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.

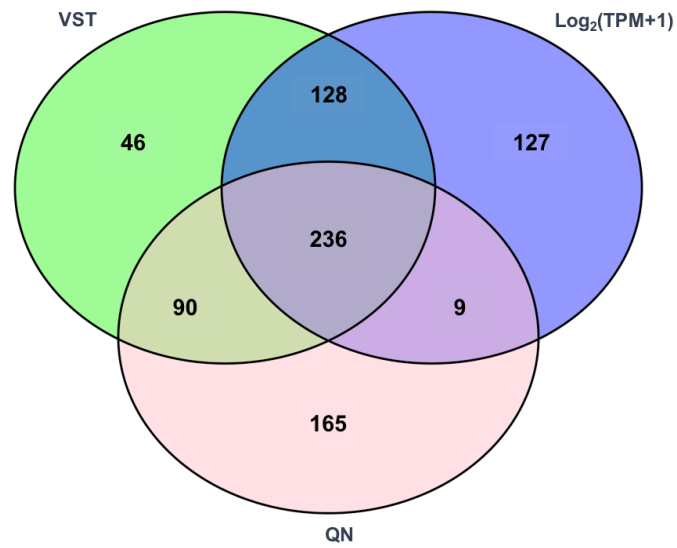

Figure 3.1 Venn diagram of overlapping most variable genes

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.

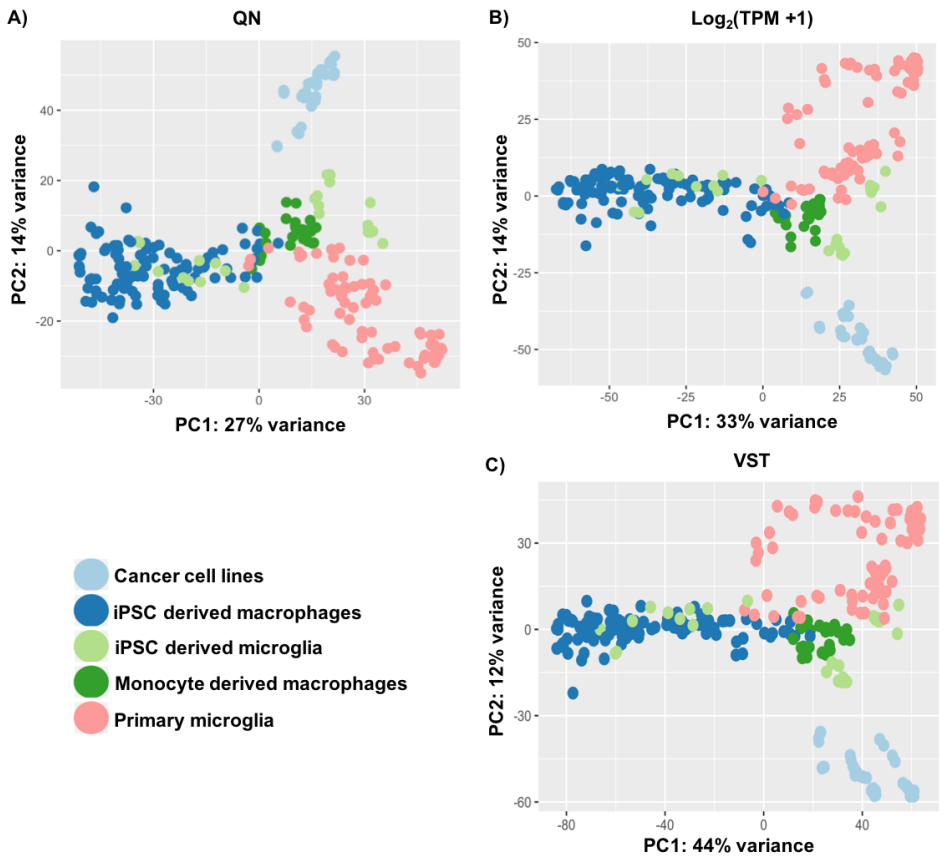

Figure 3.2 PC1 vs PC2 for three normalisation methods

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.

