# Transcriptomics of prion diseases.

Dimitriadis Athanasios

Doctor of Philosophy in

Neurodegenerative Diseases

MRC Prion Unit at UCL, Institute of Prion Diseases

University College London

I, Dimitriadis Athanasios, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

# Abstract

Despite substantial research aiming to elucidate prion disease pathogenesis, the underlying mechanisms of cellular toxicity and neurodegeneration remain poorly characterized. The human brain comprises numerous cell populations with a heterogeneous transcriptional landscape, complicating the interpretation of transcriptomic studies. To untangle this complexity, we first established and validated two single-nucleus sequencing methodologies and a bioinformatics pipeline for data analysis. We then designed a time-course case-control study of RML- and control brain homogenate-inoculated FVB mice (N = 95, time points: 20, 40, 80, 120 dpi and disease end-stage), and a human case-control study in post-mortem and biopsied brain samples (N = 26) and applied our transcriptomics pipeline. We generated 210,000 high-quality cell transcriptomes across 5 time points in mice and identified 26 subclusters of cortical neurons, interneurons, mature oligodendrocytes, oligodendrocyte precursor cells, vascular and leptomeningeal cells, and astrocytes. Glial activation was evident from 80 dpi, while our data suggested a selective transcriptomic response of individual cell clusters to disease. We identified a pattern of neuronal transcriptomic change shortly after RML-brain inoculation that quickly resolved, despite rapidly increasing prion titres in the brain, only to return at later stages when the neuropathology of prion disease was evident. Subsequent pathway analyses identified common perturbed biological pathways associated with synaptic dysfunction and ion homeostasis. Our human tissue samples did not pass quality control criteria, highlighting the need for different methodologies to assay archived samples. Here we provide the first single-cell transcriptomics study of prion diseases in mouse which found cell-type and time-specific patterns. Taken together, findings suggest that prion replication itself does not produce a transcriptomic signature in the brain, rather, a transient pattern of toxicity can be seen immediately following inoculation of prion disease brain homogenate, which becomes re-established as prion disease neuropathology develops.

# Impact statement

Prion diseases are fatal neurodegenerative pathologies affecting humans and animals. Sporadic CJD, the most common human prion disease, constitutes 85–90% of all cases however has no known cause to date. Despite substantial research aiming to elucidate prion disease pathogenesis, the underlying mechanisms of cellular toxicity and neurodegeneration are yet to be fully characterised. The transcriptional landscape of the prion-infected human brain, including changes in gene expression profiles related to tissue degeneration, has not been explored in-depth while confounding effects related to cellular heterogeneity have not been accounted for.

This thesis describes the applicability of single-cell RNA sequencing methodologies in human and mouse archived tissue and provides the first look into the single-cell transcriptomics of prion disease. It starts with evaluating droplet-based and combinatorial indexing-based single-nucleus sequencing methodologies and identifying their strengths and weaknesses when used under tightly controlled experimental conditions, especially exploring their suitability when used with BSL-3 material and infectious human prions. We provide a full working pipeline that includes tissue cutting, nuclei suspension preparation, single-cell library preparation, library multiplexing, next-generation sequencing, data manipulation, and single-cell analysis pipelines. Additionally, we provide scripts for usual exploratory analyses, statistical tests, and complex visualisations of single-cell datasets. This information can be of importance when selecting suitable similar methodologies for future experiments, saving time and energy, and simplifying data analysis tasks.

We then performed the first single-cell study of murine prion disease in animal models and generated high-quality datasets across 5 time points, characterising the disease from its earliest stages up to the end-stage. This rich resource includes numerous data visualisations, cell-type information, transcriptomic analyses and gene lists, and longitudinal and case-control comparisons coupled with additional modalities including information on immunohistochemical observations, prion infectivity, and spatially-resolved transcript expression. This opens new avenues to be exploited by future researchers to provide answers to follow-up scientific questions, facilitate the design of future targeted experiments, act as an example of correctly controlled experimental design, and raise subsequent questions and stimulate curiosity.

Our human experiments, although they did not generate useful datasets, proved that different methodologies need to be applied for archived human brain tissue and suggested ways of carefully designing and controlling similar experiments, saving time and energy from future researchers. In addition, this thesis discusses a plethora of future directions for follow-up studies that could contribute to elucidating interesting hypotheses generated through our work.

Through sharing our findings in international conferences, aiming to publish our research and make our datasets freely available, we hope that our findings will echo in the scientific community and accelerate similar innovative studies in the field of prion diseases, which is currently lagging behind other neurodegenerative disorders in terms of transcriptomics research. Finally, since common mechanisms are increasingly being identified in prion and other neurodegenerative diseases, we are confident that the work described in this thesis can have a broad impact on the wider neurodegeneration research field.

# Table of contents

## List of figures

## List of tables

# List of abbreviations

| **Abbreviation** | **Description** |
|---|---|
| AD | Alzheimer's disease |
| ALS | Amyotrophic lateral sclerosis |
| bp | Base pairs |
| BSE | Bovine spongiform encephalopathy |
| cDNA | Complementary DNA |
| CJD | Creutzfeldt-Jakob disease |
| CNS | Central nervous system |
| CPM | Counts per million |
| CWD | Chronic wasting disease |
| DAM | Disease-associated microglia |
| DE | Differential expression |
| DEG | Differentially expressed gene |
| DGE | Differential gene expression |
| EAE | Experimental autoimmune encephalomyelitis |
| FFI | Fatal familial insomnia |
| GO | Gene ontology |
| GSEA | Gene set enrichment analysis |

| | |
|---|---|
| GSS | Gerstmann-Sträussler-Scheinker disease |
| GWAS | Genome-wide association study |
| HVG | Highly variable gene |
| kDa | Kilodalton |
| KEGG | Kyoto encyclopaedia of genes and genomes |
| lncRNA | Long non-coding RNA |
| miRNA | Micro-RNA |
| MS | Multiple sclerosis |
| NGS | Next-generation sequencing |
| nt | Nucleotides |
| OPC | Oligodendrocyte progenitor cell |
| ORA | Over-representation analysis |
| PCA | Principal component analysis |
| PCR | Polymerase chain reaction |
| PK | Proteinase K |
| PMCA | Protein misfolding cyclic amplification |
| $PrP^{C}$ | Cellular prion protein |
| $PrP^{Sc}$ | Scrapie-associated prion protein |
| QC | Quality control |

| | |
|---|---|
| RIN | RNA integrity number |
| RML | Rocky Mountain laboratory |
| RNA-seq | RNA sequencing |
| rpm | Rounds per minute |
| rRNA | Ribosomal RNA |
| RT-PCR | Real-time polymerase chain reaction |
| RT-QuIC | Real-time quaking-induced conversion |
| SCA | Scrapie cell assay |
| SCEPA | Scrapie cell assay in endpoint format |
| sCJD | Sporadic Creutzfeldt-Jakob disease |
| scRNA-seq | Single-cell RNA sequencing |
| snRNA-seq | Single-nucleus RNA sequencing |
| t-SNE | t-stochastic neighbour embedding |
| UMAP | Uniform manifold approximation and projection |
| UMI | Unique molecular identifier |

# 1 Introduction

## 1.1 Prion diseases

Prion diseases are fatal neurodegenerative pathologies affecting humans and animals. They are attributed to a conformational change of the cellular prion protein ($PrP^{C}$) to disease-associated forms, including the protease-resistant $PrP^{Sc}$, which has been identified as the causative agent of scrapie (Prusiner, 1982), the first prion disease documented in sheep. Other animal prion diseases that have been described include Bovine Spongiform Encephalopathy (BSE) in cattle, Transmissible Mink Encephalopathy, Feline Spongiform Encephalopathy in cats, Exotic Ungulate Encephalopathy in nyala and kudu, Chronic Wasting Disease in cervids and Primate Transmissible Encephalopathy in lemurs (Imran & Mahmood, 2011). Human prion diseases include Creutzfeldt-Jakob Disease (CJD), Gerstmann-Sträussler-Scheinker disease (GSS), Fatal Familial Insomnia (FFI) and kuru (J Collinge, 2001).

Human prion diseases can be divided into three groups, based on their aetiology: inherited, sporadic, and acquired. Inherited, or familial prion diseases are caused by mutations in the *PRNP* gene, which encodes the human prion protein, and include GSS, familial CJD, and FFI. However, only a small percentage (10-15%) of prion diseases are attributed to genetic mutations with autosomal dominant inheritance pattern (Prusiner & Hsiao, 1994). Acquired prion diseases are caused by the transmission of prions through surgical procedures and grafts (iatrogenic CJD; iCJD), mortuary feasts (kuru) or consumption of BSE-infected food products (variant CJD; vCJD). Sporadic human prion diseases have no known cause to date and include sporadic CJD (sCJD), fatal insomnia and variably protease-sensitive prionopathy (Imran & Mahmood, 2011).

