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.

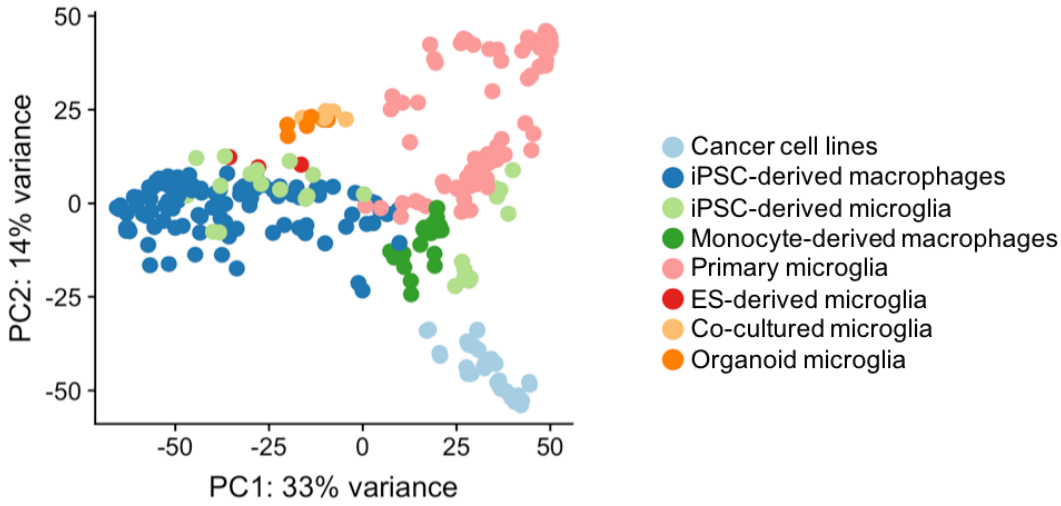

Figure 4.1 PC1 vs PC2 of model comparison dataset

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).

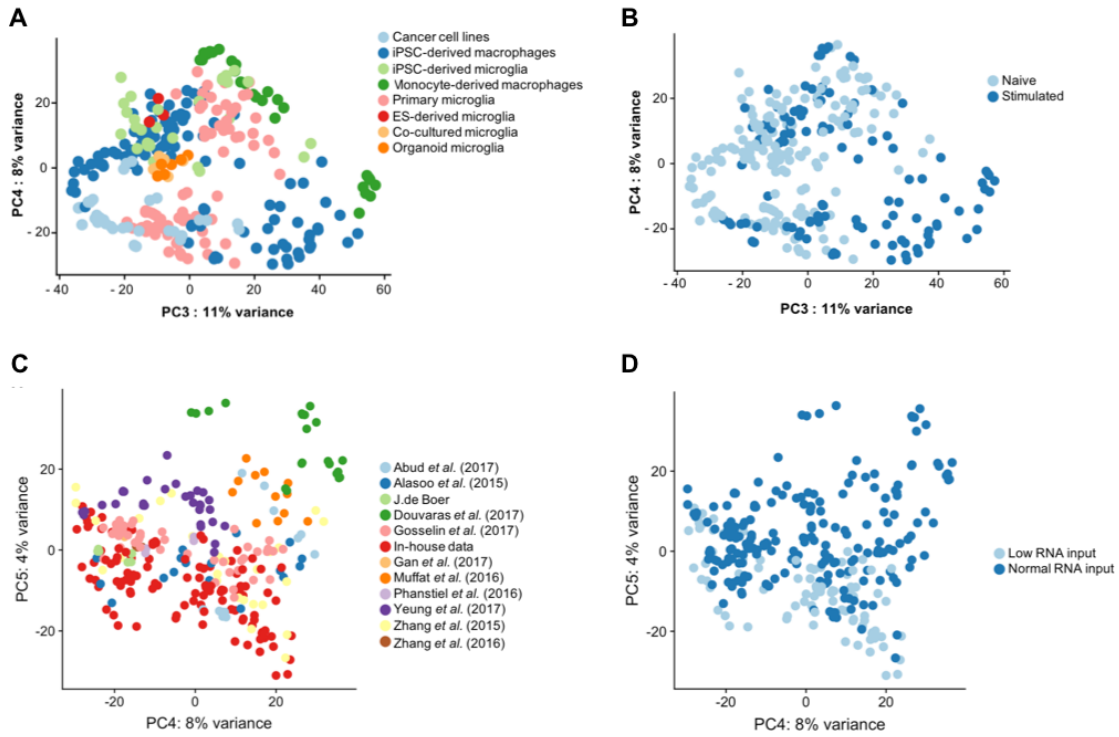

Figure 4.2 PC3 vs PC4 and PC4 vs PC5 of model comparison dataset

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.

