Chapter 2: Heterogeneity in primary adult microglial transcriptomes

Collaboration note

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

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

2.1 Introduction

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

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

2.1.1 Marker gene identification in mice and human samples

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

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

2.1.2 Fresh, primary human microglia bulk RNA-sequencing

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

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

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

2.1.3 Single cell sequencing and primary microglia

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

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

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

2.1.4 The impact on age and sex on microglial transcriptomes

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

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

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

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

2.2 Methods

2.2.1 Experimental design and sample collection

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

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

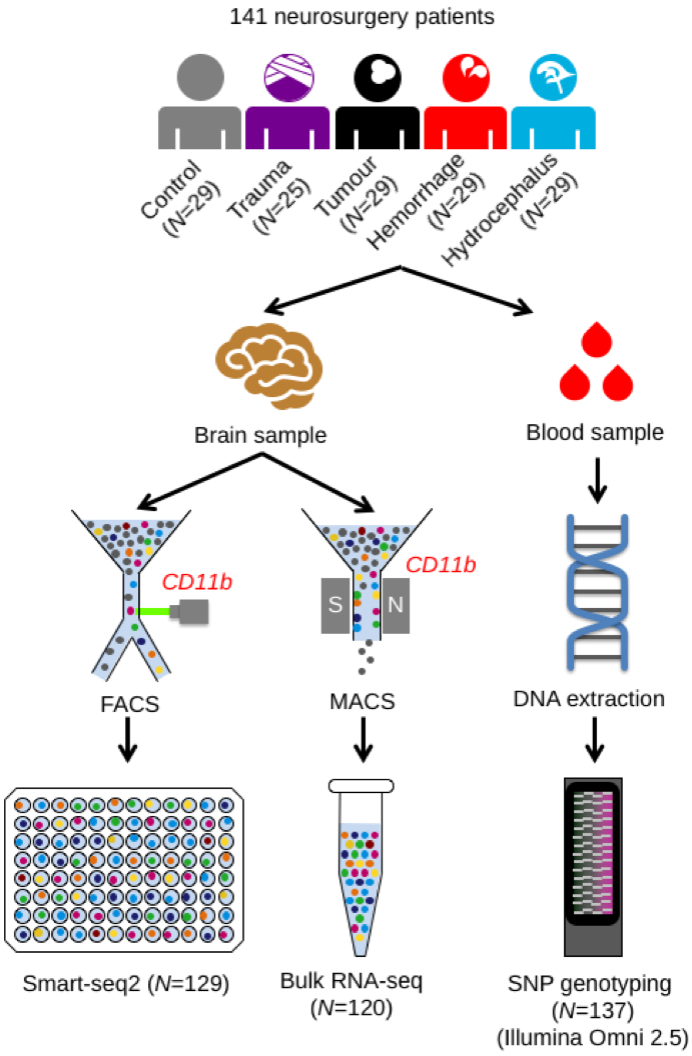

Figure 2.1 Schematic of experimental design

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

2.2.2 Tissue processing and cell sorting

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

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

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

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

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

2.2.3 RNA handling

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

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

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