Sporadic CJD, the most common human prion disease, constitutes 85–90% of all cases (Brown et al., 1994; Masters et al., 1979) with an annual occurrence of 1-2 cases per million (Ladogana et al., 2005). It affects both sexes with the same rate and the peak onset is between 55 and 75 years. Typical clinical features include dementia, visual abnormalities, muscle incoordination and gait and speech abnormalities. Pyramidal and extrapyramidal dysfunction and behavioural changes can also develop during the course of the disease. Characteristic is the rapid deterioration of the symptoms, while during the terminal stages of the disease patients develop a state of akinetic mutism (Belay, 1999). While the cause of the disease has not been elucidated yet, one of the hypotheses that exist involves the stochastic appearance of a rare somatic mutation that might result in the conversion of the normal PrP to a pathogenic isoform. This view is supported by the fact that rare mutations can occur in the population at rates comparable to the incidence of sporadic CJD (Bomba et al., 2017). Another hypothesis suggests that random misfolding of the normal prion protein to the pathogenic isoform can occur in a single neuron or a group of cells, possibly during the transcription or translation of the *PRNP* gene and trigger a cascade (Belay, 1999). No consistent risk factors have been reported for sCJD, except for age and genetic variation at the human *PRNP* gene (Simon Mead et al., 2012), while a recent Genome-Wide Association Study (GWAS) identified two new and replicated risk variants in *STX6* and *GAL3ST1* and two further unreplicated loci that were significant in genome-wide tests (*PDIA4*, *BMERB1*) (Jones et al., 2020). Some studies have also implicated surgical procedures (Ward et al., 2008) as a possible means of contamination, but the evidence remains inconclusive (Hamaguchi et al., 2009; Harries-Jones et al., 1988).

An interesting phenomenon of prion biology is the existence of different prion strains, which are defined as infectious isolates that exhibit distinct disease phenotypes when transmitted to identical hosts (Aguzzi et al., 2007). Strain differences can influence phenotypic traits, including disease incubation time and distribution of brain lesions, as well as the PrP$^{Sc}$ biochemical profile (Solforosi et al., 2013). Information relevant to prion strain specificity is believed to be encoded at the level of protein conformation, following the protein-only hypothesis that dictates that a misfolded prion protein is the essential causative agent of prion disease and transmission (Bessen et al., 1995; Tanaka et al., 2004; Telling et al., 1996). Studies of the strain phenomenon have employed experimental animals inoculated with infectious material from various species. Among them, mouse models have been particularly useful having been used to study and isolate more than 20 different prion strains (Bruce, 1993). Strains RML, ME7, 79A, 22L, Chandler, 301V, and 139A are all mouse-adapted scrapie prion strains, while S15 refers to the lysate of the SMB-S15 cell line originally established when it was cultured from a Chandler isolate-infected mouse brain (Birkett et al., 2001; Bruce, 1993; Chandler, 1961).

These strains present similar biochemical characteristics but can be differentiated when inoculated in mice by studying the characteristic profile of brain lesions and disease incubation time (Bruce et al., 1991; Fraser & Dickinson, 1973; Legname et al., 2005). In addition, it has been shown that different prion strains can elicit a distinct transcriptomic response, a phenomenon which, coupled with host genetics, complicates the interpretation of transcriptomics studies and highlights the need for tightly-controlled experimental conditions (Hwang et al., 2009).

Genotyping of the *PRNP* gene is a vital component of describing CJD strains. Of special interest is a polymorphism at codon 129 that encodes either a methionine (M) or a valine (V) residue and is considered to be an important disease modifier. Methionine homozygosity confers increased susceptibility to the sporadic, variant, and iatrogenic forms of CJD, while the three possible genotypes are associated with different clinical phenotypes (J Collinge et al., 1991; Andrew F Hill et al., 2003; Palmer et al., 1991; Zeidler et al., 1997).

The analysis of PrP$^{Sc}$ characteristics after limited proteinase K digestion of infected brain homogenates can also be used to study prion strains and is a conventional component of CJD classification. Digestion with proteinase K yields three distinct bands when electrophoresed that are associated with the three possible degrees of glycosylation of the PrP protein: unglycosylated, mono-glycosylated, or di-glycosylated (J Collinge et al., 1996; Parchi et al., 1996). The codon 129 status and the main PrP$^{Sc}$ band patterns after PK digestion are used in the two main CJD classification systems, referred here to as the "Italian classification system" and the "London classification system". The Italian system refers to "type 1" PrPSc with an unglycosylated band at 21 kDa, and "type 2" PrPSc with a band at 19 kDa (Parchi et al., 1996, 1999). In contrast, the London system refers to the banding pattern of 21 kDa as "type 2", the band of 19 kDa as "type 3", and a band of 21.5 kDa as "type 1" (J Collinge et al., 1996; Andrew F Hill et al., 2003). The banding patterns for vCJD are separately recognised by both systems — "type 2B" in the Italian and "type 4" in the London system.

Evidence suggests that a growing number of proteins involved in neurodegeneration share certain characteristics with prions. This led to the introduction of the prion paradigm which holds that the fundamental cause of specific neurodegenerative disorders is the misfolding and seeded aggregation of certain proteins (Neves, 2019; Safar, 2016; L. C. Walker & Jucker, 2015). It is becoming increasingly clear that Alzheimer's disease, Parkinson's disease, frontotemporal dementia, and polyglutamine diseases share so-called prion-like mechanisms, including strain properties (B. Frost & Diamond, 2010; Wemheuer et al., 2017). Similar to prionopathies, all of these diseases are associated with the accumulation of fibrillar aggregates of proteins (amyloid-β, tau, a-synuclein, and polyglutamine proteins). These observations have led some researchers to generalise the use of the term "prion" to refer to any alternatively folded protein undergoing self-propagation and sharing key biophysical and biochemical characteristics with PrP prions, categorising progressive supranuclear palsy and multiple system atrophy as prion diseases (Woerman et al., 2015). In addition, Jaunmuktane et al. and Purro et al. have made a strong case for iatrogenic human transmission of amyloid-β pathology similar to prion transmission, with implications for both the treatment and prevention of AD (Jaunmuktane et al., 2015; Purro et al., 2018). Overall, these more recent developments underline the importance of studying prion diseases and the associated mechanisms as our findings can have broader implications in the field of neurodegenerative disorders.

## 1.2 Transcriptomics of prion diseases

### 1.2.1 Microarray-based studies

While the genome can provide information about the heritability of a disease, it does not capture the dynamics that regulate the balance between normal and pathological states. The interplay of gene function through parallel expression measurements of the same genetic targets constitutes the core principle of functional genomics. Some of the most renowned technologies used to carry out transcriptional profiling are based on DNA microarrays. After the description of the DNA double helix structure by Watson and Crick in 1953 (Watson & Crick, 1953), scientists soon realised the potential of molecular hybridisation. The simple fact that single-stranded DNA binds to complementary DNA and the existence of complementary base pairs that form the structure of the two strands of DNA laid the foundations for analytical methods of DNA sequencing, including DNA microarrays. Grunstein and Hogness produced an early example of what can be broadly considered a DNA array by introducing colony hybridisation, a technique that could help isolate cloned DNA that contains a specific gene (Grunstein & Hogness, 1975). In their experiments, they used a nitrocellulose filter to imprint bacterial colonies cultured on a plate. The bacteria were then lysed, and their DNA denatured and fixed on the filter in situ. The resulting DNA print could then be hybridised to radioactive labelled RNA, complementary to the sequence of interest, and the result could be assessed by autoradiography. The colonies that contained the gene of interest could be identified and then isolated from the original reference plate. During the following decades, further technological advancements coupled with the development of robotic technology and an increase in automation led to the introduction of the first miniaturised array - the “microarray” - in 1995, when Schena et al. measured the expression of 45 Arabidopsis genes in parallel in an array prepared by robotic printing of complementary DNA strands on glass (Schena et al., 1995). Since then, the fundamentals of microarrays have remained the same, while their capacity and efficiency have increased, making them an invaluable tool for molecular biology and, more specifically, transcriptomics.

While the maturation of the microarray sequencing technologies was a key factor that catalysed further experiments, of equal importance was the selection of appropriate animal models. While many animal models have been used to study animal Transmissible Spongiform Encephalopathies, most of the research involved rodent models due to their relatively short disease incubation time, easier maintenance than larger animals, extensive genetic characterisation, and relatively easy genetic manipulation. In addition, the passage and introduction of various goat prion strains in mouse models have allowed the investigation of strain biology under tightly controlled experimental settings.