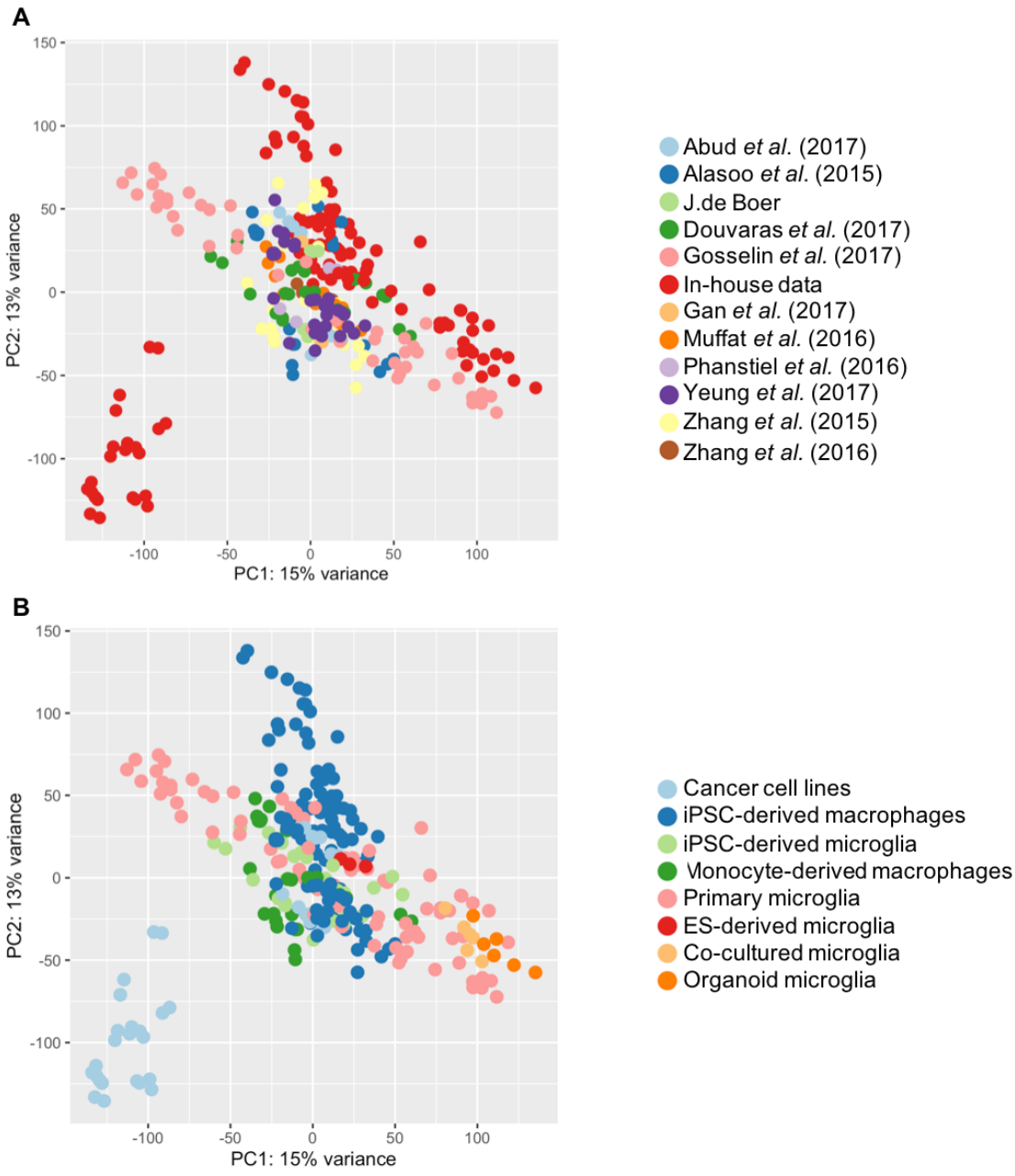

Figure 4.3 PC1 vs PC2 of residual values across all genes following removal of study based effects

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).

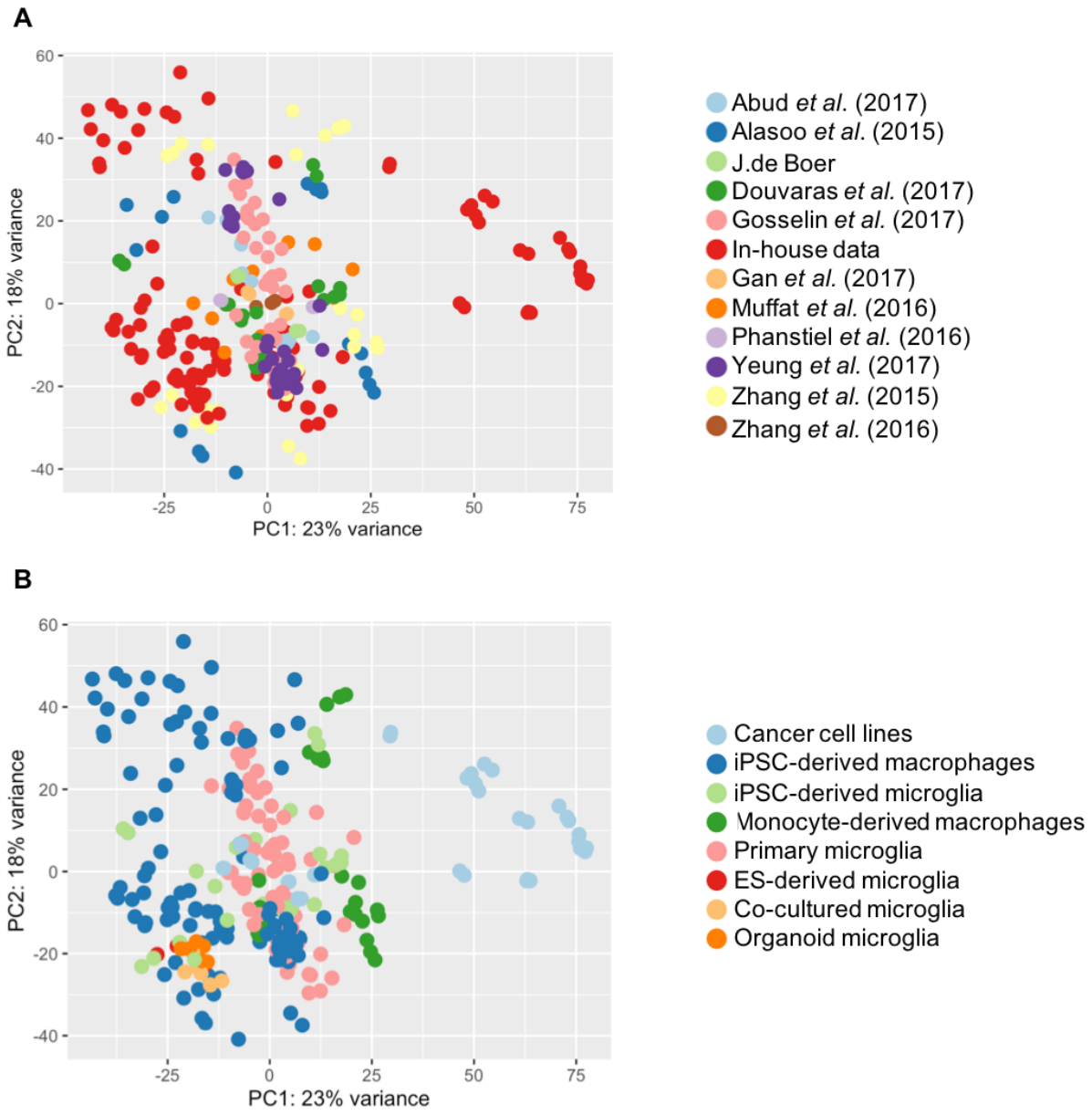

Figure 4.4 PC1 vs PC2 of residual values from the top 500 most variable genes following removal of study based effects

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.