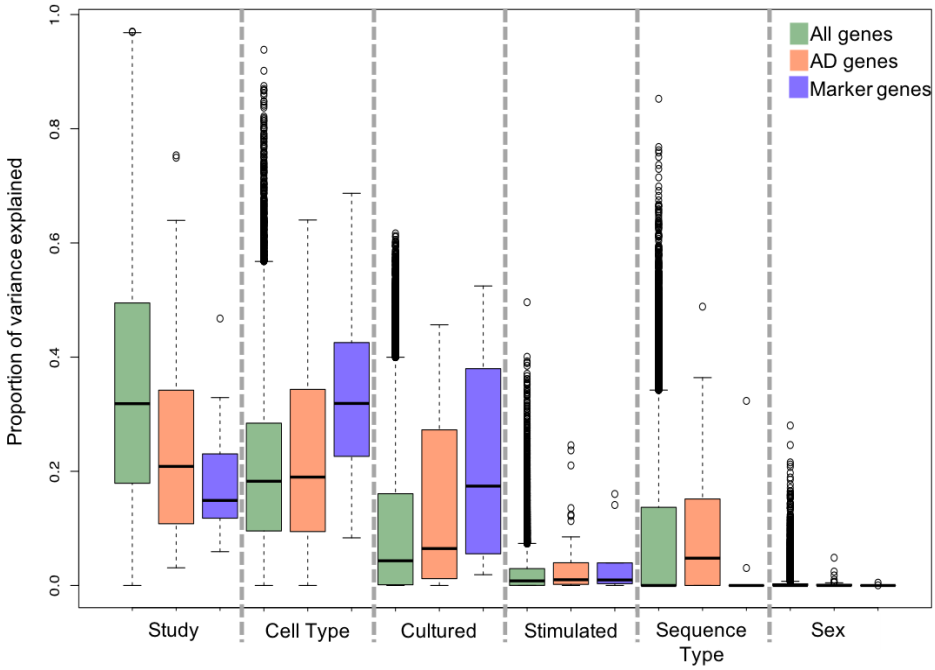

Figure 3.3 Variance components analysis

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.

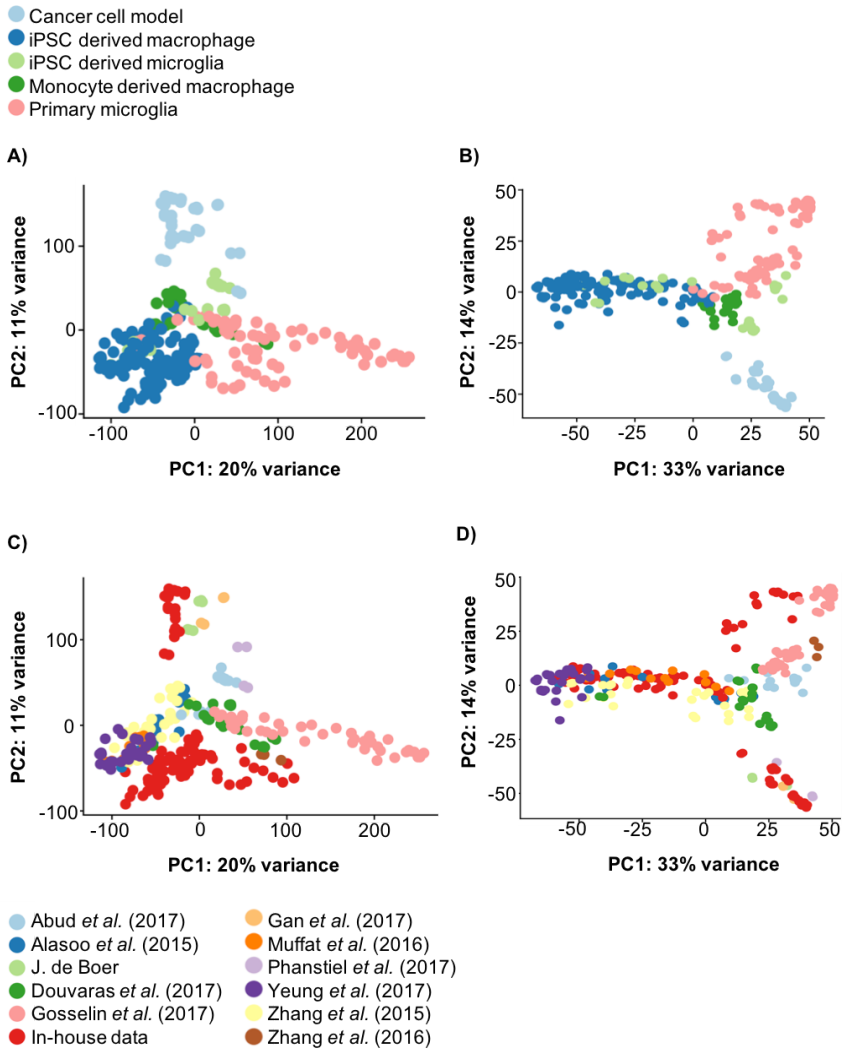

Figure 3.4 PC1 vs PC2 for all genes and top 500 genes

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.