Many studies have identified sets of perturbed genes during the early or late stages of the disease. For example, Booth et al. used cDNA microarrays to query gene expression profiles at three time points after inoculation (early, middle/preclinical, and late/clinical) of C57BL/6 mice with two mouse-adapted prion strains, ME7 and 79A (Booth et al., 2004). Most of the significantly differentially expressed genes (138 upregulated and 20 downregulated) were found in the clinical stage of the disease. A gene ontology analysis revealed that the biological processes involved include cell communication, transport, development, cell organisation and biosynthesis and others. Only a smaller set of genes were shown to be dysregulated in the early and middle stages of the disease. Interestingly, one set of genes was found to be downregulated in all tested time points, including four transcripts of genes related to the haematopoietic system, suggesting that haemopoiesis might be involved in the disease process from the early stages. Similar findings were reported by Skinner et al. in 2006, where C57BL/10 mice were inoculated with three prion strains: ME7, 22L, and Chandler/RML (Skinner et al., 2006). In that study, over 400 differentially expressed genes (DEGs) were identified in symptomatic mice, while only 22 genes were found to be significantly altered in the pre-symptomatic animals. Differences were also evident in the expression profiles of mice inoculated with different strains, underlining the heterogeneity of the transcriptomic response to different prion strains. These genes implicate cellular processes including protein folding, lysosome function, synapse function, metal ion binding, calcium regulation and cytoskeletal function.

As clinical stages of the disease might be dominated by perturbations resulting from the extensive pathology, it is crucial to focus on the early disease stages to understand disease pathogenesis and develop diagnostic assays. Following this paradigm, Kim et al. used Affymetrix microarrays to identify DEGs in the spleen and brain of intracerebrally-inoculated prion-infected mice during the early stages of the disease (H. O. Kim et al., 2008). They found 67 upregulated genes in the infected mice, prior to the onset of clinical symptoms. These were involved in many biological processes including immunity, the endosome/lysosome system, hormone activity and the cytoskeleton. More importantly, they identified 14 genes that were shown to be altered in expression in the spleen, before the onset of clinical symptoms; four of them (*Atp1b1*, *Gh*, *Anp32a*, and *Grn*) were altered only 46 days post-infection, entertaining the possibility of serving as surrogate markers for disease diagnosis.

A systems approach to studying prion diseases was adopted by Hwang and colleagues (Hwang et al., 2009). The researchers postulated that the disease emerges due to the perturbation of multiple and interconnected transcriptional targets which form biological networks in the brain. They used microarray technology to profile the global gene expression in the brains of mice from six different genetic backgrounds (B6, B6.I, FVB) that were inoculated with either one of two different prion strains (RML or 301V). This strategy provided data from 8 distinct mouse strain-prion strain combinations at 8-10 time points. This comprehensive experimental design allowed the investigation of the effects of host genetics, prion strain and PrP concentration on disease incubation time and transcriptomics. Subtractive analyses identified 333 DEGs that were commonly dysregulated and appeared to be central to the disease (Figure 1.1). The authors then integrated gene expression data with information regarding pathology, aggregated PrP deposition, gene ontology and protein interactions to generate protein networks that seemed to be related to disease pathology. Further grouping of the mice according to prion strain and disease incubation time revealed 39 DEGs associated with the RML strain, and 55 DEGs associated with short incubation time, respectively. The researchers concluded that their research highlights the power of systems approaches and provides insights that could potentially shape novel disease diagnosis and treatment approaches.

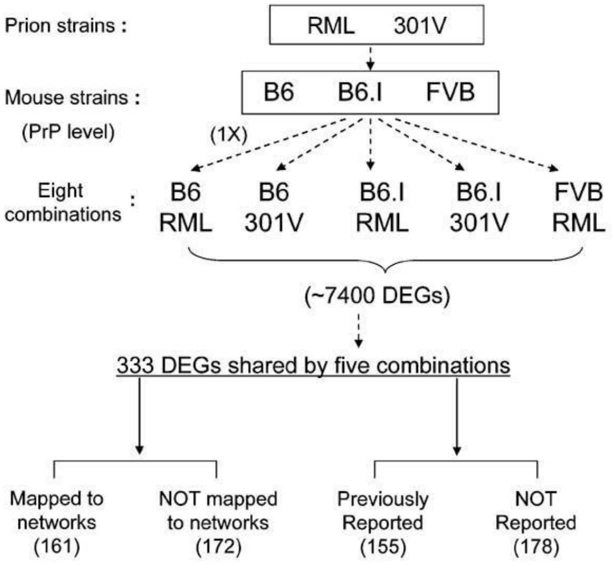


***Figure 1.1: Strategies for the identification of 333 core DEGs in mouse prion diseases.*** *The authors used microarray technology to profile the global gene expression in the brains of mice from three different genetic backgrounds (B6, B6.I, FVB) that were inoculated with either one of two different prion strains (RML or 301V). A subtractive analysis identified 333 DEGs that were commonly dysregulated and appeared to be central to the disease. Of those, 161 genes could be mapped to perturbed biological gene networks, while 178 were reported for the first time. Figure adapted from Hwang et al. 2009.*

Another approach to studying prion pathology while controlling for changes due to neuroinflammation is based on the use of cuprizone, a prion disease mimetic drug. This copper-chelating compound, when given orally for several weeks, causes chronic astrocytosis and spongiform changes that qualitatively mimic the cell population changes observed in prion diseases. Moody and colleagues used cuprizone-treated animals as experimental controls and compared their brain expression profiles to RML-inoculated mice in preclinical and clinical time points (Moody et al., 2009). Their study identified 164 DEGs during prion infection versus non-treated controls, while 307 transcripts were found to be differentially regulated in the cuprizone-treated mice versus non-treated controls. More importantly, a comparative analysis between the prion-infected and the cuprizone-treated mice identified 17 transcripts that are not affected by the drug but increase in expression from preclinical to clinical prion infection. Nine of these genes (*Hsbp1*, *Socs3*, *Ifi44*, *D12Ertd647e*, *Casp4*, *Agrp*, *Plce1*, *Ptpbl* and *Ddx58*) were found to be upregulated preclinically and could provide insight into disease progression.

Except for coding genes, recent studies have established links between micro RNAs (miRNAs) and neurodegenerative diseases, including prion diseases. miRNAs are small non-coding transcripts that are involved in fine-tuning gene expression by post-transcriptionally regulating mRNA stability. Saba et al. identified 15 miRNAs that were dysregulated in prion-infected mice using microarrays and RT-PCR (Saba et al., 2008). A group of 7 miRNAs (*miR-342-3p*, *miR-320*, *let-7b*, *miR-328*, *miR-128*, *miR-139-5p* and *miR-146a*) were shown to be over 2.5-fold upregulated, while another group of 2 miRNAs (*miR-338-3p* and *miR-337-3p*) were shown to be over 2.5-fold downregulated. Computational analyses identified potential gene targets, including 119 genes that have previously been reported to be dysregulated in mouse scrapie. These gene targets were found to be involved in intracellular protein-degradation pathways and signalling pathways related to cell death, synapse function and neurogenesis. More recently, Majer et al. used laser capture microdissection to isolate hippocampal CA1 neurons from mice infected with the RML prion strain and determine their preclinical transcriptional response during infection (Majer et al., 2012). Interestingly, it was found that a major cluster of genes was dysregulated in the preclinical disease stage, while expression returned to basal levels or was reversed during the clinical stage. Dysregulated miRNAs *miR-132-*

*3p*, *miR-124a-3p*, *miR-16-5p*, *miR-26a-5p*, *miR-29a-3p* and *miR-140-5p* exhibited this alternating pattern of expression.