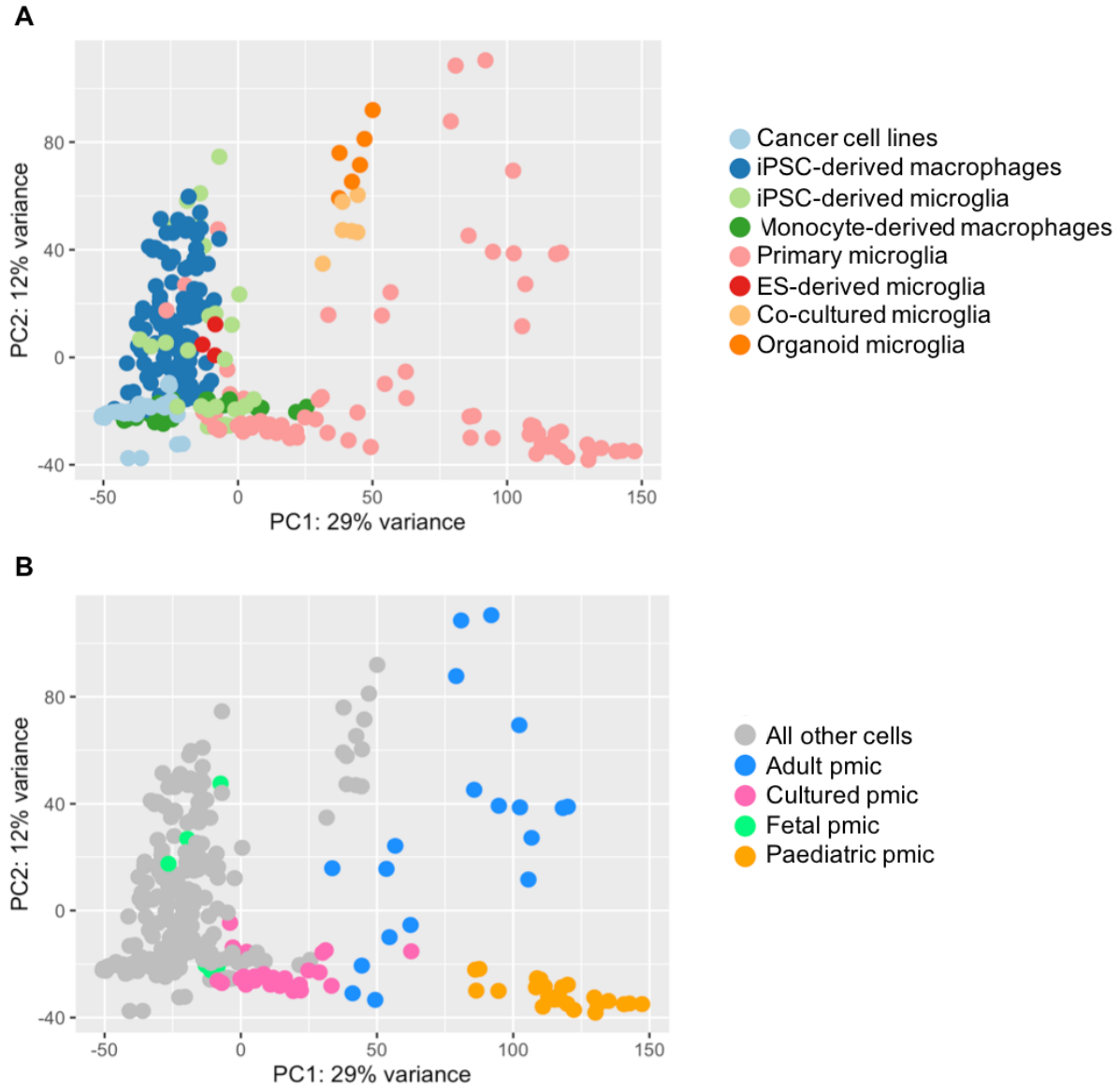

Figure 4.5 PC1 vs PC2 of all samples using the PMM input gene list

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.

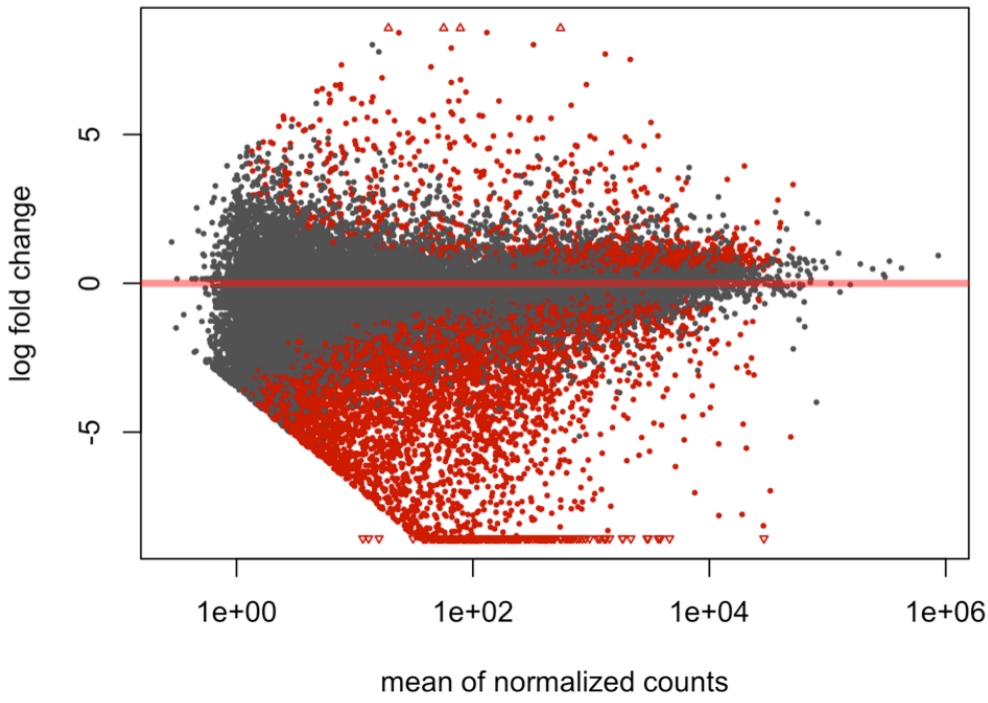

Figure 4.6 MA plot of differentially expressed genes comparing monoculture vs complex stem cell derived microglia

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).

