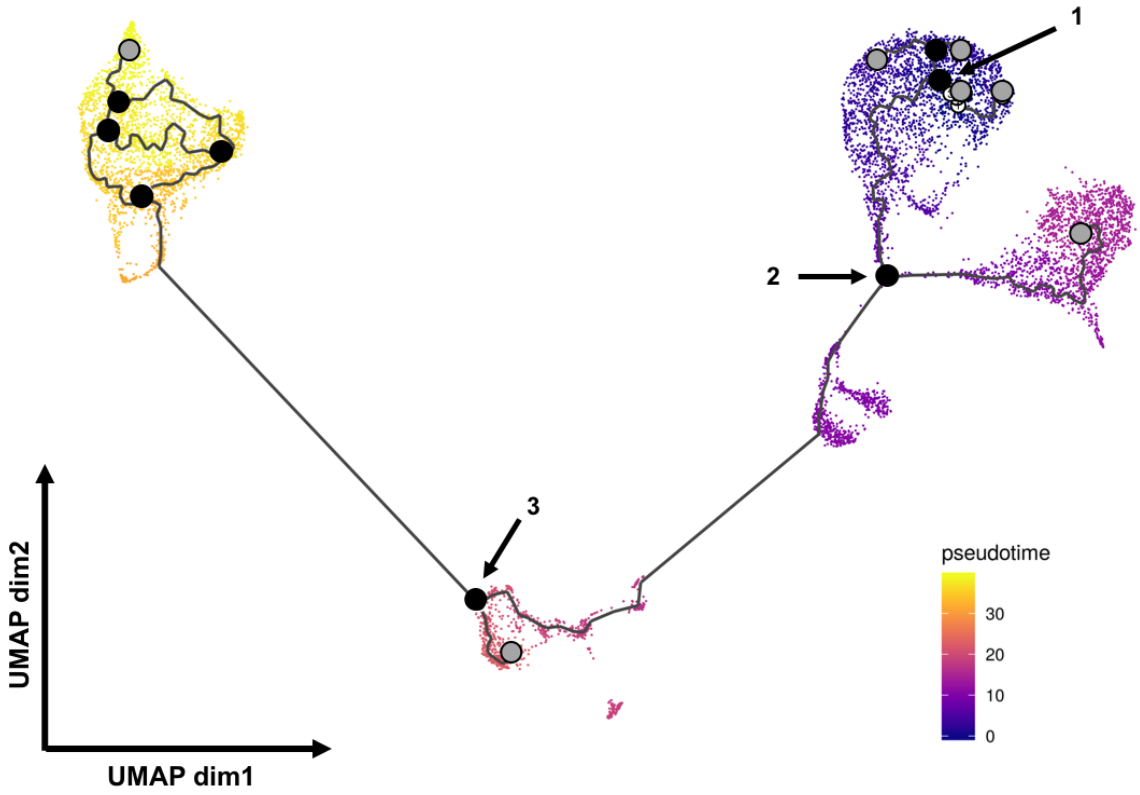


Figure 4.14 UMAP of myeloid cells in Monocle3 coloured by order in pseudotime

UMAP analysis following Monocle3 preprocessing. UMAP run using the "reduce_dimension" function of Monocle3. Cells coloured by order within pseudotime, identified using the "learn_graph" followed by "order_cells" functions in Monocle3. Light grey circles within the pseudotime represent different cell fates while black cells are branch nodes.

Monocle3 also identifies key points of cell differentiations along the trajectory it determines, determining both cell fates (grey circles in Figure 4.14) and branch nodes (black circles). Branch nodes represent points within the developmental trajectory where cells can travel down differing paths. Three major branch nodes are

highlighted in Figure 4.14, each representing a node within the trajectory where cells either move further along the differentiation trajectory or transition towards a cell fate end point (grey circles).

4.5.2 Gene expression changes along pseudotime

As well as generating the standard trajectory graph, I also used the Monocle3 package to identify genes whose expression dynamically changes along the pseudotime. I was able to identify genes, such as *MMP9* and *IL7R,* which had a significant reduction in expression along the pseudotime of differentiation (Figure 4.15). IL7R has recently been linked to the early stages of the differentiation of tissue resident macrophages from fetal precursors in mice[265]. This supports the theory that the monoculture systems represented at the beginning of this pseudotime are more similar to fetal macrophages (as suggested by bulk-RNA sequencing data analysis shown in Figure 3.5 C) and that as the cells move closer towards adult microglia the early differentiation regulators such as IL7R are switched off.

I was also able to identify genes with dynamic expression along the trajectory, such as *PRDX2* and *STMN1* which both increased expression in the intermediate portion of the pseudotime but decreased in the later stages of the trajectory (Figure 4.15). These two genes are potentially interesting as they have both been individually linked to microglia in a more activated state. For instance, single cell sequencing of the adult mouse brain identified a population of cells with increased expression of genes, including *PRDX2*, linked to energy production that could suggest the cells were in a more “immune-alert state”[266]. *STMN1* has also been shown to have increased expression in amoeboid microglial cells, which are associated with increased immune activity, when compared to ramified cells which are linked to more homeostatic functions[267].

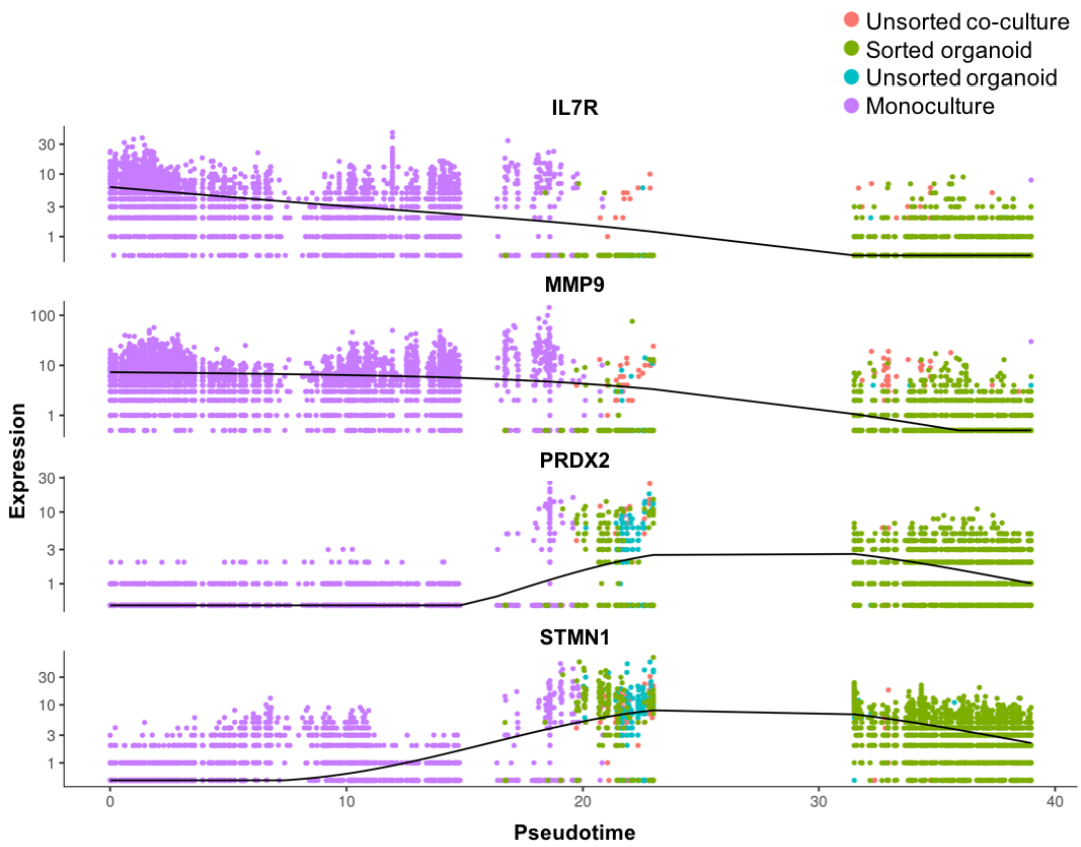


Figure 4.15 Expression of genes along pseudotime

Genes whose expression was significantly linked to a cell's position within the pseudotime trajectory, identified using the "graph_test" function of Monocle3.

The trajectory analysis also highlighted genes whose expression increased along the pseudotime trajectory (Figure 4.16). For instance *APOC1* and *FOS* represented genes that appeared to have a gradual increase along the pseudotime, with *APOC1* continuing to increase at the end stages, while *FOS* expression reached a plateau. *C1QB* was a gene not identified as a partition marker, potentially because the increase in gene expression appeared earlier in the pseudotime analysis and appeared to reach a plateau after the intermediate stage. *NR4A1*, appeared to have a very specific increase in gene expression along the pseudotime with a sharp increase in the first phase of partition 1 towards the end of the trajectory. *NR4A1*, has been suggested to play an important role in the regulation of the activation of microglia in mice and is thought to help maintain the resting state profile of the cells[268].