While mouse studies can be pivotal in uncovering the underlying transcriptomic signature of the disease, they do have limitations and require a considerable amount of time from inoculation to results. To address some of these limitations, researchers have also focused on murine cell lines. Murine cells have been shown to propagate prions in vitro while providing a tightly controlled setup that allows for more experimental freedom at the cost of less generalisable conclusions. Greenwood et al. used Neuroblastoma (N2a) and hypothalamic neuronal cells (GT1), which can be persistently infected with mouse scrapie prions, to contrast the transcriptional landscape of these cell lines when infected with prion strain RML (Greenwood et al., 2005). It was reported that the RNA profiles of the infected cells (ScN2a and ScGT1) show differences between them and between gene expression changes reported in human microglia and brain studies, while there was some overlap. In addition, curing the ScN2a cells with pentosan polysulphate led to the reversion of only some differentially expressed genes. The authors argue that this evidence supports the hypothesis that the same prion strains might have a different transcriptomic impact on different cells. Contradictory evidence was published in the same journal by Julius and colleagues 3 years later when the researchers analysed the transcriptional response of three murine neural cell lines to persistent prion infection in vitro (Julius et al., 2008). The cell lines (N2aPK1, CAD, and GT1) were infected with prion strain RML and infectivity was validated by colony spot immunochemistry (64-100% of the cells were infected). This study only found the *Nav1* gene marginally modulated in only one cell line, while no other transcript was significantly altered. The authors attribute the results to the experimental stringency of the study, which was designed to minimise genetic drift. More recently, Marbiah et al. screened prion-resistant revertant clones, isolated from susceptible PK1 cells (Marbiah et al., 2014). Their transcriptomic signature, which was associated with susceptibility and differentiation, included several genes that encode proteins involved in extracellular matrix remodelling such as fibronectin 1 (*Fn1*) and integrin α8 (*Itga8*), a cellular component in which disease-related PrP is deposited. Finally, silencing nine of these genes was able to increase prion infection susceptibility.

Human studies are more limited in number, possibly owing to the low prevalence of the disease that limits the number of samples available, the infectious nature of human prions that constrains the methodology that can be used, and the more complex experimental design due to the inherent variability of human brain tissue. One of the early studies using human brain tissue was published in 2005 by Xiang and colleagues and used Affymetrix microarrays to compare the transcriptome of the frontal cortex of 15 sCJD patients to 5 normal controls (Xiang et al., 2005). After stringent quality control, they identified 79 upregulated and 275 downregulated genes. The upregulated genes were found to be coding immune and stress-response factors and elements involved in cell death and cell cycle, while the downregulated genes mostly encoded synaptic proteins. Interestingly, the degree of increased expression was found to be correlated with the degree of neuropathological alterations in particular molecular subtypes of sCJD. Later, Tian and colleagues analysed the global expression patterns of the thalamus and parietal cortex of three FFI patients (Tian et al., 2013). They found a total of 1314 DEGs in the thalamus and 332 in the parietal lobe. 255 of those genes were shown to be modulated in the same direction in both regions (99 upregulated and 156 downregulated). The most significantly altered molecular functions included transcription and DNA-dependent regulation of transcription, RNA splicing, and mitochondrial electron transport. A KEGG pathway analysis identified 102 pathways that were changed in both brain areas. A year later, the same group published a broader study that included sCJD and AD patients (11 sCJD, 3 FFI, 3 AD, and 4 normal controls) (Tian et al., 2014). By analysing the overlap of the differential expression data in all 3 neurodegenerative disorders, they were able to identify common dysregulated biological processes (signal transduction, synaptic transmission, and neuropeptide signalling pathway) and pathways (MAPK signalling pathway, Parkinson’s disease, and oxidative phosphorylation) (Table 1).

| sCJD1 | sCJD2 | FFI | AD |
|---|---|---|---|
| Alzheimer’s disease | MAPK signalling pathway | MAPK signalling pathway | MAPK signalling pathway |
| Glutamate metabolism | Alzheimer’s disease | Regulation of autophagy | Alzheimer’s disease |

| | | | |
|---|---|---|---|
| MAPK signalling pathway | Parkinson’s disease | Epithelial cell signalling in Helicobacter pylori infection | Parkinson’s disease |
| Calcium signalling pathway | Oxidative phosphorylation | Parkinson’s disease | Oxidative phosphorylation |
| Oxidative phosphorylation | Taurine and hypotaurine metabolism | Oxidative phosphorylation | Focal adhesion |
| Phosphatidylinositol signalling system | Focal adhesion | Reductive carboxylate cycle (CO2 fixation) | Amyotrophic lateral sclerosis (ALS) |
| Parkinson’s disease | Amyotrophic lateral sclerosis (ALS) | Glyoxylate and dicarboxylate metabolism | Epithelial cell signalling in Helicobacter pylori infection |
| Regulation of actin cytoskeleton | Glutamate metabolism | Focal adhesion | Renal cell carcinoma |
| Taurine and hypotaurine metabolism | Regulation of actin cytoskeleton | Regulation of actin cytoskeleton | Melanoma |
| Citrate cycle | T cell receptor signalling pathway | Urea cycle and metabolism of amino groups | Calcium signalling pathway |

***Table 1: Top 10 changed biological pathways in sCJD, FFI and AD identified by Tian et al. 2014.*** *The researchers used Affymetrix Human Genome microarrays to profile the transcriptome of 3 FFI, 11 sCJD, and 3 AD patients. The sCJD patients were further split into two groups of sCJD1 and sCJD2, with more and less* $PrP^{Sc}$ *accumulation, respectively. The pathways Alzheimer’s disease, regulation of actin cytoskeleton, and focal adhesion were found to be perturbed in 3 groups, while the pathways MAPK signalling, oxidative phosphorylation, and Parkinson’s disease were found to be perturbed in all three neurodegenerative diseases. Table adapted from Tian et al. 2014.*

Taking into consideration the aforementioned, it becomes evident that prion diseases are associated with transcriptional perturbations; these have been identified from the earliest to the latest stages of the disease. Differences in experimental approaches have highlighted a multifaceted transcriptional response, where disease stages, host genetics, prion strains and experimental models all contribute to transcriptomic variability. Most DEGs were identified in the clinical disease stages, possibly stemming from an underlying extensive transcriptomic disruption, implicating biological mechanisms associated with protein folding and stress responses, lysosomal and immune function, and cell death. Numerous studies have tried to identify early perturbations both to facilitate disease diagnosis and pinpoint interesting drug targets. Some of these associated biological pathways were found to be associated with lysosomal function, cytoskeleton remodelling and iron, steroid, and prostaglandin metabolism. Cell studies have identified the extracellular matrix remodelling pathway as a susceptibility signature, while human studies have identified perturbations of mechanisms modulating transcription, cell signalling and oxidative phosphorylation. Interestingly, it has been suggested that part of this modulation might be due to the dysregulation of miRNAs, which have been shown to adopt alternating expression patterns during disease progression. While these studies were the first to shed light on the complexity of the transcriptomics of prion diseases, their insights have, unfortunately, failed to provide concrete evidence concerning disease mechanisms. Novel technical approaches, better-controlled experiments, and more consistent work with strains would later allow an even deeper exploration of the transcriptome, in an effort to identify pieces of the puzzle that microarrays might have missed.

### 1.2.2 Next-generation sequencing-based studies

While microarray technology has been in the spotlight for years, technological advancements and a reduction of sequencing cost over the last decade have led to a shift in transcriptomics from microarray technology, which can only quantify specific and finite targets, towards nucleic acid sequencing. DNA sequencing, in general, is the process of determining the sequence of nucleotides in a section of DNA. A major difference between microarrays and sequencing is that the former can only be used for known, predetermined features that need to be printed on the array in advance, while sequencing does not require a priori knowledge, allowing the discovery of novel targets. The first generation of commercialised DNA sequencing was Sanger sequencing. While being a breakthrough of its time, Sanger sequencing offers very low throughput, albeit with high precision. Further technological discoveries led to the introduction of a group of techniques that offer orders of magnitude higher throughput by sequencing DNA fragments in parallel. These technologies are commonly described by the umbrella term “Next-Generation Sequencing'' (NGS) and include the next two generations of sequencing methods (i.e. the second and third) (Feng et al., 2015; Płoski, 2016). Second-generation instruments require clonal amplification of DNA molecules (GS FLX+, SOLiD, Ion, HiSeq etc.), while third-generation technology enables sequencing at the single-molecule level (Helicos and Pacific Biosciences instruments). Some argue about the existence of fourth-generation sequencing technology, but its definition is not widely accepted yet (Feng et al., 2015; Ke et al., 2016; Suzuki, 2020).

Most relevant to transcriptomic studies is RNA sequencing (RNA-seq), which has become an indispensable tool for studying the many distinct aspects of RNA biology, from gene expression to translation and structure. Most RNA sequencing approaches use a second-generation sequencing methodology that involves cDNA synthesis before sequencing and are considered NGS. RNA-seq is considered to be more sensitive than microarrays, has a broader dynamic range, and can detect splice variants and non-coding RNAs that would otherwise be missed (Perkins et al., 2014; S. Zhao et al., 2014). However, that does not mean that RNA-seq does not have inherent limitations, which can lead to biases and overvaluation of the results (Conesa et al., 2016; Hayer et al., 2015; Lahens et al., 2014; Swindell et al., 2014).