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

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

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

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

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

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

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

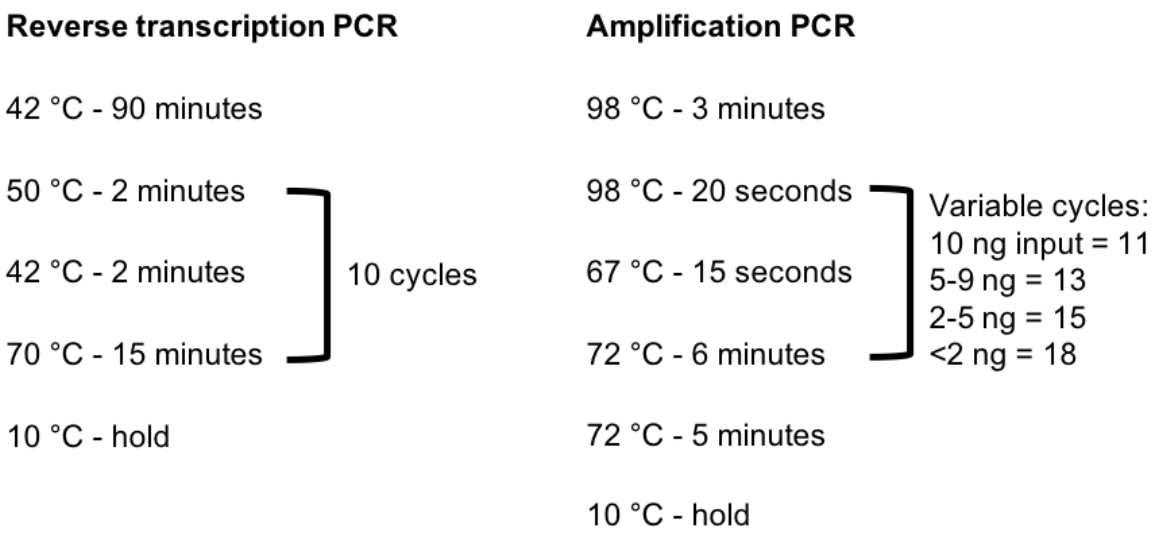

Figure 2.2 PCR reactions in low-input RNA-sequencing library preparation

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

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

2.2.4 Initial processing and quality control of sequencing data

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

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

2.2.5 Comparison of bulk data to publicly available datasets

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

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

2.2.6 Classification of microglial cells using publicly available datasets

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

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

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

2.2.7 Variance components analysis

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

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

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

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

2.2.9 Pathway enrichment analysis

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

2.3 Quality control analysis across datasets

2.3.1 Bulk RNA-sequencing quality control

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

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

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

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

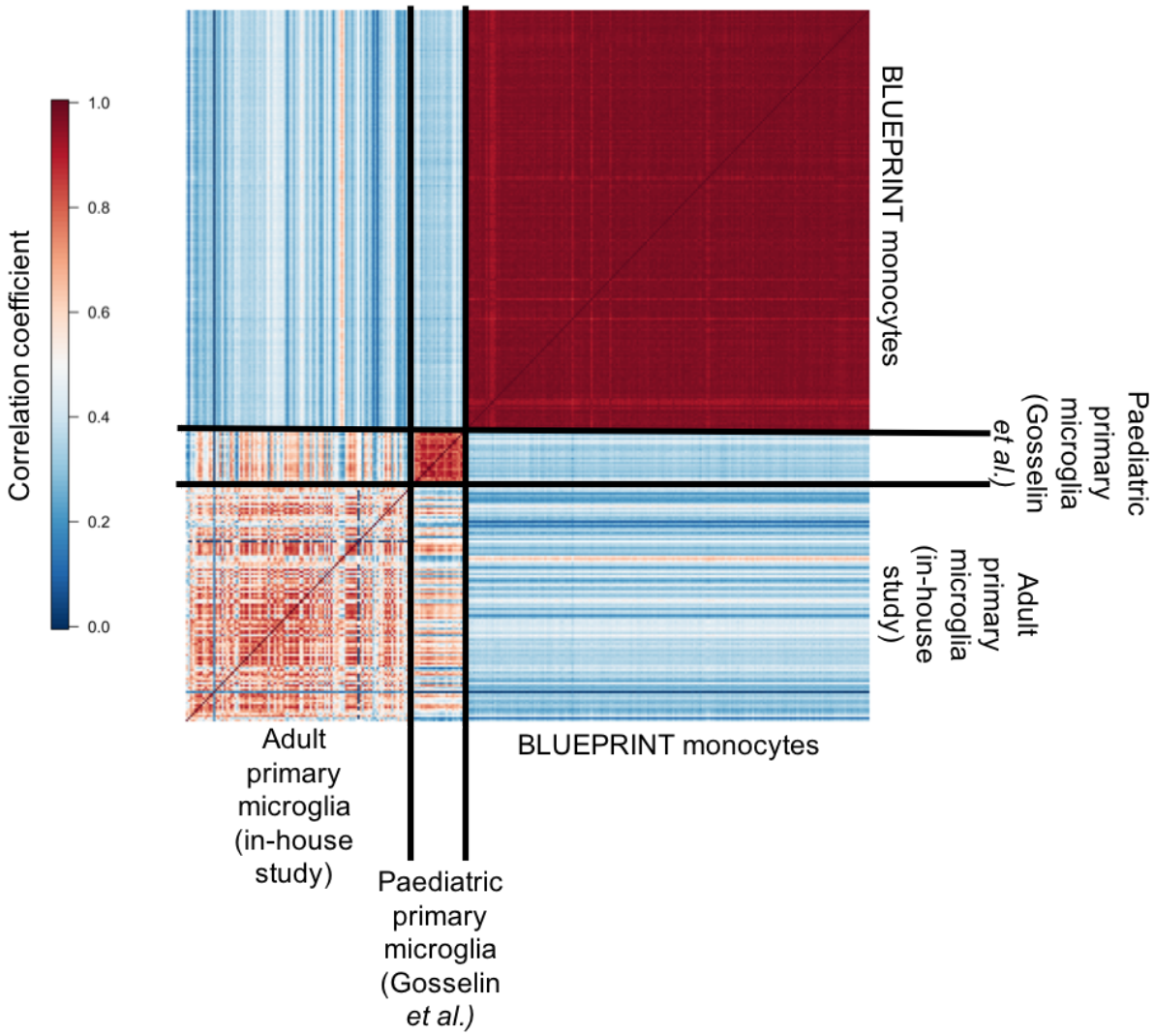

Figure 2.3 Heatmap of correlation of bulk RNA-seq gene expression between samples in primary microglia and BLUEPRINT monocytes