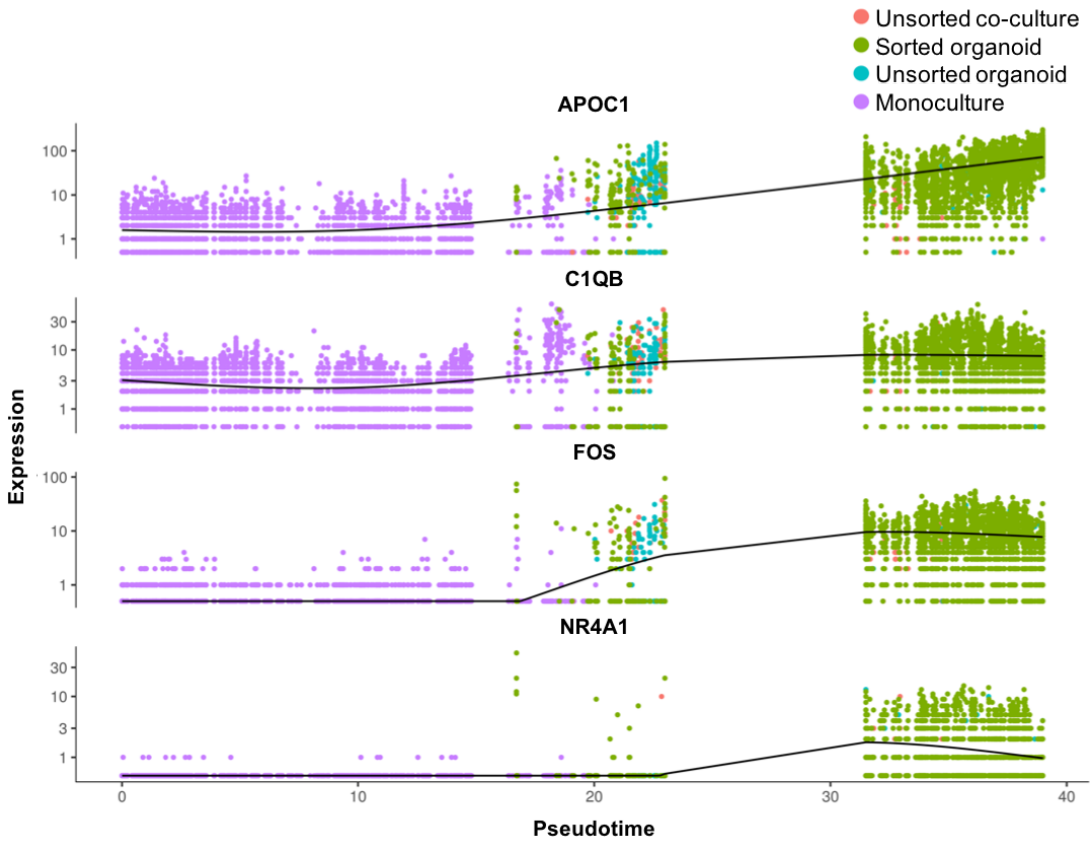


Figure 4.16 Genes with increasing expression along pseudotime

UMAP analysis following Monocle3 preprocessing. UMAP run using the "reduce_dimension" function of Monocle3. Cells coloured by order within pseudotime, identified using the "learn_graph" followed by "order_cells" functions in Monocle3. Light grey circles within the pseudotime represent different cell fates while black cells are branch nodes.

4.6 Discussion

The results in this chapter have suggested that culturing stem cell derived microglia with neuronal cells may move them closer to the primary cell type, with PCA analysis of bulk RNA-sequencing data, using the PMM gene set identified in Chapter 3, showing complex model system samples closer to the primary cells than their monocultured counterparts. Differential expression between monocultured iPSC-derived microglia and those deriving from complex models highlighted an increased gene expression of CNS linked gene sets following culturing with neurons.

This suggested that in an *in-vitro* setting microglia-like cells modified their transcriptome in response to the environment they were in. Although, comparison to primary microglia highlighted specialised neuronal functions, such as oligodendrocyte differentiation and myelination, that were still not captured by the more complex models.

Single cell analysis also allowed for the identification of specific subpopulations of cells that expressed PMM genes. These populations of cells showed increased expression genes enriched for cell migratory functions, suggesting they represent a cell type that are more motile within a dish. Interestly, monocultured microglial cells that showed the most heterogeneity across the single cell populations. Cells from complex model systems were found in two identified partitions where monoculture populations were seen in four partitions. Of the four partitions monoculture cells were found in 3 contained cells only from this culture system, suggesting they represent distinct populations only present in monoculture iPSC-derived microglia. This may mean that as the cells move to a more differentiated state they also converge towards a specific transcriptional phenotype, whereas the monocultured cells are in a more dynamic transcriptional state. The trajectory analysis allowed for individual cells to be ordered along a developmental pseudotime and for the identification of genes whose expression changed dynamically across the trajectory. Evidence from the trajectory analysis also suggested a shift from microglia in a more active state at the intermediate stage, to a more homeostatic cell type towards the end of the trajectory.

However, the single cell dataset only included cells from the cultured systems and the conclusion that the complex models moved cells along a trajectory towards the primary cell type was based on comparisons of differentially expressed genes. Ideally, this experiment would also have included single cell data collected from primary microglia. The data generated from primary microglia in Chapter 2 of this thesis used smartseq2 rather than the 10X technology used here. Batch correction methods have been developed to integrate datasets across differing sequencing technologies, such as within Seurat's updated analysis pipeline[269]. However, this relies on the batch effect not being correlated with biological factors of interest. Combining the primary microglia from Chapter 2 with the model system data

described in this chapter would leave sequencing technology confounded with cell type. As part of the project described in Chapter 2, primary microglia samples were collected and processed through the 10X pipeline. However, the samples were of poor quality and when compared to the smartseq dataset the cells had an activated phenotype that suggested an activation response to the processing pipeline. The samples were therefore not used in analysis as they were determined to not accurately represent cells within the brain.

While partition markers and differential expression analysis highlighted a potential shift towards primary microglia, expression of many AD genes did not increase in the complex model systems. Of the 9 AD linked genes identified in Chapter 3, whose expression was shown to be higher in primary microglia than any of the monoculture model systems, only *APOE* was shown to have a statistically significant increase in expression with organoid derived microglia. This suggests that the other AD linked genes may be involved in highly specialised microglial functions that are not well captured by any model system.

Chapter 5: Discussion

In this thesis I have used multiple RNA-sequencing technologies to generate a transcriptional map of human adult primary microglia and to compare these cells to available *in-vitro* model systems. I have demonstrated that microglia are constantly responding to the CNS environment. In the brain they react to trauma or disease to respond in a disorder-specific manner and it is the complex CNS environment that appears to give rise to the unique transcriptional signature of the primary cells.

5.1 Sequencing primary human microglia

In the second chapter of this thesis, I described the analysis of the largest RNA-sequencing dataset of fresh, adult primary microglial cells to date and demonstrated that microglia display pathology specific activation patterns, particularly following traumatic brain injury. The scale of this study also allowed for comparisons across a variety of clinical factors and demonstrated only a small impact of age or sex on microglial transcriptomes.

Data described in Chapter 2 of this thesis identified potential pathology driven activation patterns in microglial cells through single cell RNA-sequencing. Identification of marker genes for these subpopulations of cells will allow researchers to understand how different microglial phenotypes impact disease outcome or how the activated microglia may play differing roles in microglial responses to trauma or disease. One limitation of this work is that we have not conducted functional validation to verify potential marker genes or to map the functional consequences for each of the populations. Spatial transcriptomics provides a method to combine transcriptional data with *in-situ* hybridization and allows for the identification of cells expressing specific gene markers within a tissue[270,271]. If brain tissue slices could be collected from patients with particular pathologies, such as traumatic brain injury, spatial transcriptomics could be used to not only verify the marker gene sets identified but also see how particular cell populations are distributed within a brain region.

Transcriptomic studies of any cell come with multiple experimental caveats and challenges. The largest challenge is balancing sample access and control of experimental or technical factors that may unknowingly impact microglial transcriptomes. For instance, certain microglial transcriptomes can never be captured using fresh samples. In Chapter 2, we collected “control” patients but it's important to note that these were unlikely to be truly healthy samples. Additionally, tissue samples from certain disease pathologies, such as Alzheimer’s disease, cannot be collected fresh. In order to sequence microglia from these specific cohorts, they must be collected from post-mortem brain tissue. It is not clear how post-mortem delay may impact microglial transcriptomes, especially as data in this thesis has demonstrated that an active CNS environment is vital for the maintenance of the microglial transcriptional signature. While collecting fresh surgery samples removes the potential impact of post-mortem delay on the transcriptome, there are still stages of the single cell sequencing process, such as tissue dissociation, that might introduce transcriptional changes or cell biases. Single-nucleus sequencing may provide a method to overcome some of the technical biases introduced in single cell sequencing, but these technologies are even more costly.