One of the first NGS-based transcriptomic studies in prion diseases was carried out by Basu and colleagues and used tag profiling Solexa sequencing to compare gene expression in bovine medulla tissue between BSE-infected cattle and healthy controls (U. Basu et al., 2011). Even though the throughput of this technology was very low compared to today's standards (5-6 million reads generated per sample), the study identified 190 DEGs. Of these, 73 were found to be upregulated and 117 downregulated, while 16 were involved in 38 KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathways. While this number of genes might seem low nowadays, it should be pointed out that databases were also much sparser, including a lower number of genes and less thorough annotations. Another interesting study on the normal function of PrP was published in the same year by Khalifé and colleagues (Khalifé et al., 2011). To assess the involvement of PrP in embryogenesis, the group performed a comparative transcriptomic analysis between FVB/N *Prnp* KO mice and FVB/N mice during the early embryonic stages. They identified 73 DEGs at development stage E6.5 and 263 DEGs at E7.5, while proteolysis, protease inhibition, biological adhesion, nervous system development, apoptosis, cell proliferation, and inflammatory and innate immune response were the most represented functional groups. A few years later, Muñoz-Gutiérrez et al. investigated the contribution of cellular factors to prion infection (Muñoz-Gutiérrez et al., 2016). They used two closely related ovine microglia clones with no detectable differences in $PrP^C$ expression levels, but with different prion susceptibility. After inoculation with scrapie positive and negative sheep brainstem homogenates and passaging, the cells were sequenced using Illumina technology. 22 DEGs were identified, most of which were found to be upregulated in poorly permissive microglia (selenoprotein P, endolysosomal proteases, and proteins involved in extracellular matrix remodelling). Some of the upregulated transcripts in permissive microglia included transforming growth factor beta-induced, retinoic acid receptor response 1, and phosphoserine aminotransferase 1. A Gene Set Enrichment Analysis (GSEA) identified proteolysis, translation, and mitosis as the most affected pathways.

More recently, Kanata and colleagues profiled the transcriptome and RNA editome of humanised transgenic mice (sCJD tg340-PRNP129MM) that recapitulate human disease pathology at preclinical and clinical disease stages (Kanata et al., 2019). In the early disease stage, 1,356 DEGs were identified while neuronal and synaptic pathways and signalling cascades associated with oxidative or ER stress were implicated. In contrast, 655 DEGs were identified during the clinical disease stage and dysregulated pathways included cell survival, proliferation, differentiation, lysosome function, and immune system. Interestingly, 58 genes were found to be dysregulated in the same direction at both time points. A genome-wide atlas of gene expression, splicing and editing alterations during the course of disease in prion-infected mice was generated by Sorce and colleagues, aiming to shed light on prion transcriptomics during the disease, including time points much earlier than the appearance of clinical symptoms (Sorce et al., 2020). The authors underline that prion infection induced changes in mRNA abundance and processing well ahead of any neuropathological signs (Figure 1.2). In addition, the gene expression patterns were different for microglia-enriched genes, which were found to be upregulated simultaneously with the appearance of clinical symptoms, and for neuronal-enriched transcripts, which remained at the same levels until the end-stage of the disease. Thus, they hypothesise that glial pathophysiology represents the final driver of disease.

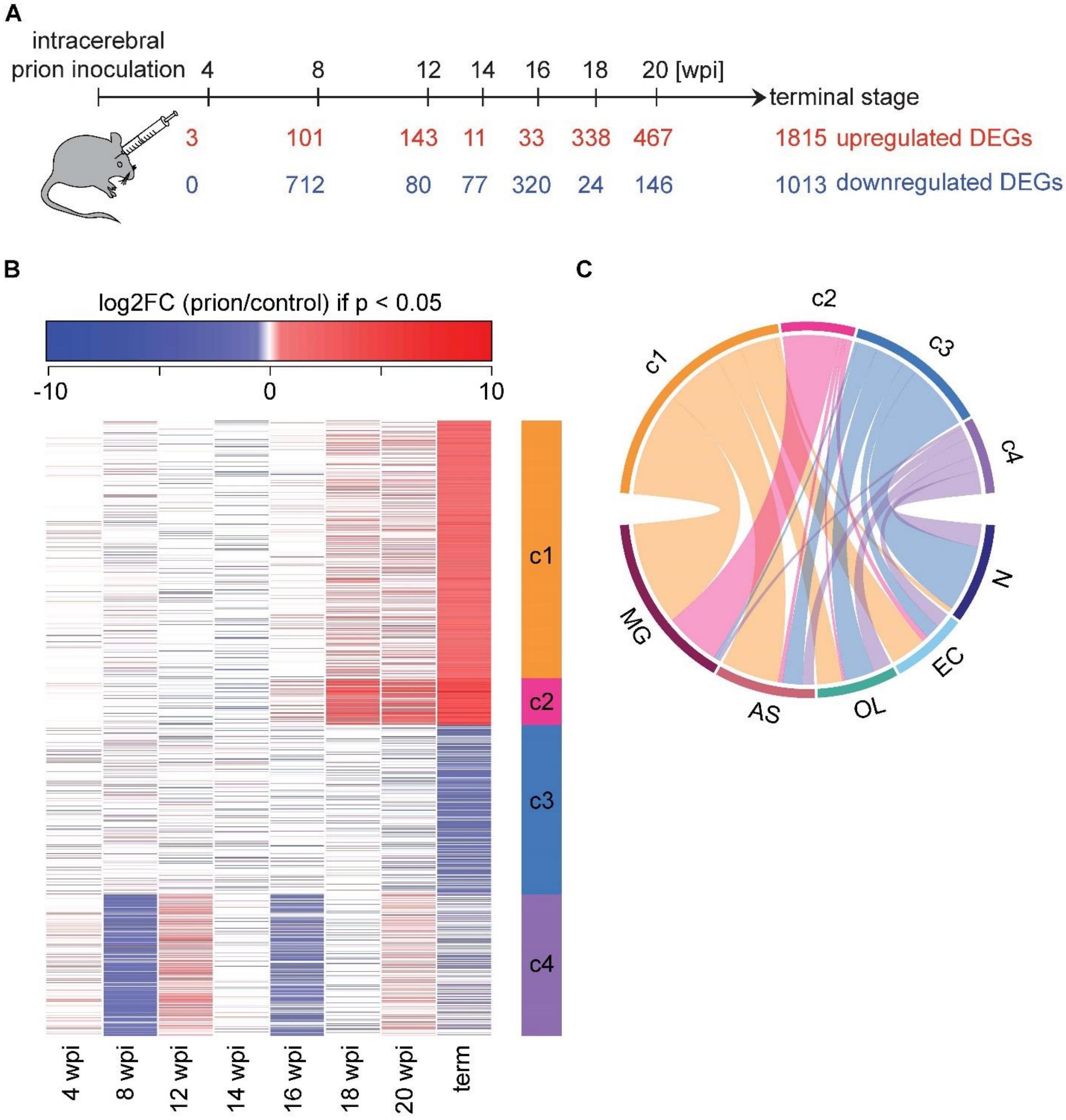


***Figure 1.2: Identification of DEGs during prion disease progression by Sorce et al.*** *The authors designed a time-course experiment, intracerebrally inoculating wild-type C57BL/6J mice with either RML or non-infectious brain homogenate.* ***(A)*** *Timeline of prion inoculations and numbers of upregulated and downregulated DEGs (|log2FC| > 0.5 and FDR < 0.05).* ***(B)*** *Heatmap displaying the log2FC of 3,723 genes that are differentially expressed at least at one time point. Only log2FC values with p < 0.05 are coloured. An unsupervised k-means clustering analysis (k = 4 clusters) identified four patterns (c1-c4) of log2FC oscillations over time (right sidebar).* ***(C)*** *Circos plot summarizing the cell type-enriched genes within each*

*cluster. MG: microglia, AS: astrocytes, OL: oligodendrocytes, EC: vascular cells, N: neurons. Figure adapted from Sorce et al. 2020.*

The advent of NGS also enabled high-throughput querying of the expression of miRNAs, as complementary transcriptional regulators. Gao et al. used deep sequencing to profile the expression of miRNAs in mice infected with different scrapie agents (139A, ME7 and S15) at the terminal disease stage (Gao et al., 2016). The comparison to age-matched normal controls revealed 57, 94 and 135 DE miRNAs in pooled brain samples of 139A-, ME7- and S15-infected mice, respectively. Interestingly, 22 and 14 of them were found to be commonly upregulated and downregulated, respectively, in all three models, while a KEGG pathway analysis highlighted the involvement of 12 similar pathways. A few years later, Norsworthy and colleagues published a blood miRNA signature study in sCJD patients (Norsworthy et al., 2020). In that study, they profiled miRNA expression from blood of 57 sCJD patients and 50 healthy controls and identified 5 DE miRNA transcripts (*hsa-let-7i-5p*, *hsa-miR-16-5p*, *hsa-miR-93-5p*, *hsa-miR-106b-3p* and *hsa-let-7d-3p*). This signature was found to discriminate sCJD from Alzheimer's disease patients, while the rate of decline in miRNA expression significantly correlated with disease progression. The authors argue that this novel signature can provide information to facilitate disease diagnosis and monitoring in a non-invasive manner.