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

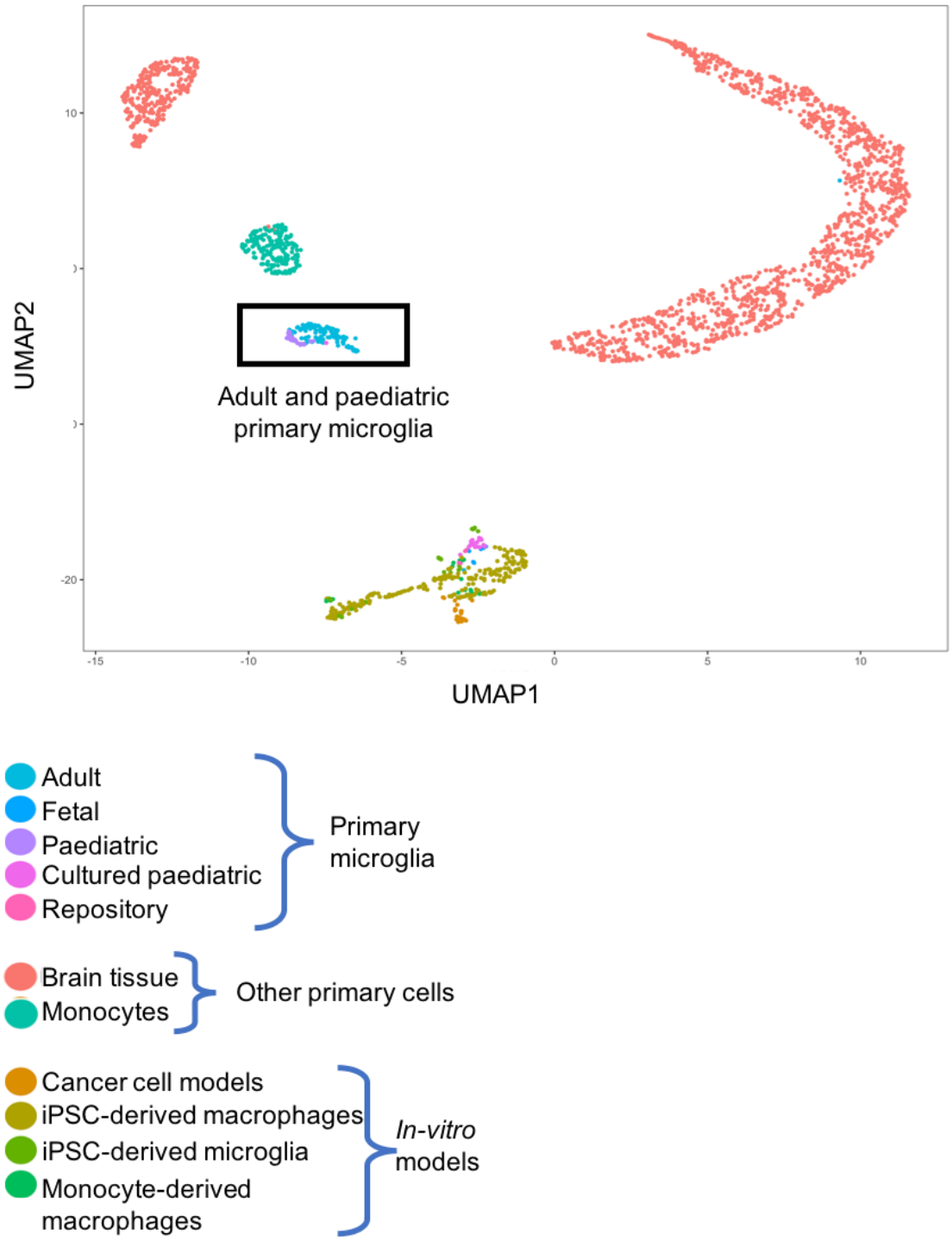

Figure 2.4 UMAP analysis of bulk primary microglia data and publicly available RNA-sequencing datasets