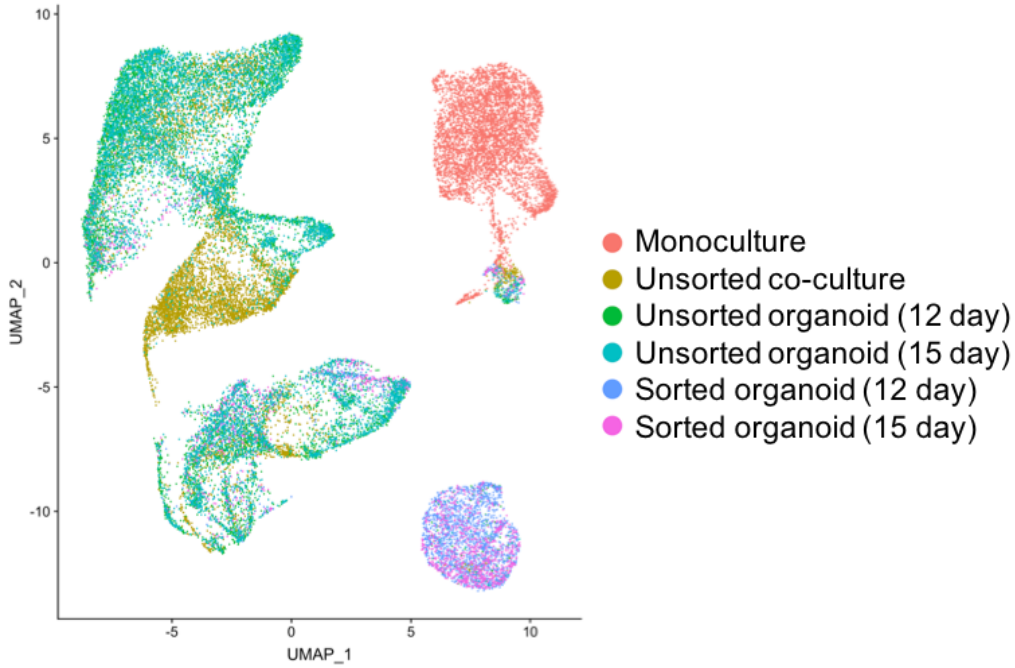

Figure 4.7 UMAP of full dataset

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.

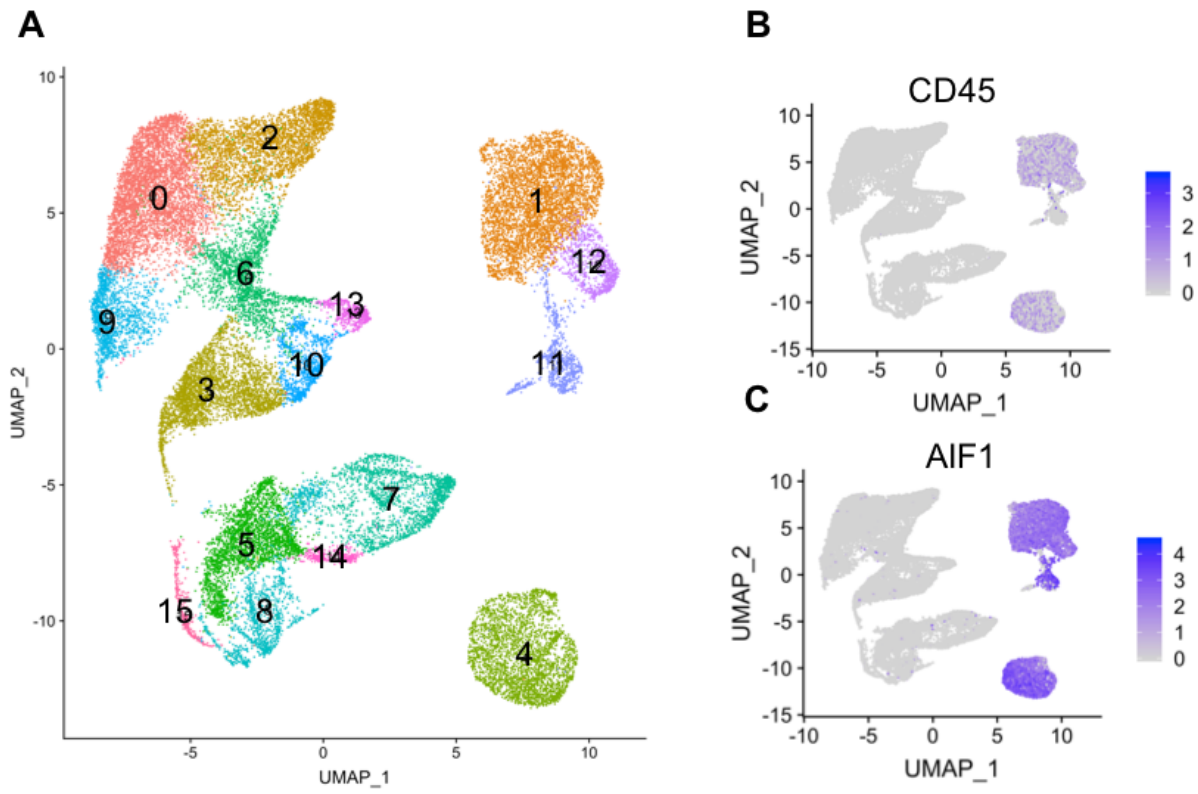

Figure 4.8 Identification of myeloid cells

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).