As mentioned above, single cell and single nucleus sequencing technologies are expensive in comparison to bulk RNA-sequencing. Deconvolution techniques allow for the identification of cell types from within bulk data[272]. This means that single cell maps such as the one generated in Chapter 2, could in future be used to deconvolute even larger collections of whole brain tissue samples to identify microglial populations. Increasing sample size within RNA-sequencing studies would allow for more complex genetic association studies, such as subtype specific eQTL studies that could identify specific cell populations that may be involved in disease. Importantly, deconvolution of bulk whole tissue samples also allows the removal of two major steps required for processing of single cell microglial samples, tissue dissociation and cell sorting, which could potentially have an unknown impact on microglial transcriptomes. However, deconvolution does not come without limitations, particularly when identifying rare populations of cells within tissues such as microglia in the brain.

5.2 Modelling primary microglia *in-vitro*

Studies such as the ones described in Chapter 3 and 4 highlight the need for transcriptional comparisons of *in-vitro* model systems to their primary counterpart in order to identify potential limitations of the culture systems. For instance, monoculture iPSC-derived microglia were shown to lack the specialised CNS-linked transcriptional signature seen in primary microglia and, therefore, some of the CNS connected cell functions may also be lacking in these systems. Organoid cultures can provide certain CNS stimuli and single cell trajectory analysis suggested that a population of organoid derived microglia cells moved further along a differentiation pathway. However, gene set enrichment analysis still suggested that certain specialised CNS linked functions were missing in the model systems, such as oligodendrocyte differentiation and myelination. Even more complex brain organoid models are being developed, such as systems with a developing vasculature network[273] or *in-vitro* systems that mimic the BBB[274]. These extensive models may begin to capture more brain functions and lead to further development of specialised cellular phenotypes such as those seen in primary microglia.

However, these complex systems also come with caveats that have to be considered when deciding which model should be used experimentally. They are time consuming to generate, require expensive equipment and reagents and can be more complicated to assay than monoculture systems. Many of these factors mean that brain organoids cannot be used at scale. Large scale genetics studies, such as quantitative trait loci (QTL) experiments, require experimental data from hundreds of samples across varying genetic backgrounds and, therefore, standard organoid differentiation pipelines would not be a feasible experimental tool for these studies. Single cell sequencing has provided a potential way to overcome this issue; it allows for the deconvolution of pools of iPSC lines from within one sample[275] and can attribute single cells back to their original donors. Pooling of iPSC lines allows for the differentiation of multiple donors within one experimental study. This not only reduces the number of required differentiations but also removes some of the batch effects that can arise from comparing different differentiation experiments across different lines.

While iPSC pooling can increase the scalability of organoid differentiations, the protocols remain expensive and complex and so it is important to understand where using these more extensive model systems is necessary. For instance, monoculture iPSC differentiated cells appear to capture some of the transcriptional profile of primary microglia and studies have shown they have comparable behavioural and morphological features of the primary cell type[197–201]. In many cases it may, therefore, be suitable to study certain aspects of microglia function with the more simple monoculture model systems. However, the monoculture models cannot accurately capture how the cells interact with neurons or how they may respond to environmental changes. In these situations more complex models may be required for studying changes in microglial function.

Large scale transcriptional comparisons such as the ones carried out in this thesis could also be used to inform these choices, particularly when studies focus on one specific gene or pathway. Before a model system is chosen, caution should be taken to ensure the gene or pathway of interest is expressed at comparable levels in the model being used to the primary cell type. While this doesn't guarantee comparable responses, it at least provides some evidence that the model system being used has a similar profile to that of primary microglia.

It is also worth noting that all of the studies described in this thesis utilise RNA-sequencing, and therefore, gene expression as a measure of classifying and characterising cell function. However, this does not account for the complicated relationship between gene and protein expression or whether gene/protein expression directly translates to a specific cell function. There are multiple processes following gene transcription that can impact protein expression[276,277] including the translation rate, a protein's half-life and the rate or method by which a protein is transported to its functional location. Variation in any of these stages can lead to a divergence between mRNA levels and protein expression. This is particularly true when cells are transitioning between states and responding to environmental stimuli[277]. This means that the gene expression changes seen in some of the studies described within this thesis may not represent correlated changes in protein levels

and, therefore, functional outputs of the cells. This may be particularly true within the primary microglial single cell dataset where the cells appeared to be dynamically responding to environmental changes.

5.3 Studying microglia in Alzheimer's disease

Microglia are thought to be pathogenic cells in the development and progression of Alzheimer's disease (AD) and therefore each chapter within this thesis has looked at expression of AD linked genes in a variety of contexts. Evidence from the single cell analysis of fresh adult primary microglia in Chapter 1 suggested that microglia respond in a pathology specific manner and studies in both mice and human brain tissue have also demonstrated AD specific activation patterns within microglia[164,166,184]. While some of the genes identified by these studies were expressed across the primary microglia studied in Chapter 1, there was no clear enrichment within a particular cluster which suggested our study did not capture AD specific microglial activation.

It should also be noted that the AD risk gene lists used throughout this thesis were in the most part curated from genes identified in genome wide association studies (GWAS) and these gene lists come with caveats. As described in section 1.5.2 GWAS often identifies a "lead SNP" and associates the SNP to the "nearest gene" despite many of the SNPs falling within the non-coding region of the genome. This may mean that the genes used in this analysis do not represent the true causal risk genes.

Identification of specific gene expression changes that occur in microglia during AD can also highlight genesets and pathways that would need to be mimicked in model systems to accurately capture AD pathology in a dish. Organoid iPSC-based systems have already been used to study AD pathology in a dish, often beginning with iPSC lines containing familial AD mutations to push the cultures towards a disease phenotype[278,279]. With identification of AD specific transcriptional profiles, it may be possible to understand how close *in-vitro* microglia capture the changes seen in

microglia throughout disease progression. One of the major problems with using iPSC differentiated cells to model AD is the maturity of the cultures, age is a major risk factor for neurodegenerative disorders such as AD and capturing that affect in a culture system is challenging as neuronal cultures in particular often more closely represent an immature cell population.

As well as using familial AD mutations within iPSC-derived cultures, it is also possible to engineer late onset AD mutations in iPSC, however there are also caveats with these experiments that should be considered. First, the analysis in this thesis has shown that certain AD risk genes were not expressed at comparable levels in any of the model systems to primary microglia. This means for certain disease genes the effects of risk alleles may not be captured. Even if the expression of the gene of interest is comparable across model systems, the model system chosen is highly dependent on the question and function of interest. For instance, basic microglial functions such as phagocytosis may be well captured by monoculture systems but if the variants are impacting interactions between cell types then more complex models may be required. Unfortunately, for many of the risk alleles associated with AD a clear function has not been identified and so it is difficult to know which model system to choose.

The variants associated with late onset AD risk also tend to have relatively small effect sizes that gradually build throughout life, meaning their effects on individual cell types may be relatively small and not easily seen in cell culture systems. For instance, mutations in the *TREM2* gene in iPSC-derived microglia have been shown to have no impact on cell differentiation, response to stimuli or the ability of microglia to phagocytose compounds[200]. Therefore, it may require the combination of AD risk genes to model AD cell changes in a dish. Polygenic risk scores are statistically based scores that combine genotypes across all risk variants of a disease to predict the likelihood of a person developing a specific trait[280]. Patient-derived cell lines, such as iPSC, could be classified by their polygenic risk scores and differentiated before running functional comparisons across a spectrum of scores. While this would not allow researchers to unpick disease causal mechanisms behind individual genes, it may mean that the subtle impacts of each SNP would combine to generate a more

realistic disease phenotype within cells. Using a spectrum of scores may allow for a greater understanding of how differing levels of disease risk could impact disease progression or development.

5.4 Concluding remarks

In summary, in this thesis I have shown that the microglial transcriptome is constantly reacting to the CNS environment. Initially to develop a unique transcriptional signature and subsequently to respond to disease or trauma. It appears to be signals from the CNS environment that are not well captured by monoculture *in-vitro* model systems. However, more complex systems that culture microglia alongside other neuronal cells and features, such as the BBB, may move the cells closer towards the primary phenotype and the combination of iPSC pooling and single cell sequencing techniques may make large scale studies of these systems more feasible in the future. The potential use of these more complicated and extensive model systems does not always mean they are required. Studies have shown that monoculture *in-vitro* models have certain comparable traits to the primary cell type, such as phagocytosis, whereas other functions of microglia that involve interaction with neuronal signals, like in learning and memory, may only be captured by complex models. It is, therefore, vital to consider the function of interest when identifying an appropriate model system to use for study. This is of particular importance when looking to understand how disease risk genes may modulate cell function. If the model system selected does not accurately capture the linked cellular phenotype then the biological function of a risk gene may be missed.