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

2.3.2 Metadata comparison

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

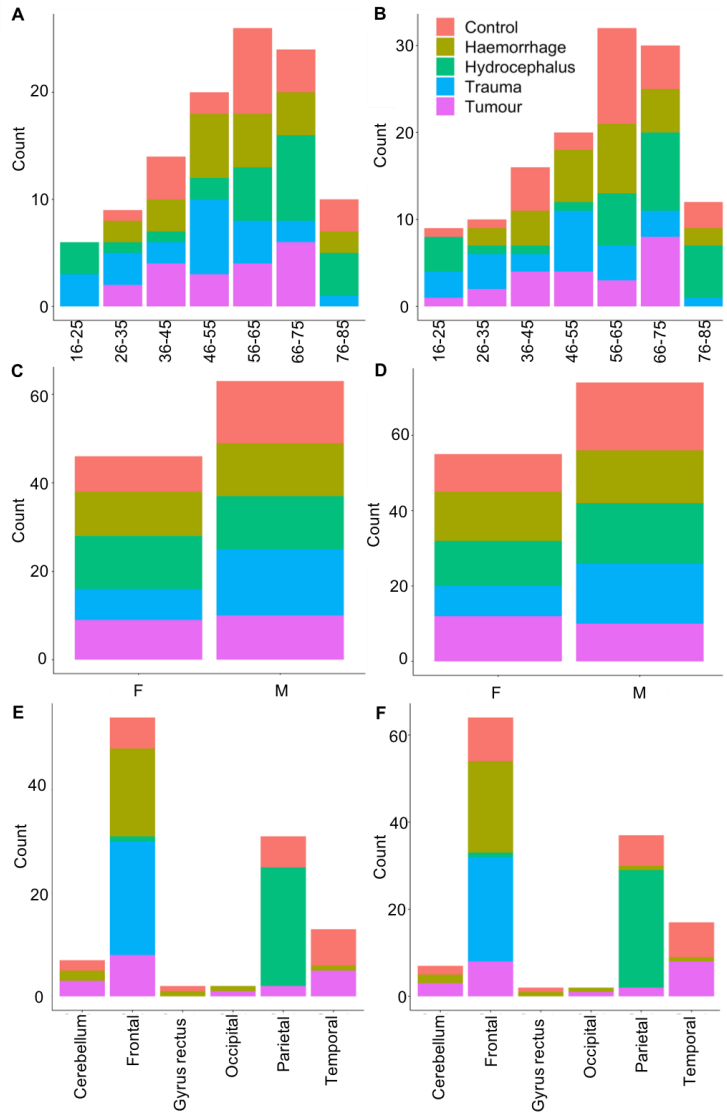

Figure 2.5 Frequency of patients from metadata groups within the bulk (A, C and E) and single cell (B, D and F) RNA-seq datasets