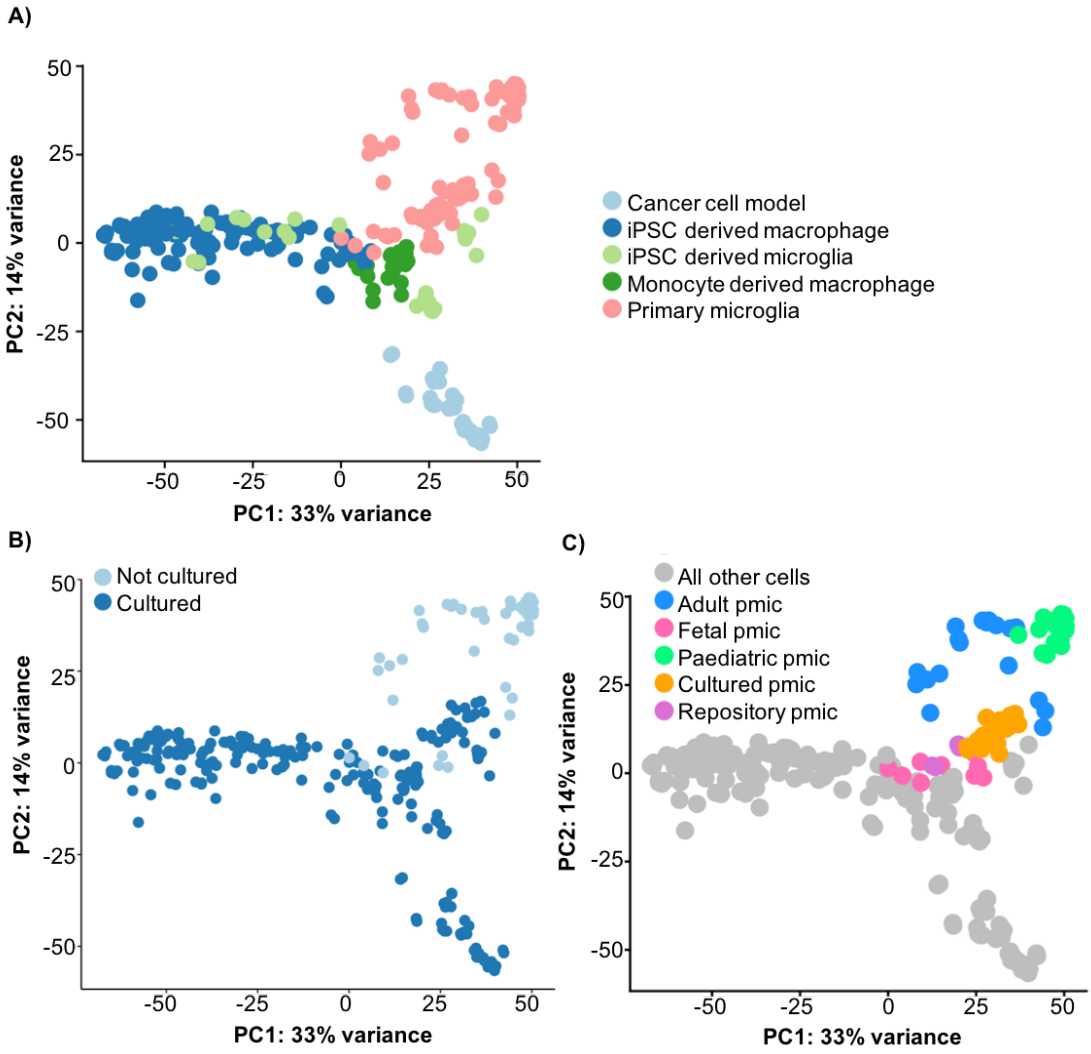

Figure 3.5 PC1 vs PC2 calculated using the top 500 genes

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