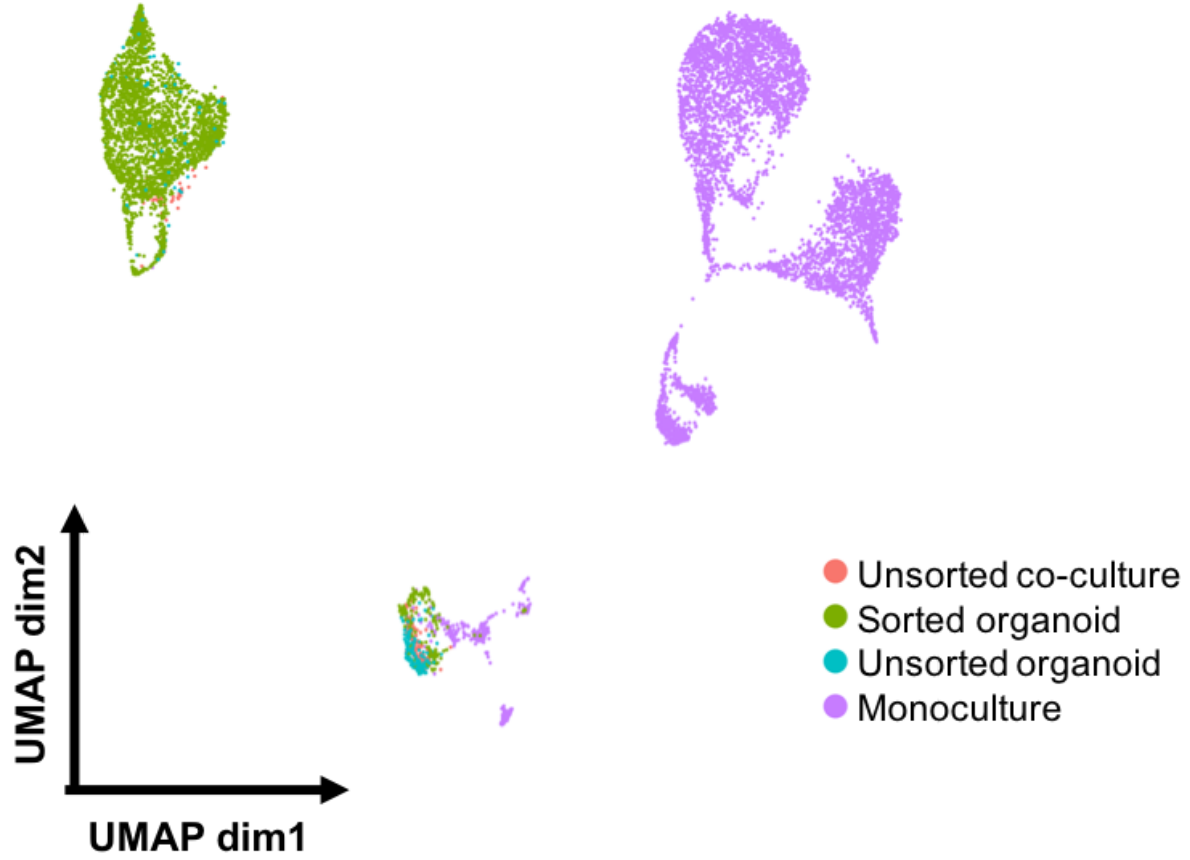

Figure 4.9 UMAP of myeloid cells in Monocle3

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).

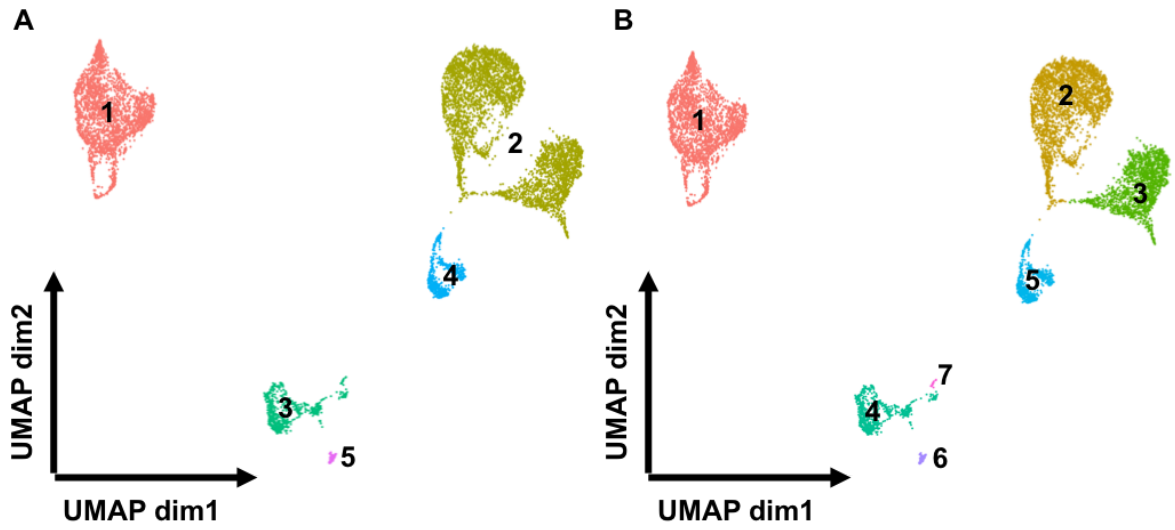

Figure 4.10 UMAP of myeloid cells in Monocle3

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.