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

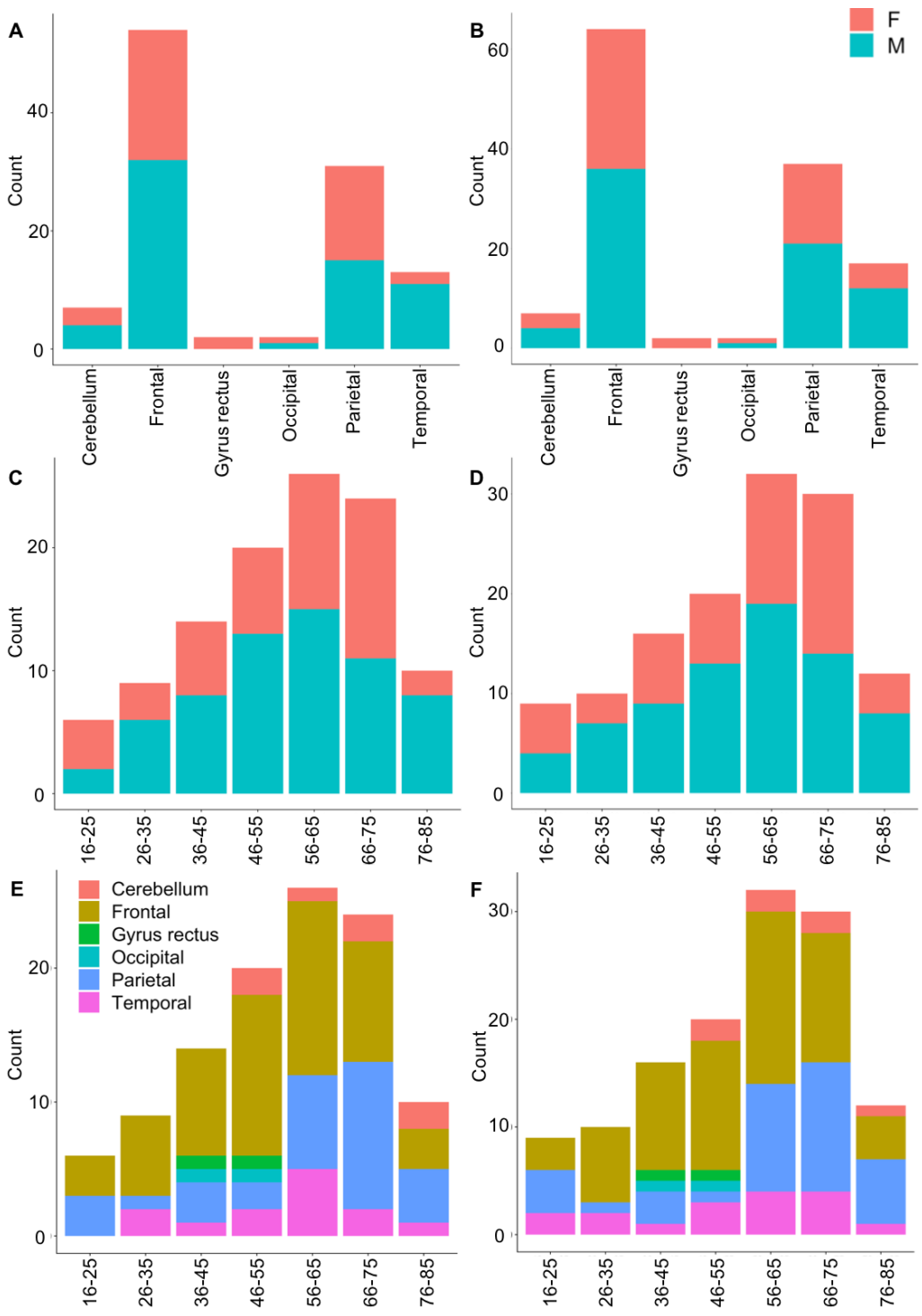

Figure 2.6 Frequency of patients from metadata groups within the bulk (A, C and E) and single cell (B, D and F) RNA-sequencing datasets

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

2.4 Single cell clustering and identification of sub-populations

2.4.1 Comparison to publicly available single cell datasets

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

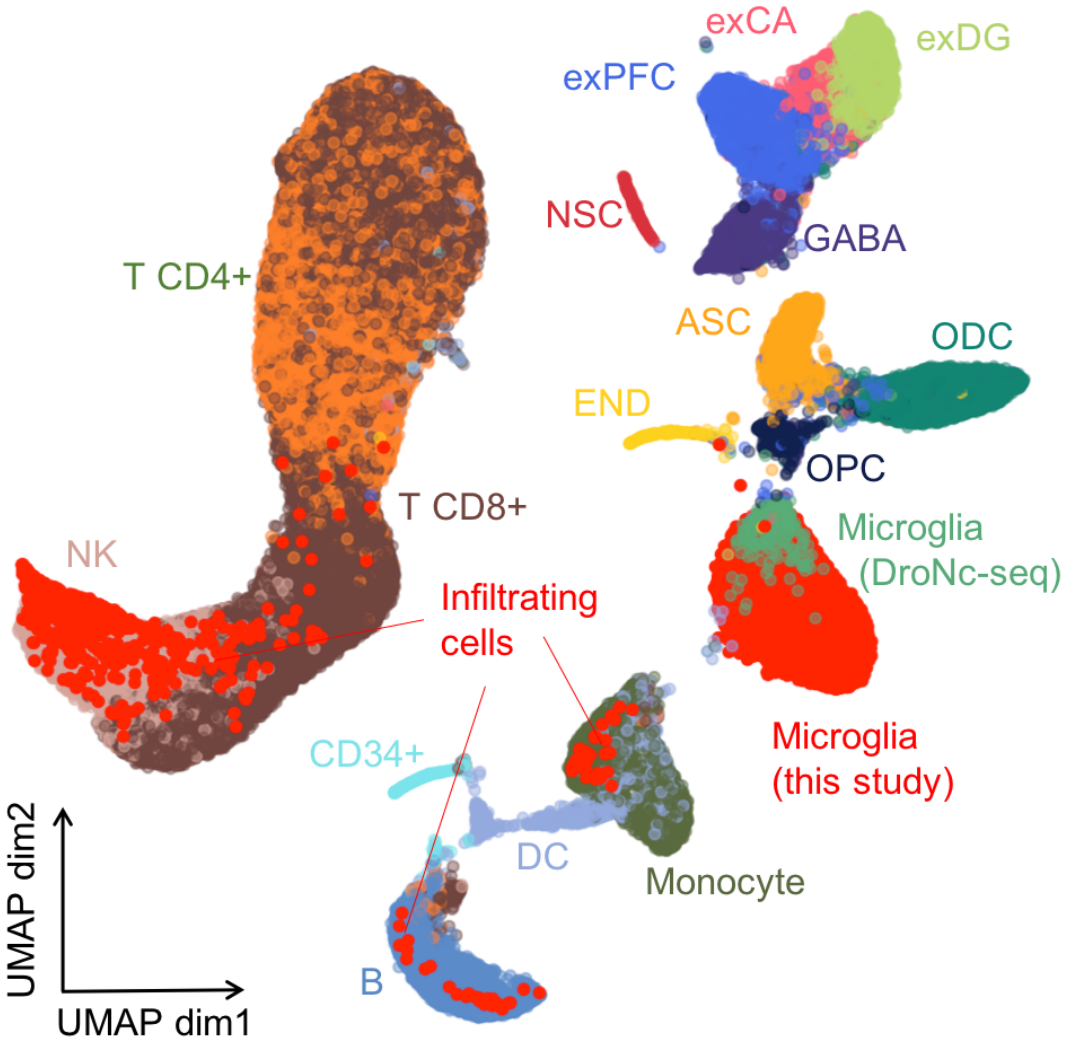

Figure 2.7 UMAP analysis of microglia single cell data and publicly available PBMC and whole brain tissue single cell datasets