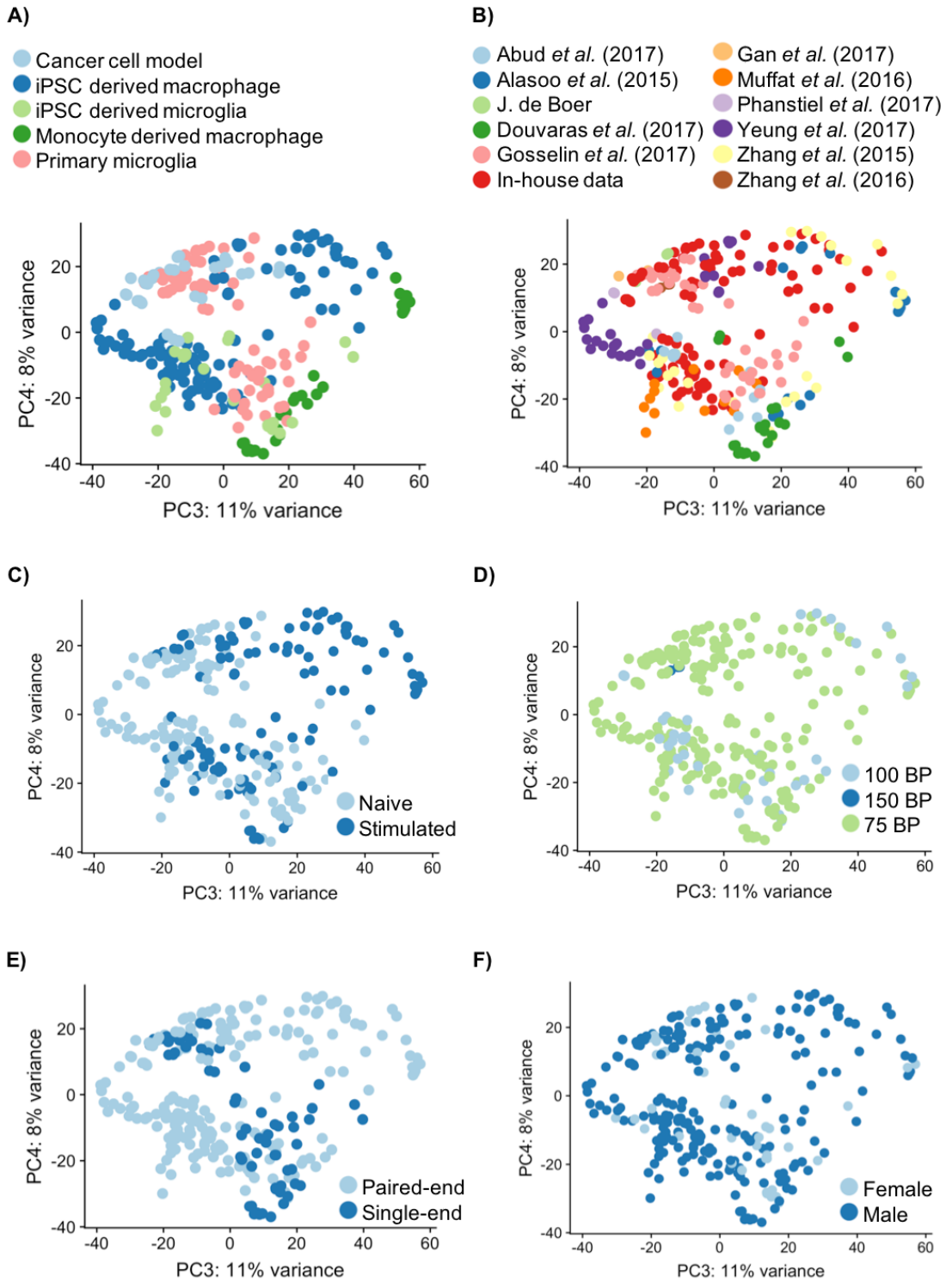

Figure 3.6 PC3 vs PC4 calculated using the top 500 genes

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.