References

1. Li, Q. & Barres, B. A. Microglia and macrophages in brain homeostasis and disease. *Nat. Rev. Immunol.* **18**, 225–242 (2018).
2. García-Marín, V., García-López, P. & Freire, M. Cajal's contributions to glia research. *Trends Neurosci.* **30**, 479–487 (2007).
3. Tremblay, M.-È., Lecours, C., Samson, L., Sánchez-Zafra, V. & Sierra, A. From the Cajal alumni Achúcarro and Río-Hortega to the rediscovery of never-resting microglia. *Front. Neuroanat.* **9**, 45 (2015).
4. Nimmerjahn, A., Kirchhoff, F. & Helmchen, F. Resting microglial cells are highly dynamic surveillants of brain parenchyma in vivo. *Science* **308**, 1314–1318 (2005).
5. Li, Y., Du, X.-F., Liu, C.-S., Wen, Z.-L. & Du, J.-L. Reciprocal regulation between resting microglial dynamics and neuronal activity in vivo. *Dev. Cell* **23**, 1189–1202 (2012).
6. Muffat, J. *et al.* Efficient derivation of microglia-like cells from human pluripotent stem cells. *Nat. Med.* **22**, 1358–1367 (2016).
7. Ginhoux, F., Lim, S., Hoeffel, G., Low, D. & Huber, T. Origin and differentiation of microglia. *Front. Cell. Neurosci.* **7**, (2013).
8. Fujita, S. & Kitamura, T. Origin of brain macrophages and the nature of the so-called microglia. *Acta Neuropathol. Suppl.* **Suppl 6**, 291–296 (1975).
9. Kitamura, T., Miyake, T. & Fujita, S. Genesis of resting microglia in the gray matter of mouse hippocampus. *J. Comp. Neurol.* **226**, 421–433 (1984).
10. Hao, C., Richardson, A. & Fedoroff, S. Macrophage-like cells originate from neuroepithelium in culture: characterization and properties of the macrophage-like cells. *Int. J. Dev. Neurosci.* **9**, 1–14 (1991).
11. Murabe, Y. & Sano, Y. Morphological studies on neuroglia. VI. Postnatal development of microglial cells. *Cell Tissue Res.* **225**, 469–485 (1982).

12. Akiyama, H. & McGeer, P. L. Brain microglia constitutively express beta-2 integrins. *J. Neuroimmunol.* **30**, 81–93 (1990).
13. McKercher, S. R. *et al.* Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *EMBO J.* **15**, 5647–5658 (1996).
14. Alliot, F., Godin, I. & Pessac, B. Microglia derive from progenitors, originating from the yolk sac, and which proliferate in the brain. *Brain Res. Dev. Brain Res.* **117**, 145–152 (1999).
15. Monier, A. *et al.* Entry and distribution of microglial cells in human embryonic and fetal cerebral cortex. *J. Neuropathol. Exp. Neurol.* **66**, 372–382 (2007).
16. De Kleer, I., Willems, F., Lambrecht, B. & Goriely, S. Ontogeny of myeloid cells. *Front. Immunol.* **5**, 423 (2014).
17. Ginhoux, F. *et al.* Fate Mapping Analysis Reveals That Adult Microglia Derive from Primitive Macrophages. *Science* **330**, 841–845 (2010).
18. Gomez Perdiguero, E. *et al.* Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors. *Nature* **518**, 547–551 (2015).
19. Schulz, C. *et al.* A lineage of myeloid cells independent of Myb and hematopoietic stem cells. *Science* **336**, 86–90 (2012).
20. Hoeffel, G. *et al.* C-Myb(+) erythro-myeloid progenitor-derived fetal monocytes give rise to adult tissue-resident macrophages. *Immunity* **42**, 665–678 (2015).
21. Kierdorf, K. *et al.* Microglia emerge from erythromyeloid precursors via Pu.1- and Irf8-dependent pathways. *Nat. Neurosci.* **16**, 273–280 (2013).
22. Louveau, A., Harris, T. H. & Kipnis, J. Revisiting the Mechanisms of CNS Immune Privilege. *Trends Immunol.* **36**, 569–577 (2015).
23. Beers, D. R. *et al.* Wild-type microglia extend survival in PU.1 knockout mice with familial amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16021–16026 (2006).
24. Kamran, P. *et al.* Parabiosis in mice: a detailed protocol. *J. Vis. Exp.* (2013) doi:10.3791/50556.

25. Ajami, B., Bennett, J. L., Krieger, C., Tetzlaff, W. & Rossi, F. M. V. Local self-renewal can sustain CNS microglia maintenance and function throughout adult life. *Nat. Neurosci.* **10**, 1538–1543 (2007).
26. Hashimoto, D. *et al.* Tissue-resident macrophages self-maintain locally throughout adult life with minimal contribution from circulating monocytes. *Immunity* **38**, 792–804 (2013).
27. Huang, Y. *et al.* Repopulated microglia are solely derived from the proliferation of residual microglia after acute depletion. *Nat. Neurosci.* **21**, 530–540 (2018).
28. Varvel, N. H. *et al.* Microglial repopulation model reveals a robust homeostatic process for replacing CNS myeloid cells. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18150–18155 (2012).
29. Elmore, M. R. P. *et al.* Colony-stimulating factor 1 receptor signaling is necessary for microglia viability, unmasking a microglia progenitor cell in the adult brain. *Neuron* **82**, 380–397 (2014).
30. Kierdorf, K. & Prinz, M. Microglia in steady state. *J. Clin. Invest.* **127**, 3201–3209 (2017).
31. Salter, M. W. & Stevens, B. Microglia emerge as central players in brain disease. *Nat. Med.* **23**, 1018–1027 (2017).
32. Tay, T. L., Savage, J. C., Hui, C. W., Bisht, K. & Tremblay, M.-È. Microglia across the lifespan: from origin to function in brain development, plasticity and cognition. *J. Physiol.* **595**, 1929–1945 (2017).
33. Yin, J., Valin, K. L., Dixon, M. L. & Leavenworth, J. W. The Role of Microglia and Macrophages in CNS Homeostasis, Autoimmunity, and Cancer. *J Immunol Res* **2017**, 5150678 (2017).
34. Oosterhof, N. *et al.* Homozygous Mutations in CSF1R Cause a Pediatric-Onset Leukoencephalopathy and Can Result in Congenital Absence of Microglia. *Am. J. Hum. Genet.* **104**, 936–947 (2019).
35. Cunningham, C. L., Martínez-Cerdeño, V. & Noctor, S. C. Microglia regulate the number of neural precursor cells in the developing cerebral cortex. *J. Neurosci.* **33**, 4216–4233

(2013).

36. Tronnes, A. A. *et al.* Effects of Lipopolysaccharide and Progesterone Exposures on Embryonic Cerebral Cortex Development in Mice. *Reprod. Sci.* **23**, 771–778 (2016).
37. Ueno, M. *et al.* Layer V cortical neurons require microglial support for survival during postnatal development. *Nat. Neurosci.* **16**, 543–551 (2013).
38. Frade, J. M. & Barde, Y. A. Microglia-derived nerve growth factor causes cell death in the developing retina. *Neuron* **20**, 35–41 (1998).
39. Marín-Teva, J. L. *et al.* Microglia promote the death of developing Purkinje cells. *Neuron* **41**, 535–547 (2004).
40. Wakselman, S. *et al.* Developmental neuronal death in hippocampus requires the microglial CD11b integrin and DAP12 immunoreceptor. *J. Neurosci.* **28**, 8138–8143 (2008).
41. Rymo, S. F. *et al.* A two-way communication between microglial cells and angiogenic sprouts regulates angiogenesis in aortic ring cultures. *PLoS One* **6**, e15846 (2011).
42. Schafer, D. P. *et al.* Microglia sculpt postnatal neural circuits in an activity and complement-dependent manner. *Neuron* **74**, 691–705 (2012).
43. Paolicelli, R. C. *et al.* Synaptic pruning by microglia is necessary for normal brain development. *Science* **333**, 1456–1458 (2011).
44. Stevens, B. *et al.* The classical complement cascade mediates CNS synapse elimination. *Cell* **131**, 1164–1178 (2007).
45. Lui, H. *et al.* Progranulin Deficiency Promotes Circuit-Specific Synaptic Pruning by Microglia via Complement Activation. *Cell* **165**, 921–935 (2016).
46. Paolicelli, R. C., Bisht, K. & Tremblay, M.-È. Fractalkine regulation of microglial physiology and consequences on the brain and behavior. *Front. Cell. Neurosci.* **8**, 129 (2014).
47. Rogers, J. T. *et al.* CX3CR1 deficiency leads to impairment of hippocampal cognitive function and synaptic plasticity. *J. Neurosci.* **31**, 16241–16250 (2011).

48. George, J., Cunha, R. A., Mulle, C. & Amédée, T. Microglia-derived purines modulate mossy fibre synaptic transmission and plasticity through P2X4 and A1 receptors. *Eur. J. Neurosci.* **43**, 1366–1378 (2016).