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

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

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

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

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

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

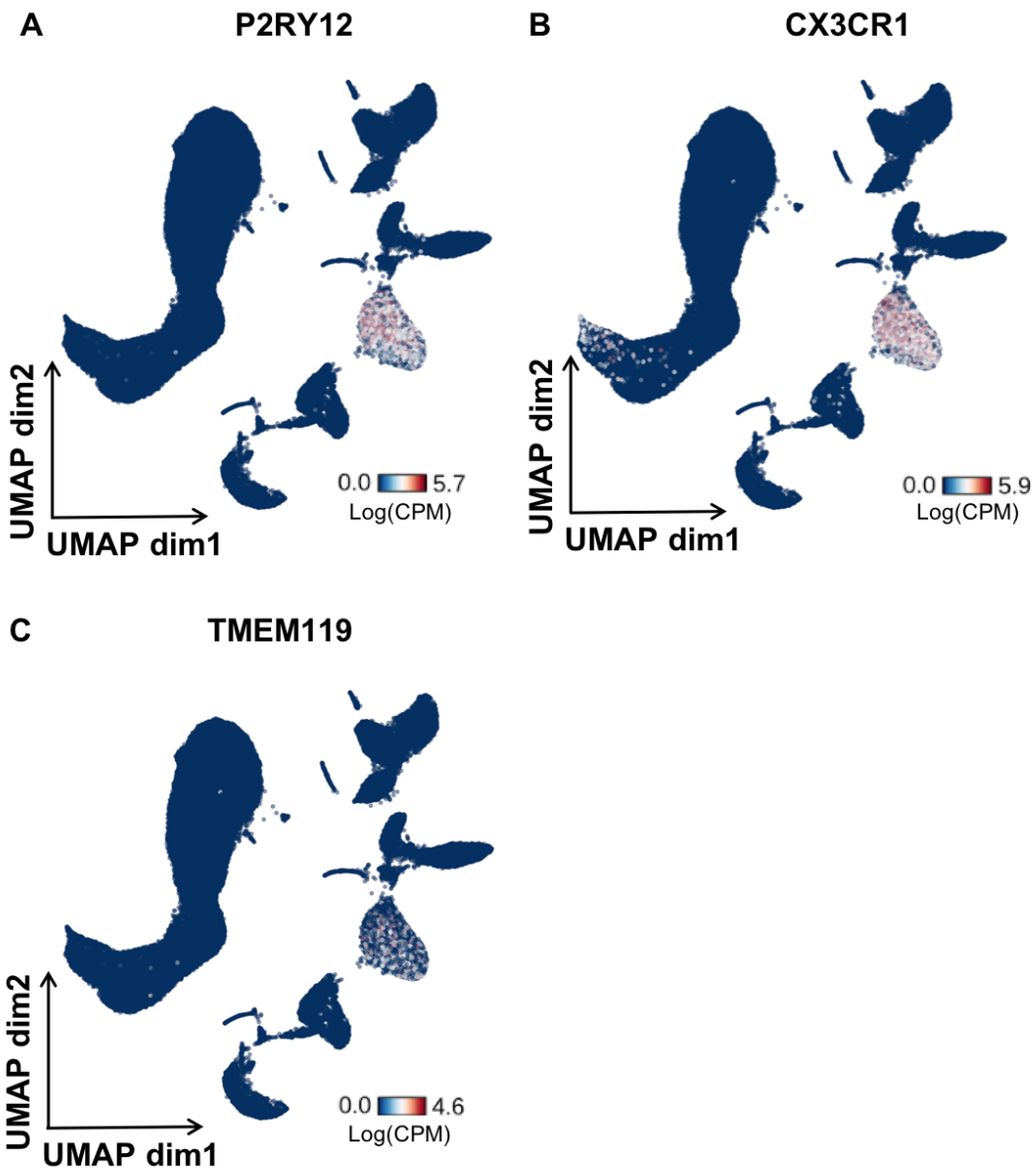

Figure 2.8 UMAP analysis of microglia single cell data and publicly available PBMC and whole brain tissue single cell datasets