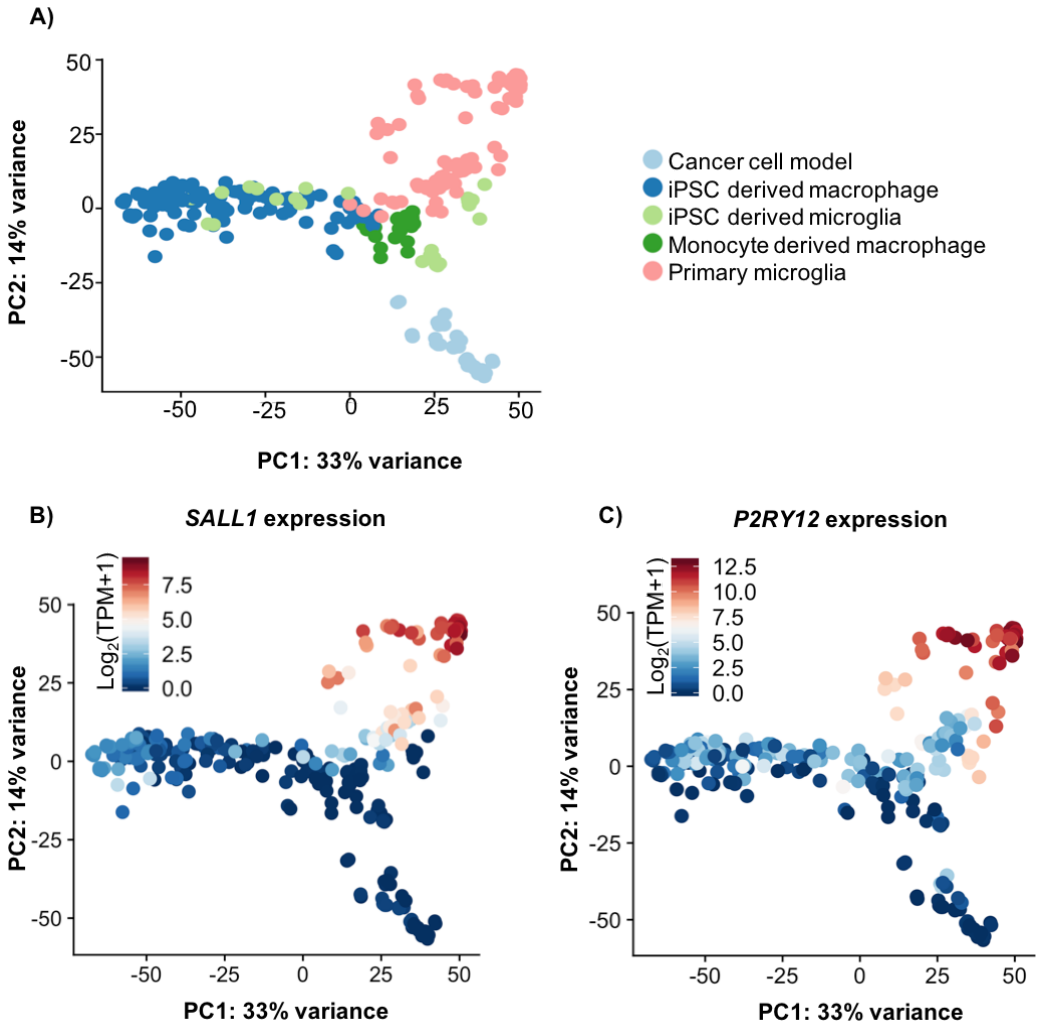

Figure 3.7 PC1 vs PC2 coloured by expression of genes heavily loading PC2

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, $p_{adj} = 2.52e^{-25}$).

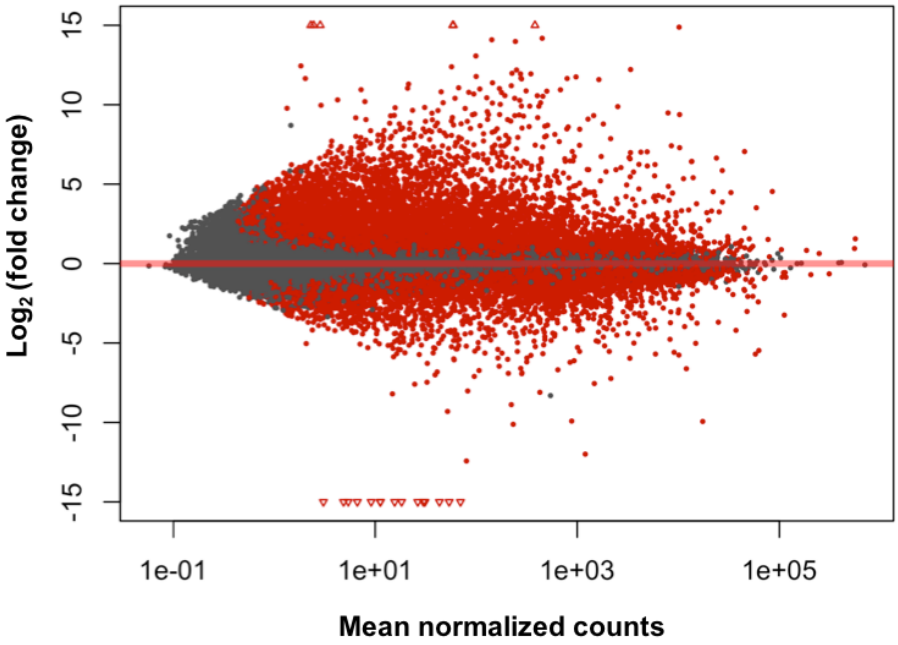

Figure 3.8 MA plot following differential expression analysis comparing primary microglia to all other cell types

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).