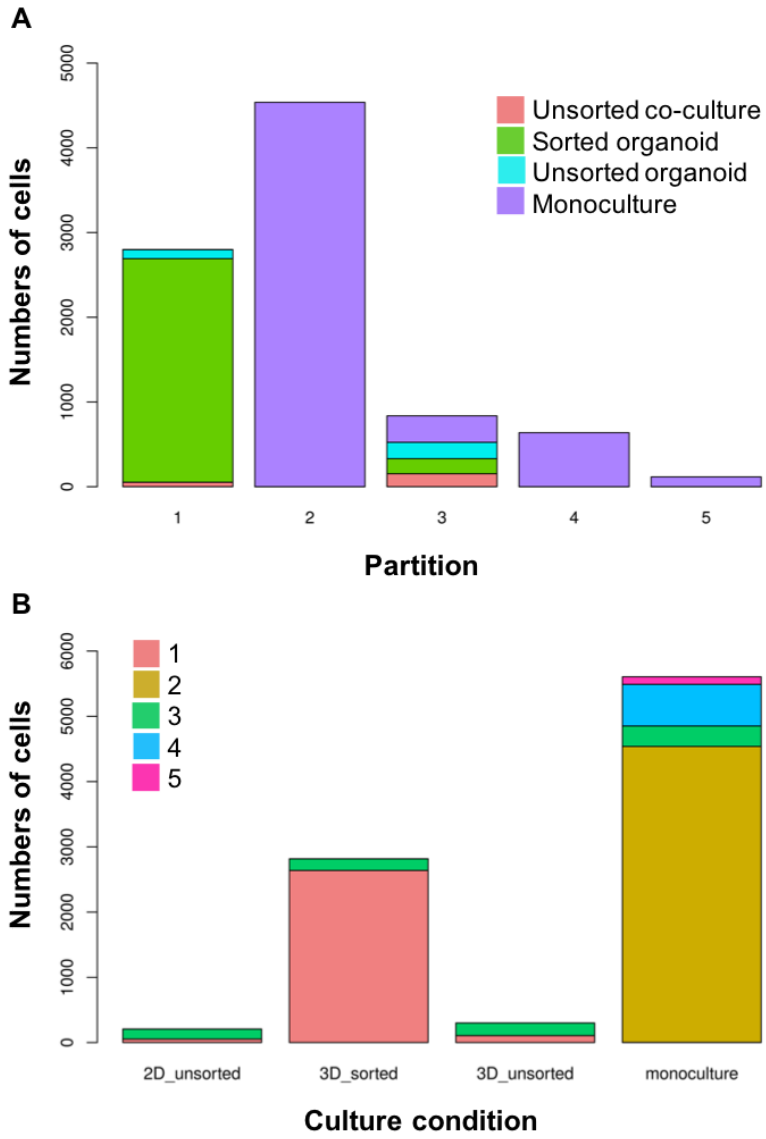

Figure 4.11 Number of cells in each partition

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)

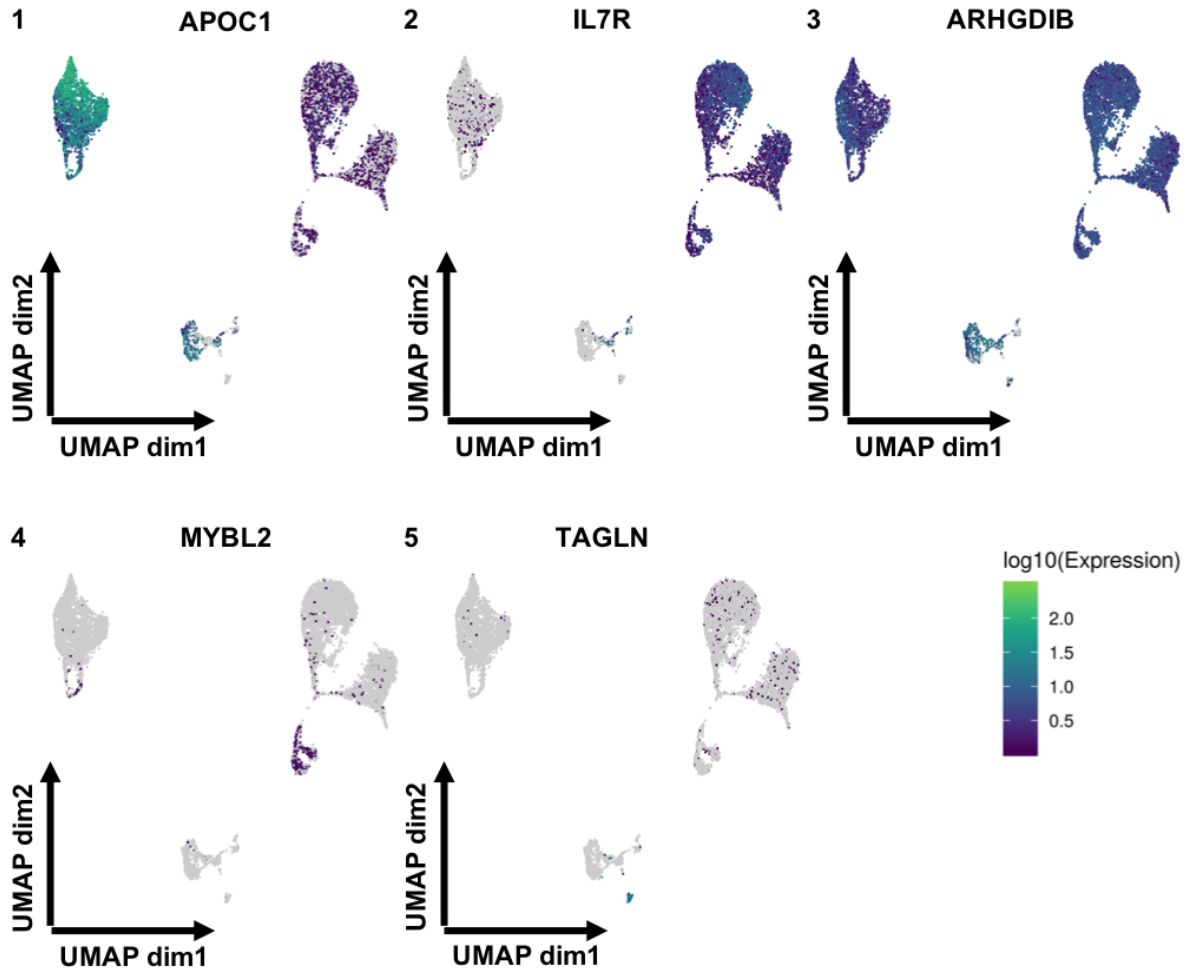

Figure 4.12 UMAP of myeloid cells in Monocle3 coloured by marker gene expression

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.