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

2.4.2 Clustering of microglial cells and cluster maker analysis

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

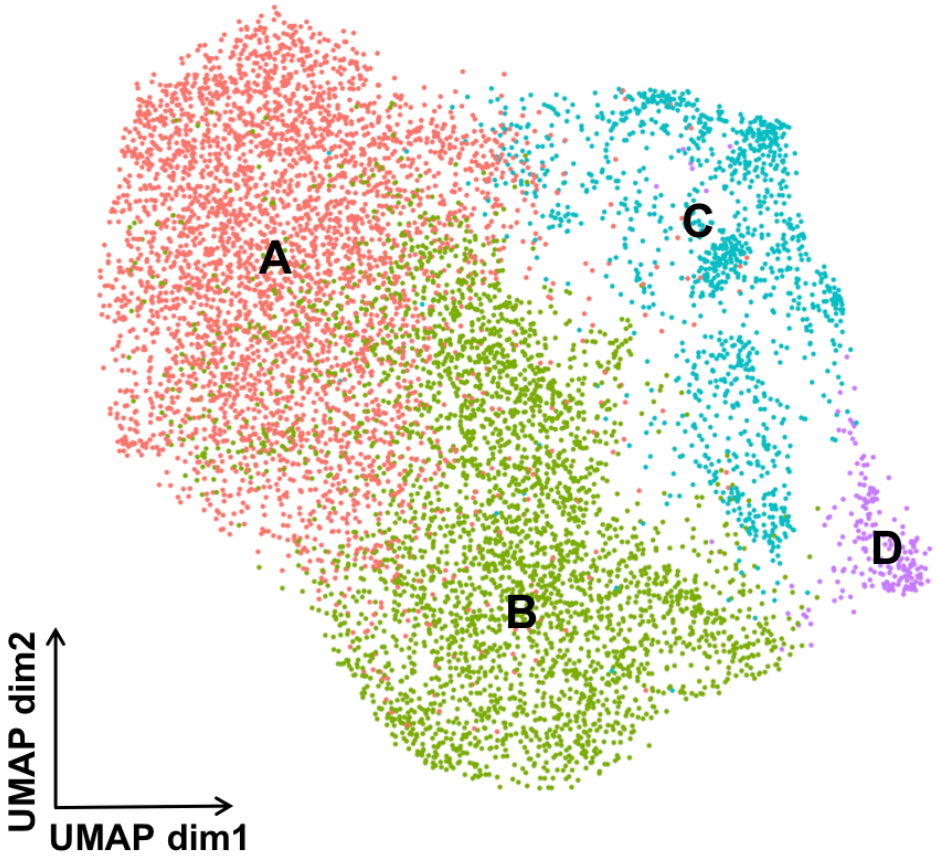

Figure 2.9 UMAP analysis of microglia cells from this study identified from previous analysis (Figure 2.7)