49. Sipe, G. O. *et al.* Microglial P2Y12 is necessary for synaptic plasticity in mouse visual cortex. *Nat. Commun.* **7**, 10905 (2016).

50. Baufeld, C., Osterloh, A., Prokop, S., Miller, K. R. & Heppner, F. L. High-fat diet-induced brain region-specific phenotypic spectrum of CNS resident microglia. *Acta Neuropathol.* **132**, 361–375 (2016).

51. Erny, D. *et al.* Host microbiota constantly control maturation and function of microglia in the CNS. *Nat. Neurosci.* **18**, 965–977 (2015).

52. Hammond, T. R., Robinton, D. & Stevens, B. Microglia and the Brain: Complementary Partners in Development and Disease. *Annu. Rev. Cell Dev. Biol.* **34**, 523–544 (2018).

53. Menon, D. K., Schwab, K., Wright, D. W., Maas, A. I. & Demographics and Clinical Assessment Working Group of the International and Interagency Initiative toward Common Data Elements for Research on Traumatic Brain Injury and Psychological Health. Position statement: definition of traumatic brain injury. *Arch. Phys. Med. Rehabil.* **91**, 1637–1640 (2010).

54. Davalos, D. *et al.* ATP mediates rapid microglial response to local brain injury in vivo. *Nat. Neurosci.* **8**, 752–758 (2005).

55. Roth, T. L. *et al.* Transcranial amelioration of inflammation and cell death after brain injury. *Nature* **505**, 223–228 (2014).

56. Lou, N. *et al.* Purinergic receptor P2RY12-dependent microglial closure of the injured blood-brain barrier. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 1074–1079 (2016).

57. Hsieh, C. L. *et al.* Traumatic brain injury induces macrophage subsets in the brain. *Eur. J. Immunol.* **43**, 2010–2022 (2013).

58. Smith, C. *et al.* The neuroinflammatory response in humans after traumatic brain injury. *Neuropathol. Appl. Neurobiol.* **39**, 654–666 (2013).

59. Ziebell, J. M. & Morganti-Kossmann, M. C. Involvement of pro- and anti-inflammatory cytokines and chemokines in the pathophysiology of traumatic brain injury. *Neurotherapeutics* **7**, 22–30 (2010).

60. Faden, A. I. & Loane, D. J. Chronic neurodegeneration after traumatic brain injury: Alzheimer disease, chronic traumatic encephalopathy, or persistent neuroinflammation? *Neurotherapeutics* **12**, 143–150 (2015).

61. Loane, D. J. & Kumar, A. Microglia in the TBI brain: The good, the bad, and the dysregulated. *Exp. Neurol.* **275 Pt 3**, 316–327 (2016).

62. Donat, C. K., Scott, G., Gentleman, S. M. & Sastre, M. Microglial Activation in Traumatic Brain Injury. *Front. Aging Neurosci.* **9**, 208 (2017).

63. Graham, N. S. & Sharp, D. J. Understanding neurodegeneration after traumatic brain injury: from mechanisms to clinical trials in dementia. *J. Neurol. Neurosurg. Psychiatry* **90**, 1221–1233 (2019).

64. Li, Y. *et al.* Head Injury as a Risk Factor for Dementia and Alzheimer's Disease: A Systematic Review and Meta-Analysis of 32 Observational Studies. *PLoS One* **12**, e0169650 (2017).

65. Bramlett, H. M., Kraydieh, S., Green, E. J. & Dietrich, W. D. Temporal and regional patterns of axonal damage following traumatic brain injury: a beta-amyloid precursor protein immunocytochemical study in rats. *J. Neuropathol. Exp. Neurol.* **56**, 1132–1141 (1997).

66. Franz, G. *et al.* Amyloid beta 1-42 and tau in cerebrospinal fluid after severe traumatic brain injury. *Neurology* **60**, 1457–1461 (2003).

67. Goldman, S. M. *et al.* Head injury, α-synuclein Rep1, and Parkinson's disease. *Ann. Neurol.* **71**, 40–48 (2012).

68. McKee, A. C. *et al.* The first NINDS/NIBIB consensus meeting to define neuropathological criteria for the diagnosis of chronic traumatic encephalopathy. *Acta Neuropathol.* **131**, 75–86 (2016).

69. Xiong, Y., Mahmood, A. & Chopp, M. Current understanding of neuroinflammation after traumatic brain injury and cell-based therapeutic opportunities. *Chin. J. Traumatol.* **21**, 137–151 (2018).

70. Johnson, V. E. *et al.* Inflammation and white matter degeneration persist for years after a single traumatic brain injury. *Brain* **136**, 28–42 (2013).

71. Scott, G. *et al.* Minocycline reduces chronic microglial activation after brain trauma but increases neurodegeneration. *Brain* **141**, 459–471 (2018).

72. Li, L. *et al.* The Association Between Apolipoprotein E and Functional Outcome After Traumatic Brain Injury: A Meta-Analysis. *Medicine* **94**, e2028 (2015).

73. Lucchinetti, C. *et al.* Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. *Ann. Neurol.* **47**, 707–717 (2000).

74. Singh, S. *et al.* Microglial nodules in early multiple sclerosis white matter are associated with degenerating axons. *Acta Neuropathol.* **125**, 595–608 (2013).

75. Ohl, K., Tenbrock, K. & Kipp, M. Oxidative stress in multiple sclerosis: Central and peripheral mode of action. *Exp. Neurol.* **277**, 58–67 (2016).

76. Gray, E., Thomas, T. L., Betmouni, S., Scolding, N. & Love, S. Elevated myeloperoxidase activity in white matter in multiple sclerosis. *Neurosci. Lett.* **444**, 195–198 (2008).

77. Fischer, M. T. *et al.* NADPH oxidase expression in active multiple sclerosis lesions in relation to oxidative tissue damage and mitochondrial injury. *Brain* **135**, 886–899 (2012).

78. Centonze, D. *et al.* Inflammation triggers synaptic alteration and degeneration in experimental autoimmune encephalomyelitis. *J. Neurosci.* **29**, 3442–3452 (2009).

79. Lampron, A. *et al.* Inefficient clearance of myelin debris by microglia impairs remyelinating processes. *J. Exp. Med.* **212**, 481–495 (2015).

80. Miron, V. E. *et al.* M2 microglia and macrophages drive oligodendrocyte differentiation during CNS remyelination. *Nat. Neurosci.* **16**, 1211–1218 (2013).

81. Tetreault, N. A. *et al.* Microglia in the cerebral cortex in autism. *J. Autism Dev. Disord.*

42, 2569–2584 (2012).

82. Vargas, D. L., Nascimbene, C., Krishnan, C., Zimmerman, A. W. & Pardo, C. A. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann. Neurol.* **57**, 67–81 (2005).
83. Morgan, J. T. *et al.* Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism. *Biol. Psychiatry* **68**, 368–376 (2010).
84. Suzuki, K. *et al.* Microglial activation in young adults with autism spectrum disorder. *JAMA Psychiatry* **70**, 49–58 (2013).
85. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5**, 5748 (2014).
86. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
87. Feinberg, I. Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? *J. Psychiatr. Res.* **17**, 319–334 (1982).
88. Penzes, P., Cahill, M. E., Jones, K. A., VanLeeuwen, J.-E. & Woolfrey, K. M. Dendritic spine pathology in neuropsychiatric disorders. *Nat. Neurosci.* **14**, 285–293 (2011).
89. Gu, N. *et al.* Spinal Microgliosis Due to Resident Microglial Proliferation Is Required for Pain Hypersensitivity after Peripheral Nerve Injury. *Cell Rep.* **16**, 605–614 (2016).
90. Okubo, M. *et al.* Macrophage-Colony Stimulating Factor Derived from Injured Primary Afferent Induces Proliferation of Spinal Microglia and Neuropathic Pain in Rats. *PLoS One* **11**, e0153375 (2016).
91. Gu, N. *et al.* Microglial P2Y12 receptors regulate microglial activation and surveillance during neuropathic pain. *Brain Behav. Immun.* **55**, 82–92 (2016).
92. Guan, Z. *et al.* Injured sensory neuron-derived CSF1 induces microglial proliferation and DAP12-dependent pain. *Nat. Neurosci.* **19**, 94–101 (2016).
93. Stelzmann, R. A., Norman Schnitzlein, H. & Reed Murtagh, F. An english translation of

alzheimer's 1907 paper, ?�ber eine eigenartige erkankung der hirnrinde? *Clin. Anat.* **8**, 429–431 (1995).