NGS made in-depth transcriptomic queries possible and has allowed scientists to investigate the role of PrP, identify factors contributing to prion susceptibility and characterise novel miRNA signatures and mRNA alterations, ultimately leading to the accumulation of knowledge concerning prion diseases. Studies using cell lines identified retinoic acid receptor response 1, transforming growth factor β-induced and phosphoserine aminotransferase 1 as a permissive prion disease signature, while humanised mouse models facilitated the identification of oxidative and ER stress as having a central role in the late stages of the disease. Glial pathophysiology has also been implicated, while novel blood miRNA signatures might contribute to disease diagnosis. However, while these studies generated unprecedented amounts of data, they have failed to substantially elucidate the underlying disease mechanisms. Importantly, even though many different experimental models have been used - including cattle, cell lines and transgenic mice - no NGS information is available regarding the human prion-infected brain. Further leaps in our understanding of the underlying complexity of other neurodegenerative diseases would be made possible due to the introduction of single-cell transcriptomics.

## 1.3 Single-cell transcriptomics

### 1.3.1 The need for single-cell resolution

As next-generation sequencing technologies matured, they justifiably became the method of choice for the majority of transcriptomic studies. Years of development from private companies and academic institutions alike led to further optimisation of the chemistry involved, new materials were utilised to produce denser substrates, innovations in nanotechnology led to the introduction of nanopore-based sequencing and the development of new open-source software and algorithms democratised data analysis and interpretation. All these discoveries have culminated in an unprecedented increase in data throughput, high levels of sensitivity and specificity, the introduction of various methods that specifically target different aspects of biology (like splicing, RNA editing, RNA methylation or other modifications etc), and most importantly the generation of vast amounts of useful data that has been placed in the epicentre regarding the elucidation of complex diseases, such as cancer and neurodegeneration.

However, sequencing approaches that use bulk tissue as input material assume and represent all the cells as a homogeneous mixture, while in reality, ex vivo material can consist of different cell types which can potentially have dissimilar gene expression patterns (Raj & van Oudenaarden, 2008). One of the major arguments against bulk tissue sequencing is that it averages the gene expression of all input cells and, thus, all intra-cellular heterogeneity and population-specific genetic signatures are lost, possibly hindering the extraction of meaningful conclusions.

This notion has been strengthened by the advent of new single-cell RNA sequencing (scRNA-seq) technologies that enabled the dissection of gene expression and the preservation of valuable information about the cells of origin. Indeed, newer single-cell sequencing studies have revealed biologically meaningful and previously underestimated intracellular gene expression variability (G. Chen et al., 2016; Jaitin et al., 2014; Rosenberg et al., 2018), as well as various previously unidentified cell types (Cao et al., 2017; Grün et al., 2015; Rosenberg et al., 2018). Some representative examples worthwhile mentioning include two independent studies by Mathys et al. and Rosenberg et al. Mathys and colleagues sequenced 80,660 single nuclei from 48 individuals with varying levels of Alzheimer's disease pathology (Mathys et al., 2019). The data revealed that groups of genes, called marker genes, are differentially expressed in specific cell populations (Figure 1.3). The authors argue that this selective response to disease can provide valuable information and shed light on disease emergence and manifestation, highlighting the fact that bulk RNA-seq analysis of the same samples could not uncover this heterogeneity. This study is of great interest and relevance to this project and will be discussed in more detail in the following sections. In another pivotal study, Rosenberg and colleagues used a scRNA-seq technique called SPLiT-seq to analyse 156,049 single-nucleus transcriptomes from postnatal day 2 and 11 mouse brains and spinal cords (Rosenberg et al., 2018). They identified 73 distinct clusters when grouping the transcriptomes using unsupervised clustering; most of those were neuronal cells (54 clusters), while the rest were assigned to four astrocyte types, six oligodendrocyte types, one oligodendrocyte precursor cell type, two vascular and leptomeningeal cell types, endothelial cells, smooth muscle cells, microglia, macrophages, ependymal cells, and olfactory ensheathing cells. While this study identified molecular markers for specific cell types and identified subtypes for the first time, functional differences of the majority of cellular subtypes remain unclear.

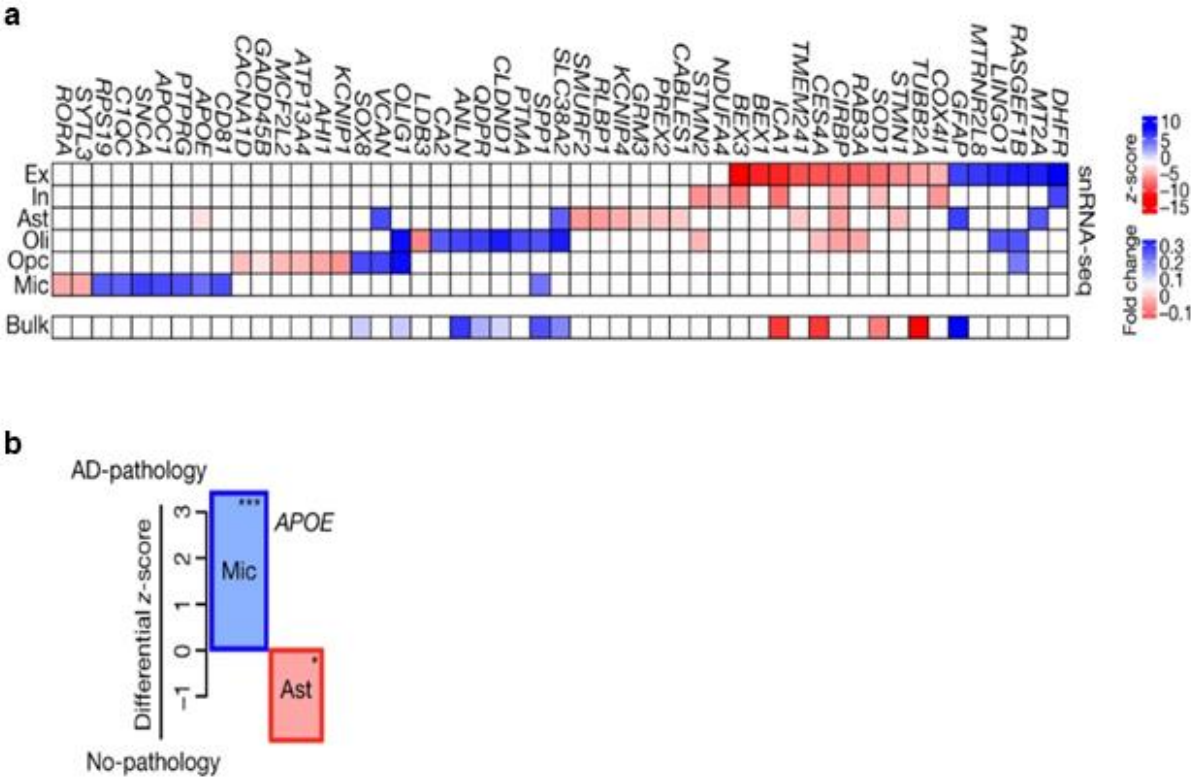


***Figure 1.3: Single-cell sequencing of the prefrontal cortex of Alzheimer's disease patients uncovers heterogeneity in gene expression patterns of different cell populations that cannot be detected by bulk RNA sequencing. (a)*** *Comparison of differential expression signature of marker genes in 6 different cell types versus global differential expression patterns identified by bulk sequencing. Single-cell analysis reveals that these marker genes are selectively dysregulated in specific cell populations, while they show little deviation in others. A bulk RNA-seq analysis of the same samples can identify only the strongest signatures (the downregulated signature of oligodendrocytes and the upregulated signature of excitatory neurons), while population-specific information is lost. Ex: excitatory neurons, In: inhibitory neurons, Ast: astrocytes, Oli: oligodendrocytes, Opc: oligodendrocyte precursor cells, Mic: Microglia.* ***(b)*** *scRNA-seq identifies opposite differential expression direction of gene APOE in microglia (Mic) and astrocyte (Ast) cell populations. APOE was found to be upregulated in microglia while being downregulated in astrocytes. Figure adapted from Mathys et al., 2017.*

### 1.3.2 Single-cell sequencing technologies

The transcriptome is a major determinant of cell function; distinct gene expression programs can regulate both cellular activity and identity (Y. Li et al., 2017; C. Sun et al., 2018; Zhou et al., 2020). The appeal of single-cell transcriptomics stems from the fact that they can be used to characterise and classify cells at the molecular level, offering an unbiased approach without being constrained to specific features, such as proteomics. Even though the field is still in its infancy, the ability to perform high throughput sequencing of the transcriptome of a single cell has existed for more than a decade. The technique was first introduced by Tang and colleagues in 2009 (F. Tang et al., 2009). By leveraging the ability to perform untargeted single-cell mRNA amplification, the researchers developed and adapted technologies to incorporate high throughput DNA sequencing into the equation and demonstrated the first transcriptome-wide querying of the mRNA from a single cell. In their study, they detected the expression of 75% more genes than microarray-based techniques had at the time and identified 1,753 previously unknown splice junctions in a single mouse blastomere.