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

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

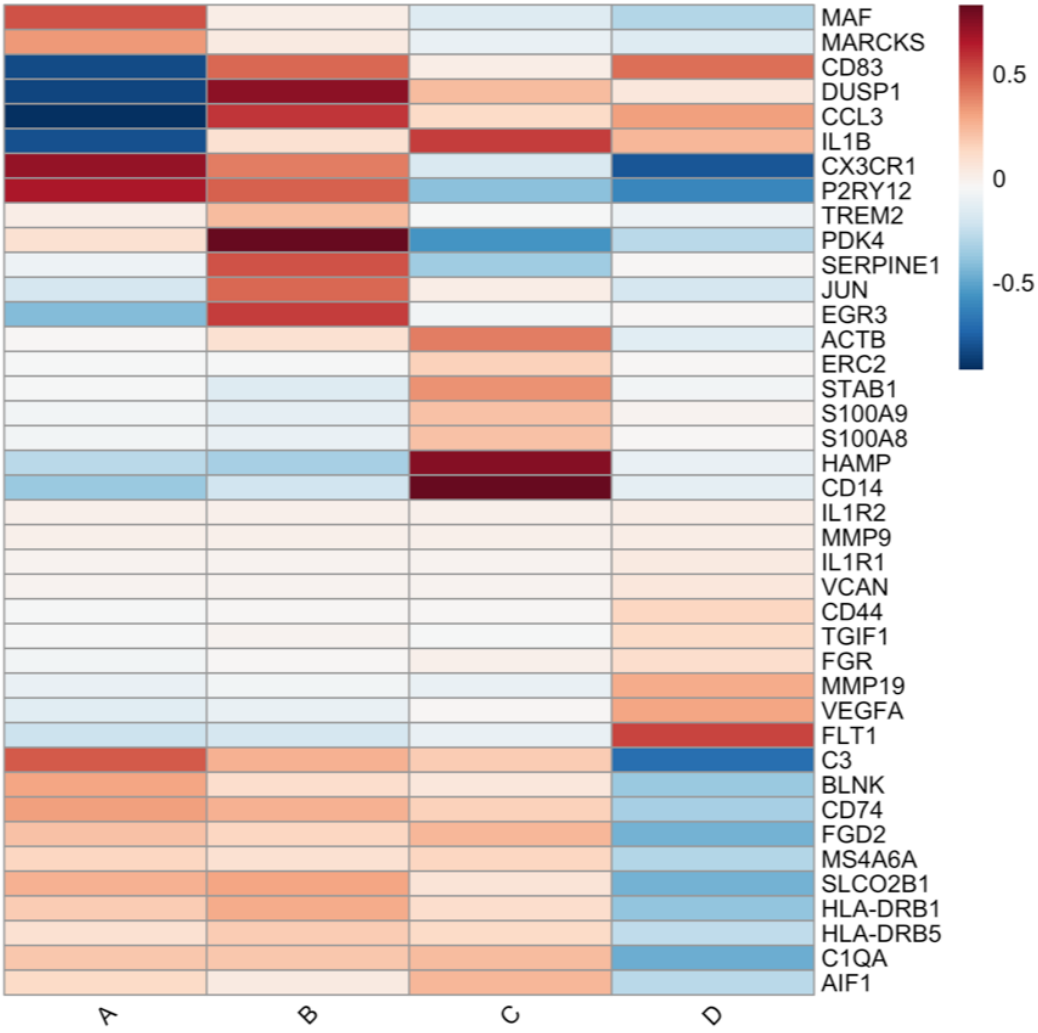

Figure 2.10 Cluster maker genes for microglia single cell data

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

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

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

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

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

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

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

Table 2.3 Top enriched biological process terms for cluster marker genes

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

2.5 Clinical metadata and microglial transcriptome signatures

2.5.1 Variance components analysis

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

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

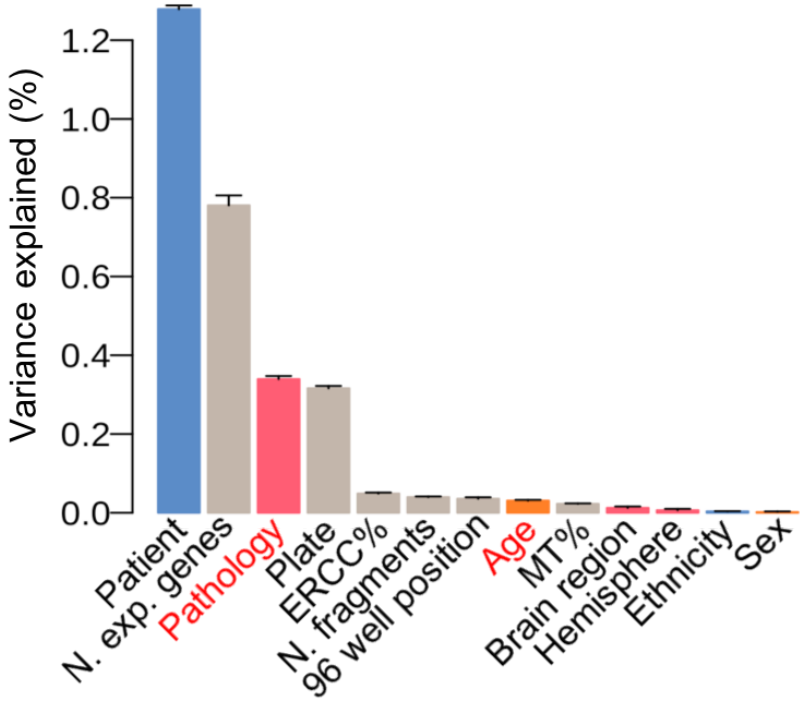

Figure 2.11 Variance components analysis

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

2.5.2 Gene expression linked to clinical metadata

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

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

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

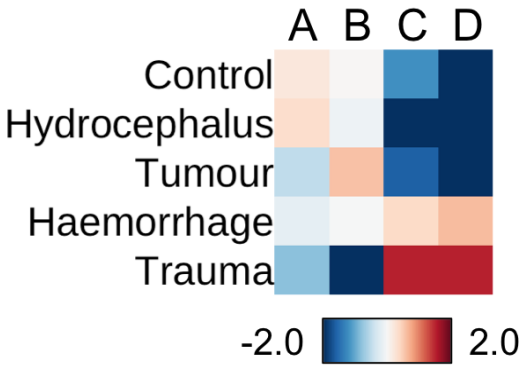

Figure 2.12 Odds ratios from Fisher's exact tests across clinical pathologies for each cluster.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.6 Microglia and disease

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

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

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

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

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

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

2.7 Discussion

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

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

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

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

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