94. Hippius, H. & Neundörfer, G. The discovery of Alzheimer's disease. *Dialogues Clin. Neurosci.* **5**, 101–108 (2003).
95. Barber, R. C. The genetics of Alzheimer's disease. *Scientifica* **2012**, 246210 (2012).
96. Davies, P. & Maloney, A. J. Selective loss of central cholinergic neurons in Alzheimer's disease. *Lancet* **2**, 1403 (1976).
97. Francis, P. T., Palmer, A. M., Snape, M. & Wilcock, G. K. The cholinergic hypothesis of Alzheimer's disease: a review of progress. *J. Neurol. Neurosurg. Psychiatry* **66**, 137–147 (1999).
98. Liu, P.-P., Xie, Y., Meng, X.-Y. & Kang, J.-S. History and progress of hypotheses and clinical trials for Alzheimer's disease. *Signal Transduct Target Ther* **4**, 29 (2019).
99. Hardy, J. A. & Higgins, G. A. Alzheimer's disease: the amyloid cascade hypothesis. *Science* **256**, 184–185 (1992).
100. Goate, A. *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**, 704–706 (1991).
101. Ryan, N. S. *et al.* Clinical phenotype and genetic associations in autosomal dominant familial Alzheimer's disease: a case series. *Lancet Neurol.* **15**, 1326–1335 (2016).
102. Sherrington, R. *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754–760 (1995).
103. Levy-Lahad, E. *et al.* Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* **269**, 973–977 (1995).
104. Ricciarelli, R. & Fedele, E. The Amyloid Cascade Hypothesis in Alzheimer's Disease: It's Time to Change Our Mind. *Curr. Neuropharmacol.* **15**, 926–935 (2017).
105. Mucke, L. & Selkoe, D. J. Neurotoxicity of amyloid β-protein: synaptic and network dysfunction. *Cold Spring Harb. Perspect. Med.* **2**, a006338 (2012).
106. Ferreira, S. T., Lourenco, M. V., Oliveira, M. M. & De Felice, F. G. Soluble amyloid-β

oligomers as synaptotoxins leading to cognitive impairment in Alzheimer's disease. *Front. Cell. Neurosci.* **9**, 191 (2015).

107. Lorenzo, A. & Yankner, B. A. Beta-amyloid neurotoxicity requires fibril formation and is inhibited by congo red. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 12243–12247 (1994).

108. Tomiyama, T. *et al.* A mouse model of amyloid beta oligomers: their contribution to synaptic alteration, abnormal tau phosphorylation, glial activation, and neuronal loss in vivo. *J. Neurosci.* **30**, 4845–4856 (2010).

109. McLaurin, J. *et al.* Cyclohexanehexol inhibitors of Abeta aggregation prevent and reverse Alzheimer phenotype in a mouse model. *Nat. Med.* **12**, 801–808 (2006).

110. Li, C., Ebrahimi, A. & Schluesener, H. Drug pipeline in neurodegeneration based on transgenic mice models of Alzheimer's disease. *Ageing Res. Rev.* **12**, 116–140 (2013).

111. Reitz, C. Alzheimer's disease and the amyloid cascade hypothesis: a critical review. *Int. J. Alzheimers. Dis.* **2012**, 369808 (2012).

112. Armstrong, R. A. A critical analysis of the 'amyloid cascade hypothesis'. *Folia Neuropathol.* **52**, 211–225 (2014).

113. Mullard, A. Anti-amyloid failures stack up as Alzheimer antibody flops. *Nat. Rev. Drug Discov.* (2019) doi:10.1038/d41573-019-00064-1.

114. Miller, B. W., Willett, K. C. & Desilets, A. R. Rosiglitazone and pioglitazone for the treatment of Alzheimer's disease. *Ann. Pharmacother.* **45**, 1416–1424 (2011).

115. Doody, R. S. *et al.* A phase 3 trial of semagacestat for treatment of Alzheimer's disease. *N. Engl. J. Med.* **369**, 341–350 (2013).

116. Barão, S., Moechars, D., Lichtenthaler, S. F. & De Strooper, B. BACE1 Physiological Functions May Limit Its Use as Therapeutic Target for Alzheimer's Disease. *Trends Neurosci.* **39**, 158–169 (2016).

117. Gilman, S. *et al.* Clinical effects of Abeta immunization (AN1792) in patients with AD in an interrupted trial. *Neurology* **64**, 1553–1562 (2005).

118. Salloway, S. *et al.* Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's

disease. *N. Engl. J. Med.* **370**, 322–333 (2014).

119. Rinne, J. O. *et al.* 11C-PiB PET assessment of change in fibrillar amyloid-beta load in patients with Alzheimer's disease treated with bapineuzumab: a phase 2, double-blind, placebo-controlled, ascending-dose study. *Lancet Neurol.* **9**, 363–372 (2010).

120. Mehta, D., Jackson, R., Paul, G., Shi, J. & Sabbagh, M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert Opin. Investig. Drugs* **26**, 735–739 (2017).

121. Abbott, A. Fresh push for 'failed' Alzheimer's drug. *Nature* (2019) doi:10.1038/d41586-019-03261-5.

122. Kueper, J. K., Speechley, M. & Montero-Odasso, M. The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review. *J. Alzheimers. Dis.* **63**, 423–444 (2018).

123. Saunders, A. M. *et al.* Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467–1472 (1993).

124. Strittmatter, W. J. *et al.* Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 1977–1981 (1993).

125. Corder, E. H. *et al.* Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat. Genet.* **7**, 180–184 (1994).

126. Rogaeva, E. *et al.* The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. *Nat. Genet.* **39**, 168–177 (2007).

127. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).

128. McGeer, P. L., Akiyama, H., Itagaki, S. & McGeer, E. G. Immune system response in Alzheimer's disease. *Can. J. Neurol. Sci.* **16**, 516–527 (1989).

129. Itagaki, S., McGeer, P. L., Akiyama, H., Zhu, S. & Selkoe, D. Relationship of microglia

and astrocytes to amyloid deposits of Alzheimer disease. *J. Neuroimmunol.* **24**, 173–182 (1989).

130. Sims, R. *et al.* Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet.* **49**, 1373–1384 (2017).

131. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* **41**, 1088–1093 (2009).

132. Jun, G. *et al.* Meta-analysis confirms CR1, CLU, and PICALM as alzheimer disease risk loci and reveals interactions with APOE genotypes. *Arch. Neurol.* **67**, 1473–1484 (2010).

133. Seshadri, S. *et al.* Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA* **303**, 1832–1840 (2010).

134. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

135. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).

136. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).

137. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

138. Lambert, J.-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094–1099 (2009).

139. Hollingworth, P. *et al.* Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat. Genet.* **43**, 429–435 (2011).

140. Bertram, L. *et al.* Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. *Am. J. Hum. Genet.* **83**, 623–632 (2008).

141. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to

Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).

142. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).

143. Calderon, D. *et al.* Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *Am. J. Hum. Genet.* **101**, 686–699 (2017).

144. Tansey, K. E., Cameron, D. & Hill, M. J. Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks. *Genome Med.* **10**, 14 (2018).

145. Bolmont, T. *et al.* Dynamics of the microglial/amyloid interaction indicate a role in plaque maintenance. *J. Neurosci.* **28**, 4283–4292 (2008).

146. Condello, C., Yuan, P., Schain, A. & Grutzendler, J. Microglia constitute a barrier that prevents neurotoxic protofibrillar Aβ42 hotspots around plaques. *Nat. Commun.* **6**, 6176 (2015).

147. Hickman, S. E., Allison, E. K. & El Khoury, J. Microglial dysfunction and defective beta-amyloid clearance pathways in aging Alzheimer's disease mice. *J. Neurosci.* **28**, 8354–8360 (2008).

148. Krabbe, G. *et al.* Functional impairment of microglia coincides with Beta-amyloid deposition in mice with Alzheimer-like pathology. *PLoS One* **8**, e60921 (2013).

149. Spangenberg, E. E. *et al.* Eliminating microglia in Alzheimer's mice prevents neuronal loss without modulating amyloid-β pathology. *Brain* **139**, 1265–1281 (2016).

150. Hong, S. *et al.* Complement and microglia mediate early synapse loss in Alzheimer mouse models. *Science* **352**, 712–716 (2016).

151. Sosna, J. *et al.* Early long-term administration of the CSF1R inhibitor PLX3397 ablates microglia and reduces accumulation of intraneuronal amyloid, neuritic plaque deposition and pre-fibrillar oligomers in 5XFAD mouse model of Alzheimer's disease. *Mol. Neurodegener.* **13**, 11 (2018).

152. Spangenberg, E. *et al.* Sustained microglial depletion with CSF1R inhibitor impairs parenchymal plaque development in an Alzheimer's disease model. *Nat. Commun.* **10**, 3758 (2019).

153. Asai, H. *et al.* Depletion of microglia and inhibition of exosome synthesis halt tau propagation. *Nat. Neurosci.* **18**, 1584–1593 (2015).

154. Gratuze, M., Leyns, C. E. G. & Holtzman, D. M. New insights into the role of TREM2 in Alzheimer's disease. *Mol. Neurodegener.* **13**, 66 (2018).