While the first studies were focused on sequencing a few interesting cells with known identities (Ramsköld et al., 2012; F. Tang et al., 2010), it soon became clear that larger-scale studies would require parallel profiling of multiple cells which might not be pre-sorted. Guo and colleagues' pivotal study demonstrated that cell types of mixed cell populations can be identified without pre-sorting, based on their transcriptional patterns (Guo et al., 2010). This realisation paved the way for the invention of novel protocols and technologies that allowed the exponential scaling of single-cell experiments (Figure 1.4).

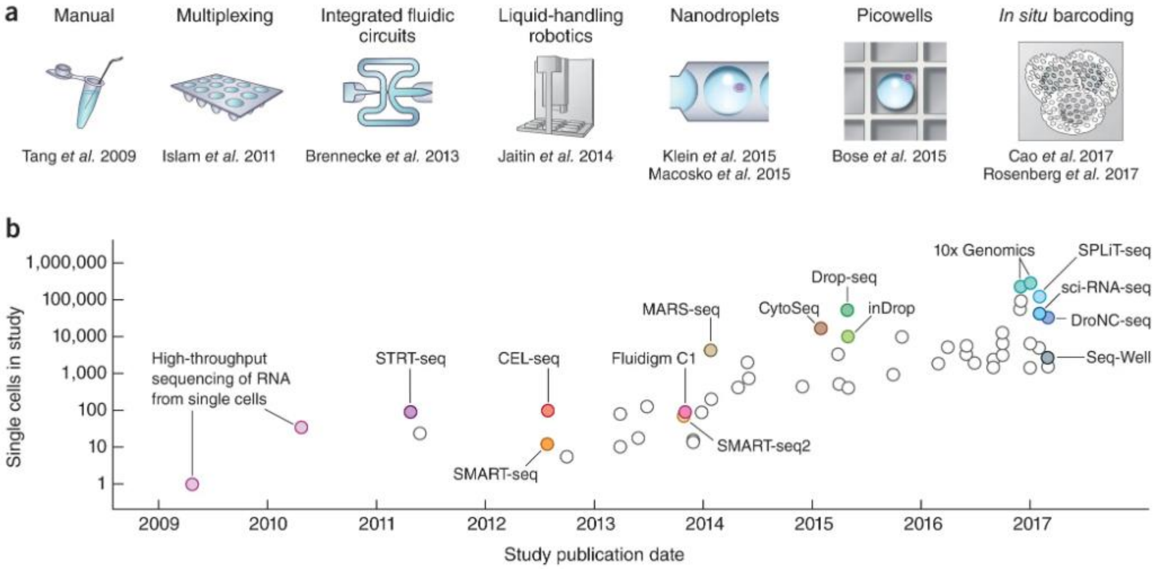


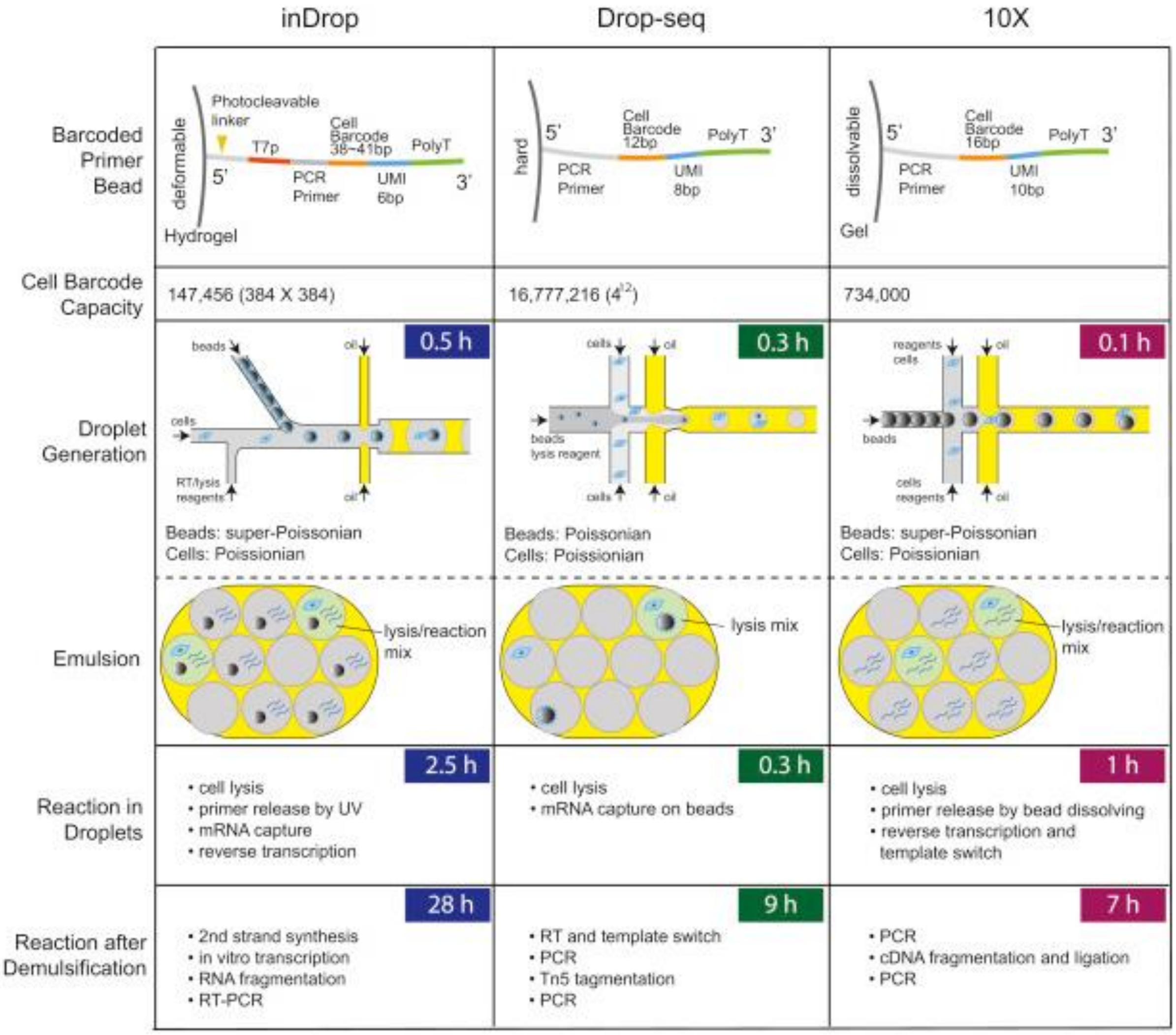
***Figure 1.4: Scaling of scRNA-seq experiments.*** *a) Key technologies that have led to an increase in single-cell experiment throughput. Sample multiplexing was the first approach that allowed the sequencing*

*of hundreds of cells. The design of integrated fluidic circuits and robotics for automation meant that thousands of cells could be profiled. Nanodroplets and picowells represent random capture technologies that increased the throughput to tens of thousands of cells. Finally, in situ barcoding has allowed the profiling of hundreds of thousands of cells, while keeping the cost to a minimum. b) The number of cells assayed has followed an exponential increase during the last decade. The graph shows the number of cells reported in manuscripts using the different technologies annotated. Figure adapted from Svensson et al., 2018.*

While various scRNA-seq technologies have been developed lately, having different strengths and weaknesses (G. Chen et al., 2019; Natarajan et al., 2019; Xiannian Zhang et al., 2019; Ziegenhain et al., 2017) and have been applied to a wide array of samples (blood cells, solid tissue, frozen or fixed tissue etc.), this thesis will focus mainly on two technical approaches that were selected for the purposes of this research project, namely droplet-based technologies, and combinatorial indexing techniques.