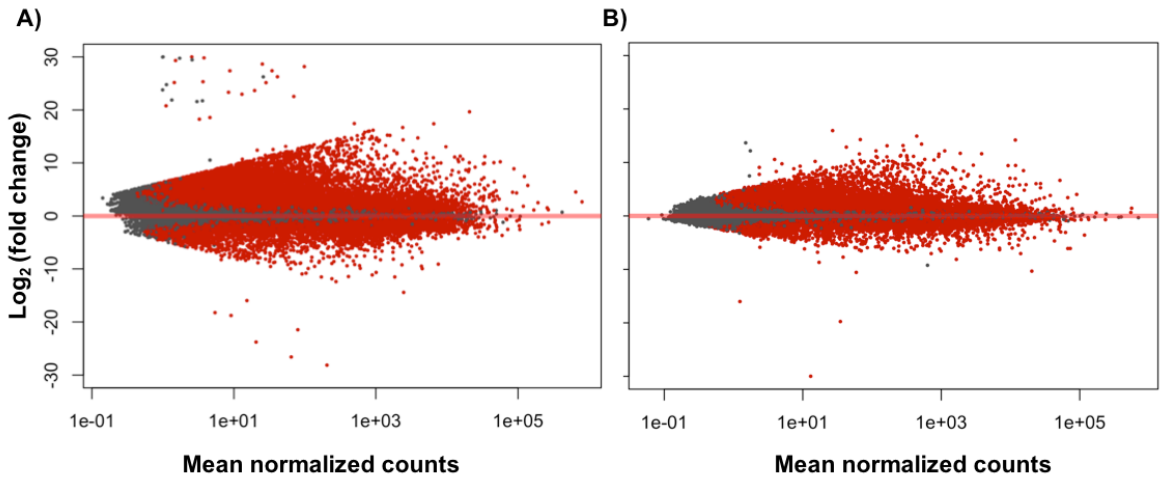

Figure 3.9 MA plots comparing primary microglia to cancer cell lines and iPSC-derived cells

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.