155. Zheng, H. *et al.* TREM2 in Alzheimer's Disease: Microglial Survival and Energy Metabolism. *Front. Aging Neurosci.* **10**, 395 (2018).

156. Colonna, M. & Wang, Y. TREM2 variants: new keys to decipher Alzheimer disease pathogenesis. *Nat. Rev. Neurosci.* **17**, 201–207 (2016).

157. Heslegrave, A. *et al.* Increased cerebrospinal fluid soluble TREM2 concentration in Alzheimer's disease. *Mol. Neurodegener.* **11**, 3 (2016).

158. Zhong, L. *et al.* Soluble TREM2 induces inflammatory responses and enhances microglial survival. *J. Exp. Med.* **214**, 597–607 (2017).

159. Zhong, L. *et al.* Soluble TREM2 ameliorates pathological phenotypes by modulating microglial functions in an Alzheimer's disease model. *Nat. Commun.* **10**, 1365 (2019).

160. Yeh, F. L., Wang, Y., Tom, I., Gonzalez, L. C. & Sheng, M. TREM2 Binds to Apolipoproteins, Including APOE and CLU/APOJ, and Thereby Facilitates Uptake of Amyloid-Beta by Microglia. *Neuron* **91**, 328–340 (2016).

161. Shi, Y. & Holtzman, D. M. Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight. *Nat. Rev. Immunol.* **18**, 759–772 (2018).

162. Griciuc, A. *et al.* TREM2 Acts Downstream of CD33 in Modulating Microglial Pathology in Alzheimer's Disease. *Neuron* **103**, 820–835.e7 (2019).

163. Deming, Y. *et al.* The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Sci. Transl. Med.* **11**, (2019).

164. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development

of Alzheimer's Disease. *Cell* **169**, 1276–1290.e17 (2017).

165. Krasemann, S. *et al.* The TREM2-APOE Pathway Drives the Transcriptional Phenotype of Dysfunctional Microglia in Neurodegenerative Diseases. *Immunity* **47**, 566–581.e9 (2017).

166. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* (2019) doi:10.1038/s41586-019-1195-2.

167. Smith, A. M. & Dragunow, M. The human side of microglia. *Trends Neurosci.* **37**, 125–135 (2014).

168. Watkins, L. R. & Hutchinson, M. R. A concern on comparing 'apples' and 'oranges' when differences between microglia used in human and rodent studies go far, far beyond simply species: comment on Smith and Dragunow. *Trends in neurosciences* vol. 37 189–190 (2014).

169. Smith, A. M. & Dragunow, M. Response to Watkins and Hutchinson. *Trends Neurosci.* **37**, 190 (2014).

170. Streit, W. J., Braak, H., Xue, Q.-S. & Bechmann, I. Dystrophic (senescent) rather than activated microglial cells are associated with tau pathology and likely precede neurodegeneration in Alzheimer's disease. *Acta Neuropathol.* **118**, 475–485 (2009).

171. Gosselin, D. *et al.* An environment-dependent transcriptional network specifies human microglia identity. *Science* **356**, (2017).

172. Olah, M. *et al.* An optimized protocol for the acute isolation of human microglia from autopsy brain samples. *Glia* **60**, 96–111 (2012).

173. Rustenhoven, J. *et al.* Isolation of highly enriched primary human microglia for functional studies. *Sci. Rep.* **6**, 19371 (2016).

174. Mizee, M. R., Poel, M. van der & Huitinga, I. Purification of cells from fresh human brain tissue: primary human glial cells. *Handb. Clin. Neurol.* **150**, 273–283 (2018).

175. Mizee, M. R. *et al.* Isolation of primary microglia from the human post-mortem brain: effects of ante- and post-mortem variables. *Acta Neuropathol Commun* **5**, 16 (2017).

176. Richardson, G. M., Lannigan, J. & Macara, I. G. Does FACS perturb gene expression? *Cytometry A* **87**, 166–175 (2015).

177. Gautier, E. L. *et al.* Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat. Immunol.* **13**, 1118–1128 (2012).

178. Butovsky, O. *et al.* Identification of a unique TGF-[beta]-dependent molecular and functional signature in microglia. *Nat. Neurosci.* **17**, 131–143 (2014).

179. Olah, M. *et al.* A transcriptomic atlas of aged human microglia. *Nat. Commun.* **9**, 539 (2018).

180. van der Poel, M. *et al.* Transcriptional profiling of human microglia reveals grey-white matter heterogeneity and multiple sclerosis-associated changes. *Nat. Commun.* **10**, 1139 (2019).

181. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 96 (2018).

182. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* **10**, 317 (2019).

183. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).

184. Del-Aguila, J. L. *et al.* A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain. *Alzheimers. Res. Ther.* **11**, 71 (2019).

185. Olah, M. *et al.* A single cell-based atlas of human microglial states reveals associations with neurological disorders and histopathological features of the aging brain. *bioRxiv* 343780 (2018) doi:10.1101/343780.

186. Masuda, T. *et al.* Spatial and temporal heterogeneity of mouse and human microglia at single-cell resolution. *Nature* **566**, 388–392 (2019).

187. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).

188. Maherali, N. *et al.* Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1**, 55–70 (2007).

189. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

190. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).

191. Sterneckert, J. L., Reinhardt, P. & Schöler, H. R. Investigating human disease using stem cell models. *Nat. Rev. Genet.* **15**, 625–639 (2014).

192. Karlsson, K. R. *et al.* Homogeneous monocytes and macrophages from human embryonic stem cells following coculture-free differentiation in M-CSF and IL-3. *Exp. Hematol.* **36**, 1167–1175 (2008).

193. van Wilgenburg, B., Browne, C., Vowles, J. & Cowley, S. A. Efficient, long term production of monocyte-derived macrophages from human pluripotent stem cells under partly-defined and fully-defined conditions. *PLoS One* **8**, e71098 (2013).

194. Alasoo, K. *et al.* Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription. *Sci. Rep.* **5**, 12524 (2015).

195. Zhang, H. *et al.* Functional analysis and transcriptomic profiling of iPSC-derived macrophages and their application in modeling Mendelian disease. *Circ. Res.* **117**, 17–28 (2015).

196. Itskovitz-Eldor, J. *et al.* Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. *Mol. Med.* **6**, 88–95 (2000).

197. Muffat, J. *et al.* Efficient derivation of microglia-like cells from human pluripotent stem cells. *Nat. Med.* **22**, 1358–1367 (2016).

198. Abud, E. M. *et al.* iPSC-Derived Human Microglia-like Cells to Study Neurological Diseases. *Neuron* **94**, 278–293.e9 (2017).

199. Douvaras, P. *et al.* Directed Differentiation of Human Pluripotent Stem Cells to Microglia.

Stem Cell Reports **8**, 1516–1524 (2017).

200. Brownjohn, P. W. *et al.* Functional Studies of Missense TREM2 Mutations in Human Stem Cell-Derived Microglia. *Stem Cell Reports* **10**, 1294–1307 (2018).

201. McQuade, A. *et al.* Development and validation of a simplified method to generate human microglia from pluripotent stem cells. *Mol. Neurodegener.* **13**, 67 (2018).

202. Haenseler, W. *et al.* A Highly Efficient Human Pluripotent Stem Cell Microglia Model Displays a Neuronal-Co-culture-Specific Expression Profile and Inflammatory Response. *Stem Cell Reports* **8**, 1727–1742 (2017).

203. Qian, X. *et al.* Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. *Cell* **165**, 1238–1254 (2016).

204. Ormel, P. R. *et al.* Microglia innately develop within cerebral organoids. *Nat. Commun.* **9**, 4167 (2018).

205. Park, J. *et al.* A 3D human triculture system modeling neurodegeneration and neuroinflammation in Alzheimer's disease. *Nat. Neurosci.* **21**, 941–951 (2018).

206. Qian, X. *et al.* Generation of human brain region-specific organoids using a miniaturized spinning bioreactor. *Nat. Protoc.* **13**, 565–580 (2018).

207. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).

208. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

209. Arora, S., Pattwell, S. S., Holland, E. C. & Bolouri, H. Uncertainty in RNA-seq gene expression data. *bioRxiv* 445601 (2018) doi:10.1101/445601.

210. Young, A., Kumasaka, N., Calvert, F. & Hammond, T. R. A map of transcriptional heterogeneity and regulatory variation in human microglia. *bioRxiv* (2019).

211. Bennett, M. L. *et al.* New tools for studying microglia in the mouse and human CNS. *Proceedings of the National Academy of Sciences* **113**, E1738–E1746 (2016).

212. Satoh, J.-I. *et al.* TMEM119 marks a subset of microglia in the human brain.

Neuropathology **36**, 39–49 (2016).

213. Hammond, T. R. *et al.* Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes. *Immunity* (2018) doi:10.1016/j.immuni.2018.11.004.

214. Montagne, A. *et al.* Blood-brain barrier breakdown in the aging human hippocampus. *Neuron* **85**, 296–302 (2015).