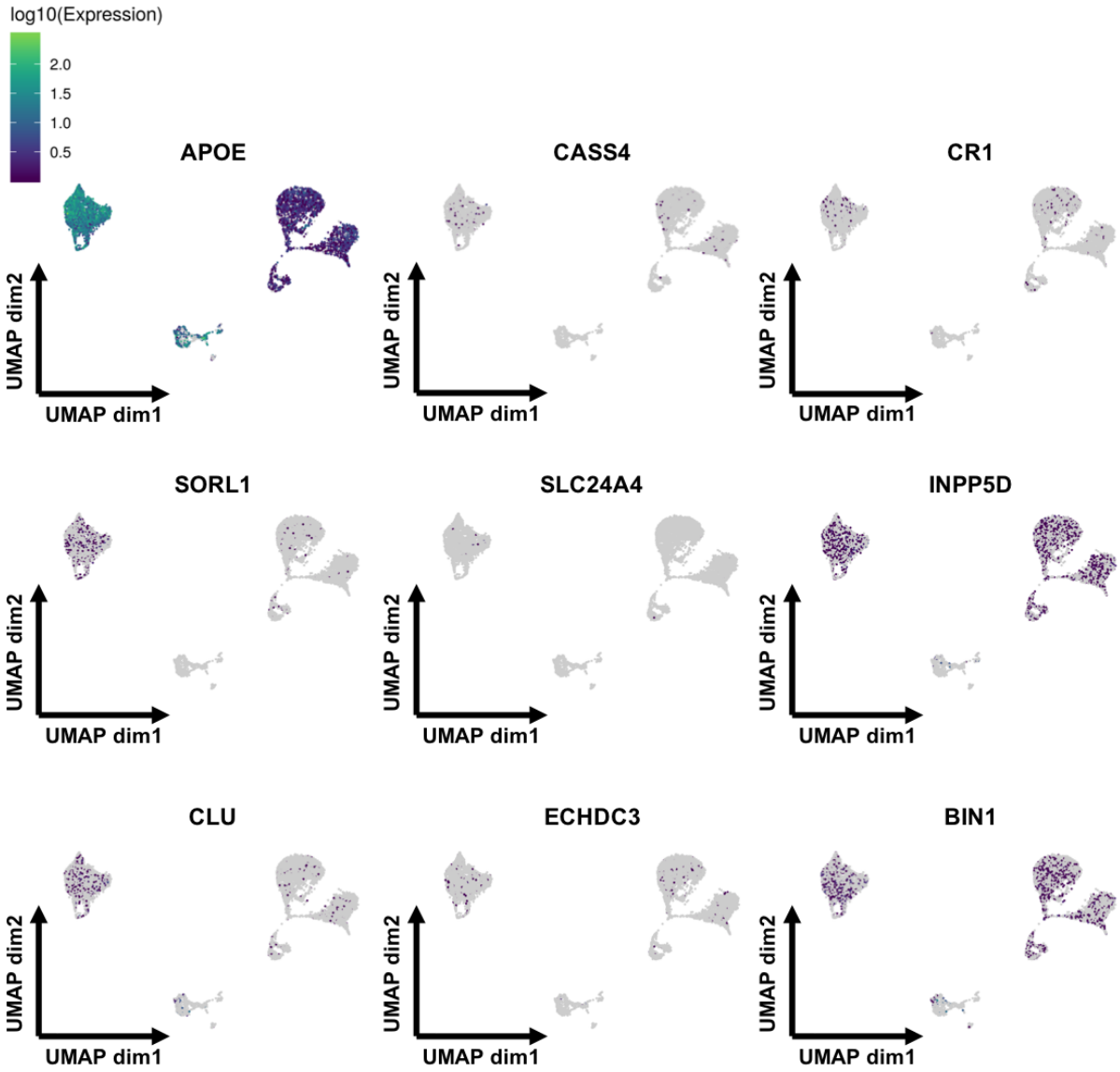

Figure 4.13 UMAP of myeloid cells in Monocle3 coloured by AD gene expression

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.

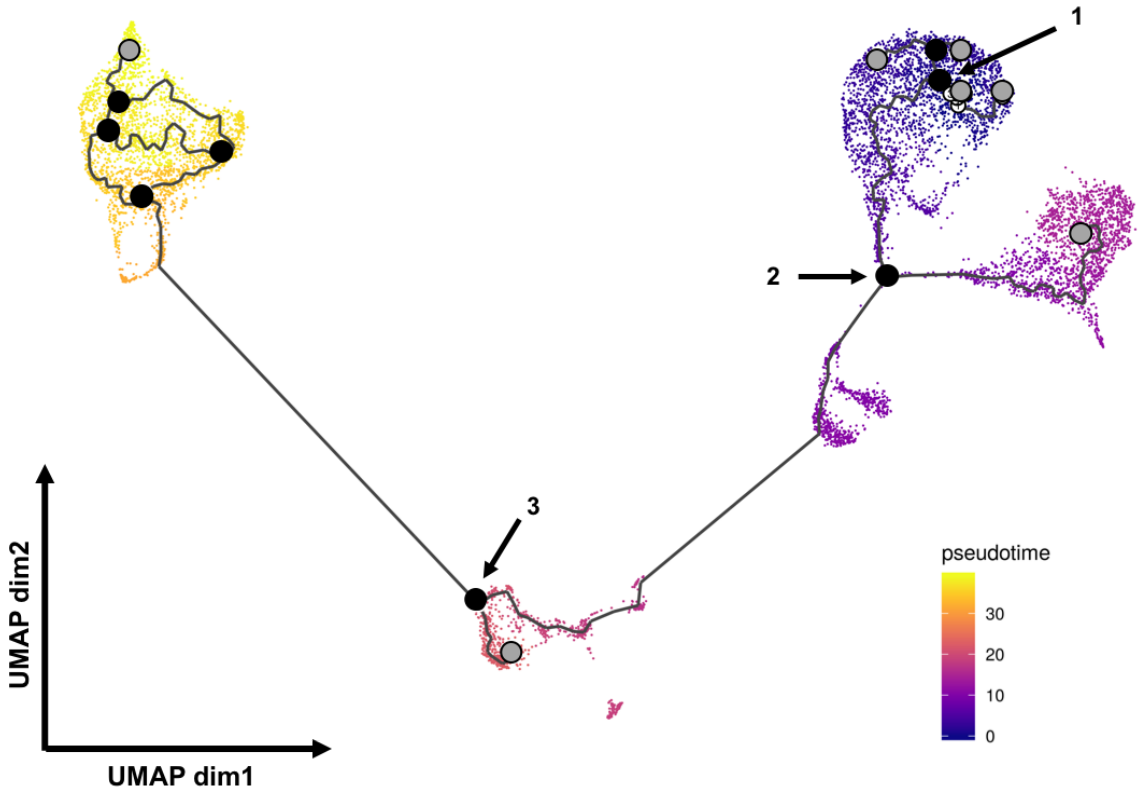

Figure 4.14 UMAP of myeloid cells in Monocle3 coloured by order in pseudotime

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].