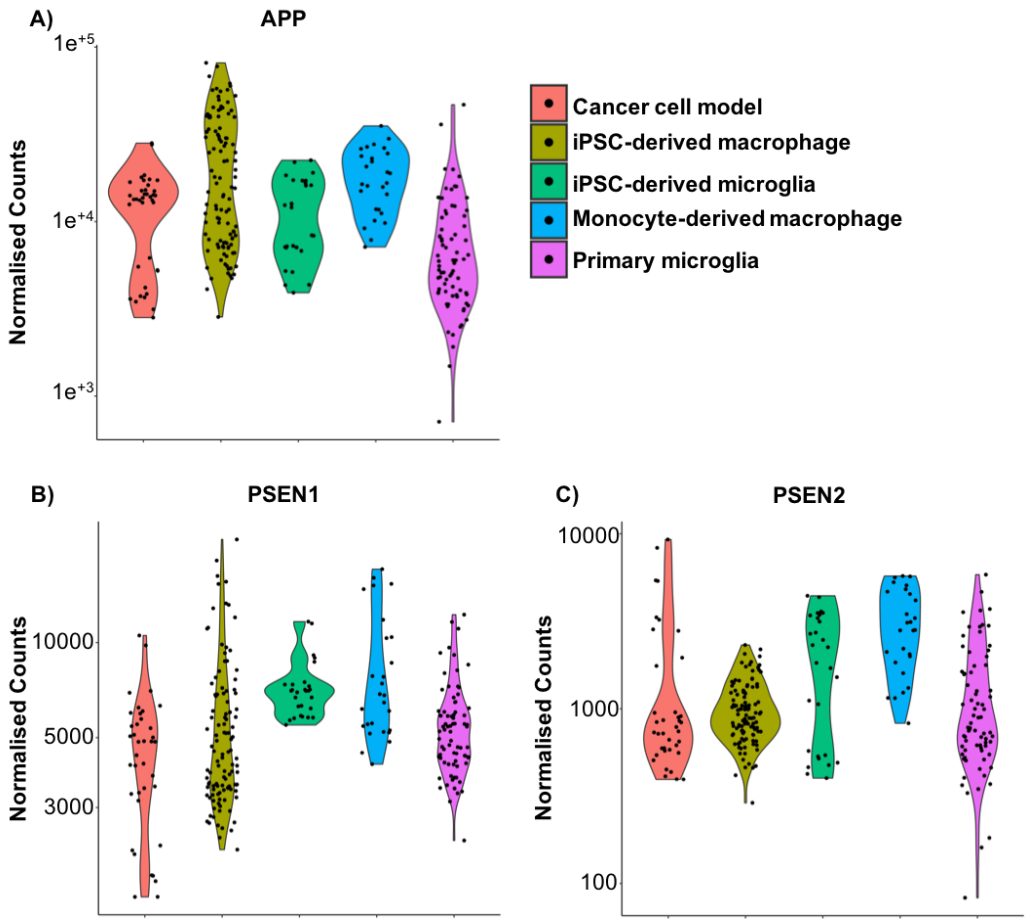

Figure 3.10 Expression of familial AD genes by cell type

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.

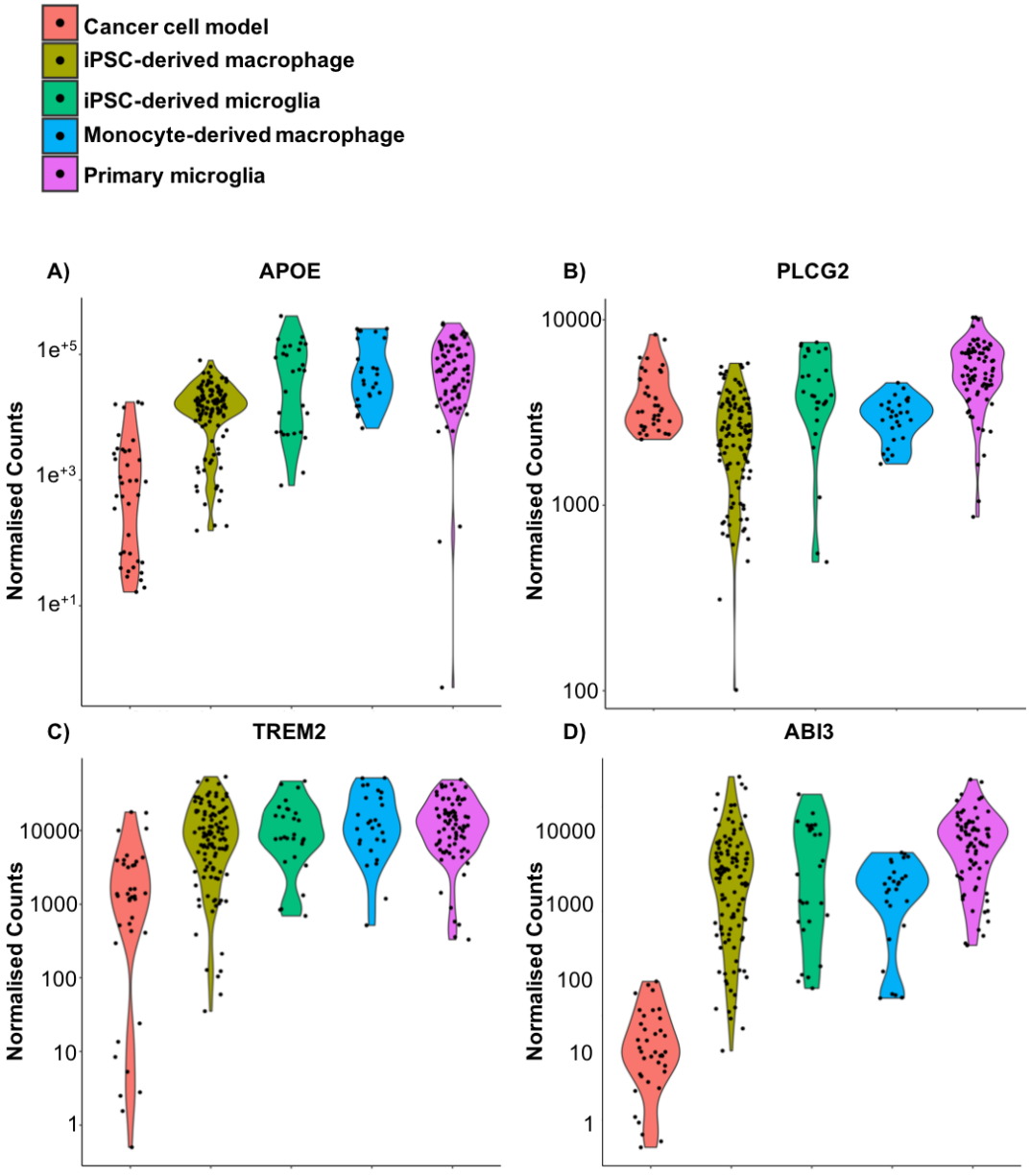

Figure 3.11 Expression of late onset AD rare and high effect size genes by cell type

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).