215. Smith, E. E., Schneider, J. A., Wardlaw, J. M. & Greenberg, S. M. Cerebral microinfarcts: the invisible lesions. *Lancet Neurol.* **11**, 272–282 (2012).

216. Villa, A. *et al.* Sex-Specific Features of Microglia from Adult Mice. *Cell Rep.* **23**, 3501–3511 (2018).

217. Hanamsagar, R. *et al.* Generation of a microglial developmental index in mice and in humans reveals a sex difference in maturation and immune reactivity. *Glia* **66**, 460 (2018).

218. Thion, M. S. *et al.* Microbiome Influences Prenatal and Adult Microglia in a Sex-Specific Manner. *Cell* **172**, 500–516.e16 (2018).

219. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

220. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 182 (2014).

221. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

222. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

223. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

224. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

225. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).

226. Reimand, J. *et al.* g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).

227. Vela, D. Hepcidin, an emerging and important player in brain iron homeostasis. *J. Transl. Med.* **16**, 25 (2018).

228. Sköld, M. K., von Gertten, C., Sandberg-Nordqvist, A.-C., Mathiesen, T. & Holmin, S. VEGF and VEGF receptor expression after experimental brain contusion in rat. *J. Neurotrauma* **22**, 353–367 (2005).

229. Krum, J. M., Mani, N. & Rosenstein, J. M. Roles of the endogenous VEGF receptors flt-1 and flk-1 in astroglial and vascular remodeling after brain injury. *Exp. Neurol.* **212**, 108–117 (2008).

230. Ryu, J. K., Cho, T., Choi, H. B., Wang, Y. T. & McLarnon, J. G. Microglial VEGF receptor response is an integral chemotactic component in Alzheimer's disease pathology. *J. Neurosci.* **29**, 3–13 (2009).

231. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).

232. Ginhoux, F. & Prinz, M. Origin of microglia: current concepts and past controversies. *Cold Spring Harb. Perspect. Biol.* **7**, a020537 (2015).

233. Lyakh, L. A. *et al.* Adenovirus type 5 vectors induce dendritic cell differentiation in human CD14(+) monocytes cultured under serum-free conditions. *Blood* **99**, 600–608 (2002).

234. Muruve, D. A. *et al.* The inflammasome recognizes cytosolic microbial and host DNA and triggers an innate immune response. *Nature* **452**, 103–107 (2008).

235. Tsuchiya, S. *et al.* Establishment and characterization of a human acute monocytic leukemia cell line (THP-1). *Int. J. Cancer* **26**, 171–176 (1980).

236. Sundström, C. & Nilsson, K. Establishment and characterization of a human histiocytic lymphoma cell line (U-937). *Int. J. Cancer* **17**, 565–577 (1976).

237. Chanput, W., Peters, V. & Wichers, H. THP-1 and U937 Cells. in *The Impact of Food Bioactives on Health: in vitro and ex vivo models* (eds. Verhoeckx, K. et al.) 147–159 (Springer International Publishing, 2015).

238. Daigneault, M., Preston, J. A., Marriott, H. M., Whyte, M. K. B. & Dockrell, D. H. The identification of markers of macrophage differentiation in PMA-stimulated THP-1 cells and monocyte-derived macrophages. *PLoS One* **5**, e8668 (2010).

239. Chanput, W., Mes, J. J. & Wichers, H. J. THP-1 cell line: an in vitro cell model for immune modulation approach. *Int. Immunopharmacol.* **23**, 37–45 (2014).

240. Schildberger, A., Rossmanith, E., Eichhorn, T., Strassl, K. & Weber, V. Monocytes, peripheral blood mononuclear cells, and THP-1 cells exhibit different cytokine expression patterns following stimulation with lipopolysaccharide. *Mediators Inflamm.* **2013**, 697972 (2013).

241. Adati, N., Huang, M.-C., Suzuki, T., Suzuki, H. & Kojima, T. High-resolution analysis of aberrant regions in autosomal chromosomes in human leukemia THP-1 cell line. *BMC Res. Notes* **2**, 153 (2009).

242. Buchrieser, J., James, W. & Moore, M. D. Human Induced Pluripotent Stem Cell-Derived Macrophages Share Ontogeny with MYB-Independent Tissue-Resident Macrophages. *Stem Cell Reports* **8**, 334–345 (2017).

243. Garcia-Reitboeck, P. *et al.* Human Induced Pluripotent Stem Cell-Derived Microglia-Like Cells Harboring TREM2 Missense Mutations Show Specific Deficits in Phagocytosis. *Cell Rep.* **24**, 2300–2311 (2018).

244. Gan, X. *et al.* Epigenetically repressing human cytomegalovirus lytic infection and reactivation from latency in THP-1 model by targeting H3K9 and H3K27 histone demethylases. *PLoS One* **12**, e0175390 (2017).

245. Phanstiel, D. H. *et al.* Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol. Cell* **67**, 1037–1048.e6 (2017).

246. Yeung, A. T. Y. *et al.* Exploiting induced pluripotent stem cell-derived macrophages to

unravel host factors influencing Chlamydia trachomatis pathogenesis. *Nat. Commun.* **8**, 15013 (2017).

247. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37–53 (2016).

248. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

249. Hicks, S. C. & Irizarry, R. A. quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biol.* **16**, 117 (2015).

250. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software* vol. 67 (2015).

251. Jonsson, T. *et al.* Variant of TREM2 Associated with the Risk of Alzheimer's Disease. *N. Engl. J. Med.* **368**, 107–116 (2013).

252. Guerreiro, R. *et al.* TREM2 Variants in Alzheimer's Disease. *N. Engl. J. Med.* **368**, 117–127 (2013).

253. Nguyen, Q. H. *et al.* Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. *Genome Res.* **28**, 1053–1066 (2018).

254. Paik, D. T. *et al.* Large-Scale Single-Cell RNA-Seq Reveals Molecular Signatures of Heterogeneous Populations of Human Induced Pluripotent Stem Cell-Derived Endothelial Cells. *Circ. Res.* **123**, 443–450 (2018).

255. McCracken, I. R. *et al.* Transcriptional dynamics of pluripotent stem cell-derived endothelial cell differentiation revealed by single-cell RNA sequencing. *Eur. Heart J.* (2019) doi:10.1093/eurheartj/ehz351.

256. Etzrodt, M., Endele, M. & Schroeder, T. Quantitative single-cell approaches to stem cell research. *Cell Stem Cell* **15**, 546–558 (2014).

257. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism.

Nature **541**, 331–338 (2017).

258. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

259. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

260. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

261. Paonessa, F. *et al.* Microtubules Deform the Nuclear Membrane and Disrupt Nucleocytoplasmic Transport in Tau-Mediated Frontotemporal Dementia. *Cell Rep.* **26**, 582–593.e5 (2019).

262. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

263. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

264. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

265. Leung, G. A. *et al.* The lymphoid-associated interleukin 7 receptor (IL7R) regulates tissue-resident macrophage development. *Development* **146**, (2019).

266. Grabert, K. *et al.* Microglial brain region-dependent diversity and selective regional sensitivities to aging. *Nat. Neurosci.* **19**, 504–516 (2016).

267. Parakalan, R. *et al.* Transcriptome analysis of amoeboid and ramified microglia isolated from the corpus callosum of rat brain. *BMC Neurosci.* **13**, 64 (2012).

268. Rothe, T. *et al.* The Nuclear Receptor Nr4a1 Acts as a Microglia Rheostat and Serves as a Therapeutic Target in Autoimmune-Driven Central Nervous System Inflammation. *J. Immunol.* **198**, 3878–3885 (2017).

269. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

270. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).

271. Burgess, D. J. Spatial transcriptomics coming of age. *Nature reviews. Genetics* vol. 20 317 (2019).

272. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 1–9 (2019).

273. Cakir, B. *et al.* Engineering of human brain organoids with a functional vascular-like system. *Nat. Methods* **16**, 1169–1175 (2019).

274. Phan, D. T. *et al.* Blood-brain barrier-on-a-chip: Microphysiological systems that capture the complexity of the blood-central nervous system interface. *Exp. Biol. Med.* **242**, 1669–1678 (2017).

275. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *bioRxiv* 630996 (2019) doi:10.1101/630996.

276. McManus, J., Cheng, Z. & Vogel, C. Next-generation analysis of gene expression regulation--comparing the roles of synthesis and degradation. *Mol. Biosyst.* **11**, 2680–2689 (2015).

277. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

278. Choi, S. H., Kim, Y. H., Quinti, L., Tanzi, R. E. & Kim, D. Y. 3D culture models of Alzheimer's disease: a road map to a 'cure-in-a-dish'. *Mol. Neurodegener.* **11**, 75 (2016).

279. Penney, J., Ralvenius, W. T. & Tsai, L.-H. Modeling Alzheimer's disease with iPSC-derived brain cells. *Mol. Psychiatry* (2019) doi:10.1038/s41380-019-0468-3.

280. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).