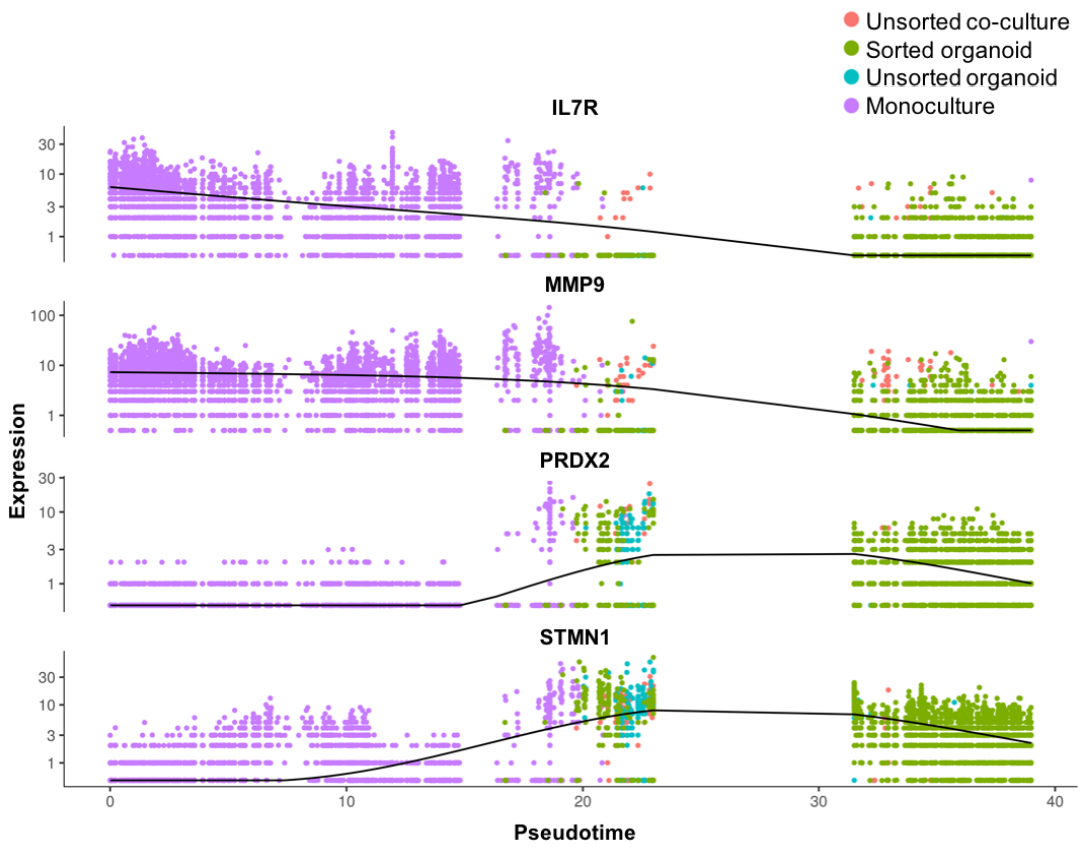

Figure 4.15 Expression of genes along pseudotime

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].

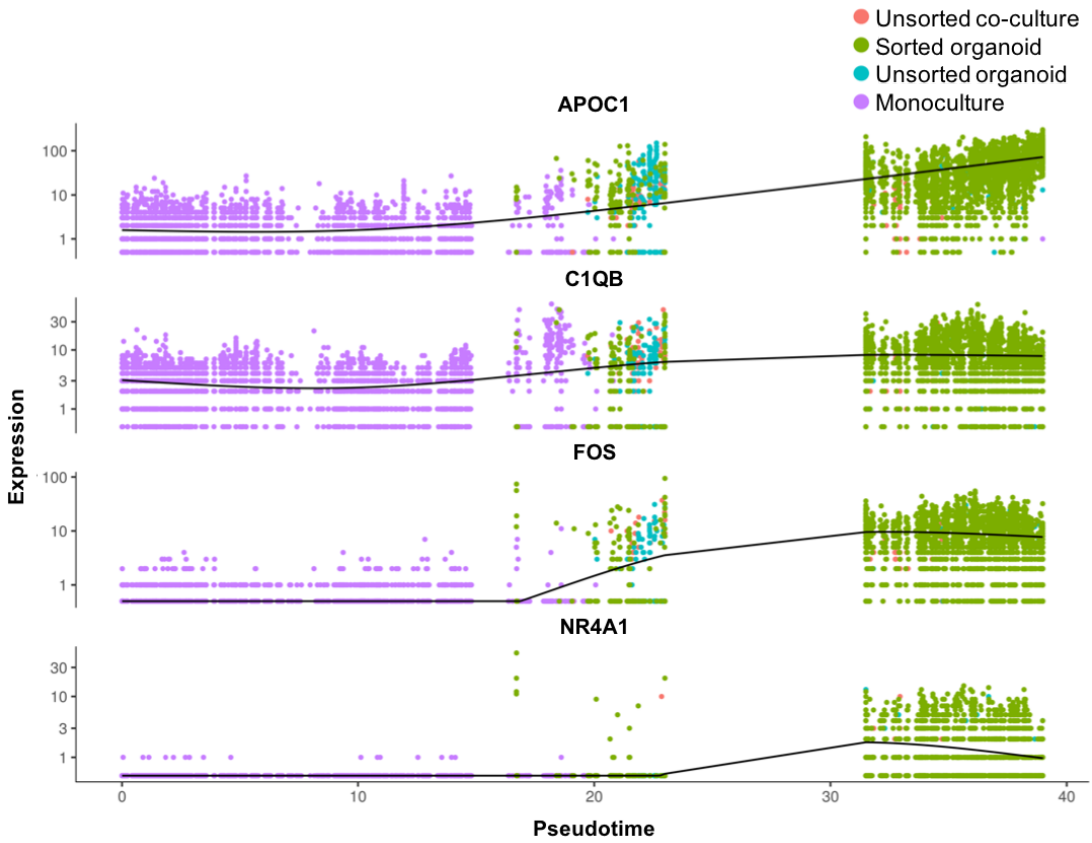

Figure 4.16 Genes with increasing expression along pseudotime

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.