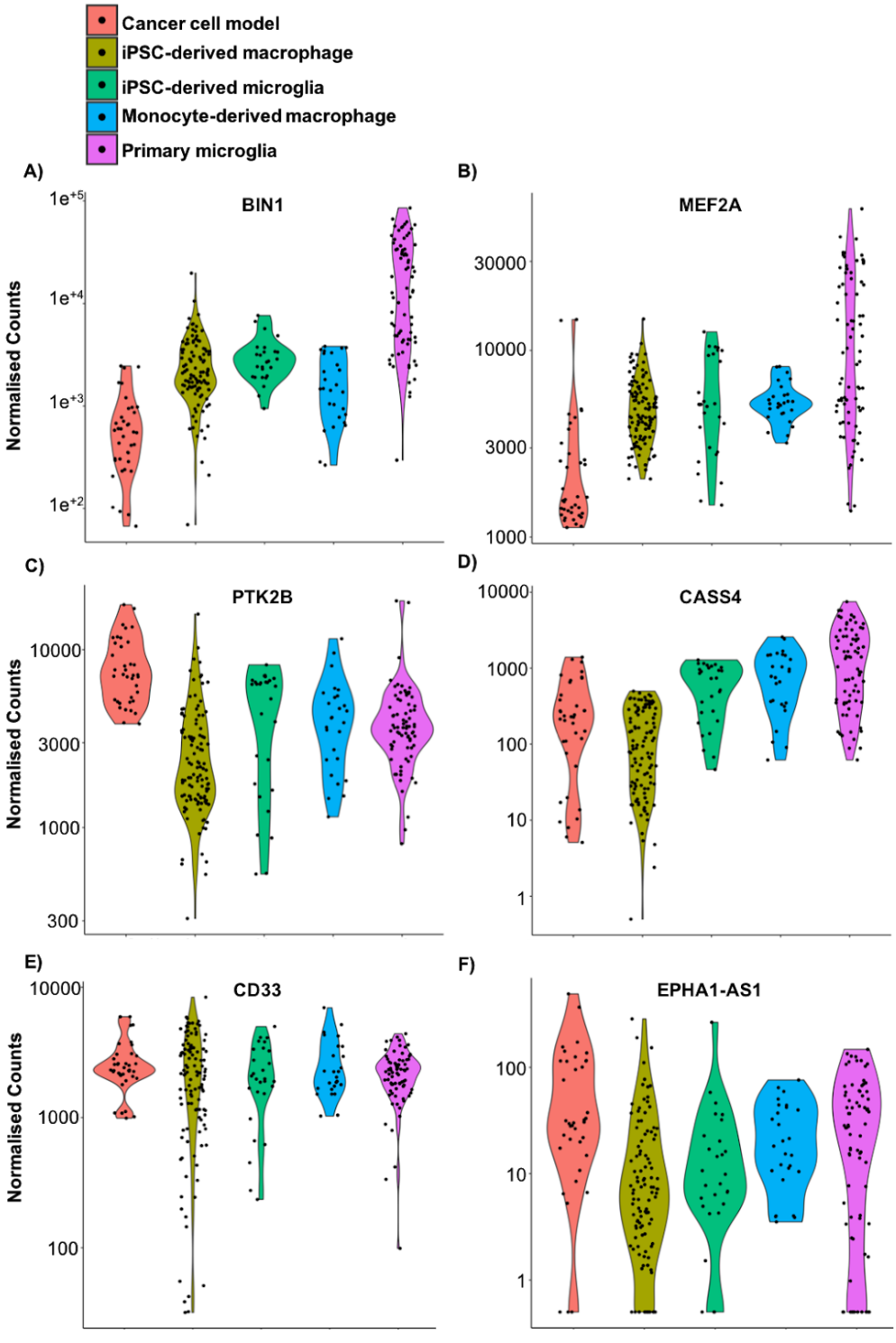

Figure 3.12 Expression of late onset AD risk genes

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.