Droplet-based technologies utilise microfluidic devices to encapsulate and compartmentalise single cells in nanolitre droplets that include cell-specific oligonucleotide primers. Simple microchannels introduce immiscible reagents at specific rates in chambers of carefully designed geometry allowing the generation of thousands of droplets per second. While most of these will be empty, a small percentage will contain a single cell and, due to the high droplet generation rates, thousands of cells can be encapsulated in just a few minutes. In parallel to the cells, the systems allow co-encapsulation of specifically designed beads that introduce the necessary primers and barcodes. Each droplet acts as a reaction chamber, containing the necessary reagents for the first steps of cDNA library preparation. This strategy greatly increases the reaction throughput justifying the classification of the methods as “ultra-high-throughput”. In addition, the nanolitre scale of the reaction volumes keeps reagent usage low, reducing the experimental cost and allowing the profiling of more cells. Currently, the most widely adopted droplet-based systems include inDrop (Klein et al., 2015), Drop-seq (Macosko et al., 2015) and 10X Genomics Chromium (Zheng et al., 2017). All of these have been demonstrated to be robust and efficient in generating single-cell libraries from thousands of cells at an acceptable cost; they use similar microfluidics technologies to generate droplets, make use of unique molecular identifiers (UMIs) for PCR bias correction, use barcoded beads to differentiate between individual cells and involve a final NGS step.

inDrop and Drop-seq are open-source protocols that were published in the same issue of Cell in 2015 (Klein et al., 2015; Macosko et al., 2015). In contrast, 10X is a proprietary technology developed a few months later based on the same principles. Despite their similarities, the protocols use different approaches regarding bead manufacturing, single-cell barcode design and cDNA library generation (Figure 1.5). All systems require the manufacturing of specific beads covered with oligonucleotide primers. These nucleotide sequences include a PCR handle, a cell barcode, a UMI, and a poly-T region. inDrop beads also include a photocleavable region and a T7 promoter. Drop-seq beads are made of hard resin, while 10X and inDrop use hydrogel beads. The distinct advantage of using deformable hydrogel is that it allows super-Poissonian loading of the beads in the droplets, leading to higher percentages of droplet occupation (Abate et al., 2009). Loading of droplets with cells for all protocols, and both beads and cells for Drop-seq, follows a Poissonian distribution. Encapsulation takes place in specially designed microfluidics channels of similar geometry. The reaction buffer incorporated inside each droplet includes lysis reagents that cause rupture of the captured cell and release of its nucleic acids. Poly-A mRNA can then bind to the poly-T region of the primers before reverse transcription. The reaction takes place inside the droplets for inDrop and 10X, while Drop-seq requires prior demulsification. After reverse transcription and the introduction of a demulsifying agent, a final library preparation step and amplification are required to make the products compatible with Illumina sequencing. Here, cDNA is fragmented, adapters are ligated, and the products are amplified and purified before sequencing. It is worth pointing out that inDrop uses in vitro transcription during this final library preparation step, increasing the protocol’s length by 28 hours. Finally, the resulting library can be sequenced on Illumina instruments, such as NextSeq and HiSeq. The generated data can then be used to demultiplex cell information and quantify transcript abundance.



| | inDrop | Drop-seq | 10X |
|---|---|---|---|
| Barcoded Primer Bead | | | |
| Cell Barcode Capacity | 147,456 (384 X 384) | 16,777,216 ($4^{12}$) | 734,000 |
| Droplet Generation | 0.5 h<br>Beads: super-Poissonian<br>Cells: Poissionian | 0.3 h<br>Beads: Poissonian<br>Cells: Poissionian | 0.1 h<br>Beads: super-Poissonian<br>Cells: Poissionian |
| Emulsion | | | |
| Reaction in Droplets | 2.5 h<br>• cell lysis<br>• primer release by UV<br>• mRNA capture<br>• reverse transcription | 0.3 h<br>• cell lysis<br>• mRNA capture on beads | 1 h<br>• cell lysis<br>• primer release by bead dissolving<br>• reverse transcription and template switch |
| Reaction after Demulsification | 28 h<br>• 2nd strand synthesis<br>• in vitro transcription<br>• RNA fragmentation<br>• RT-PCR | 9 h<br>• RT and template switch<br>• PCR<br>• Tn5 tagmentation<br>• PCR | 7 h<br>• PCR<br>• cDNA fragmentation and ligation<br>• PCR |

***Figure 1.5: Comparison of high-throughput droplet-based scRNA sequencing technologies.*** *All three technologies use barcoded primer beads to introduce cell-specific barcodes which allow demultiplexing of the generated data. The oligonucleotide constructs are specific to each technology but include PCR handles, Unique Molecular Identifiers, and poly-T tails for poly-A mRNA capture. Droplet generation and cell encapsulation use microfluidic devices of similar geometry. The reactions after droplet formation are specific to each protocol and involve cell lysis and mRNA capture, reverse transcription, amplification, and Illumina library generation. Drop-seq and 10X protocols can be concluded in under 10 hours, while inDrop requires substantially more time due to utilising in vitro transcription for nucleic acid amplification. Figure adapted from Xiannian Zhang et al., 2019.*

While droplet-based techniques use nanolitre droplets for compartmentalising different cells and introducing cell-specific oligonucleotide barcodes in each reaction chamber, combinatorial indexing techniques build upon this logic to take the protocol one step further, abolishing the need for droplet generation and utilising each cell's body as a reaction chamber. At the core of combinatorial indexing approaches are multiple split-pool barcoding steps. In summary, a large number of entities to be barcoded are split across different reactions, each of which will incorporate a specific barcode. Then the entities are pooled together and mixed, before being split again in a different batch of reactions, where new barcodes will be incorporated. These cycles of split-pooling ultimately lead to the tethering of multiple sequential barcodes in each entity. Even though each barcode cannot provide enough information for effective demultiplexing in isolation, the adapter combination should have a very low probability of being created more than once in well-designed experiments and can thus uniquely characterise the entity.

Early single-molecule combinatorial indexing approaches have been used for de novo genome assembly and haplotype-resolved genome sequencing (Adey et al., 2014; Amini et al., 2014). The first single-cell applications of this principle were used for profiling chromatin accessibility (Cusanovich et al., 2015), genome sequence (Vitak et al., 2017), chromosome conformation (Ramani et al., 2017), and DNA methylation (Mulqueen et al., 2018) in single cells. The first to implement this methodology to uniquely label the transcriptomes of single cells were Cao and colleagues in 2017, when they applied their protocol, termed sci-RNA-seq, to profile the transcriptome of the whole multicellular organism Caenorhabditis elegans at the L2 stage (Cao et al., 2017). Less than a year later, Rosenberg and colleagues developed their own version of a split-pool barcoding protocol, termed SPLiT-seq (split-pool ligation-based transcriptome sequencing) and used it to transcriptionally profile more than a hundred thousand mouse brain and spinal cord cells (Rosenberg et al., 2018). Finally, Cao and colleagues improved their protocol and introduced sci-RNA-seq3 in 2019, making it the highest-throughput single-cell transcriptomics approach to date, capable of profiling more than 2 million cells per experiment (Cao et al., 2019).

A proprietary method based on combinatorial indexing was first made available to the research community in 2021. Following their publication in Science in 2018, Rosenberg and Roco launched Parse Biosciences, which was successfully funded during the following years. Their whole transcriptome kit was launched in February 2021 and is based on the principles of SPLiT-seq. While no peer-reviewed information is available, the company claims that extensive optimisation has led to an improved method that offers higher sensitivity and throughput, lower doublet rates, and an easier protocol with reduced hands-on time. Depending on the number of cells to be sequenced, three products are available that can be used for up to 1 million cells. The Evercode™ Whole Transcriptome is the medium-sized kit that can be used to profile up to 100,000 cells and 48 samples and is the kit that has been used as part of this study.

All split-pool barcoding techniques follow the same principles, but each protocol has its own specificities (Figure 1.6). The protocols start with cell fixation and permeabilization, after isolating and dissociating the number of cells that will be used. Careful experimental planning allows the calculation of the maximum number of cells based on the accepted collision probability, i.e. the maximum accepted probability that two or more cells will have the same barcode. It is important that the nucleic acids are fixed inside the cell body to prevent diffusion to a neighbouring cell during the split-pool rounds. Then follow several split-pool barcoding rounds, the number of which, and the number of compartments used for splitting, depends on the protocol used. A higher number of barcoding rounds and compartments used for splitting will increase the number of barcode combinations possible, which means the experiment's throughput is crucially dependent on this step. The cells or nuclei can then be lysed, and their barcoded nucleic acids purified. Subsequent tagmentation reactions introduce the final barcodes and the library is PCR-amplified before sequencing. The data resulting from sequencing these complex constructs are demultiplexed and analysed to generate transcript abundance information.

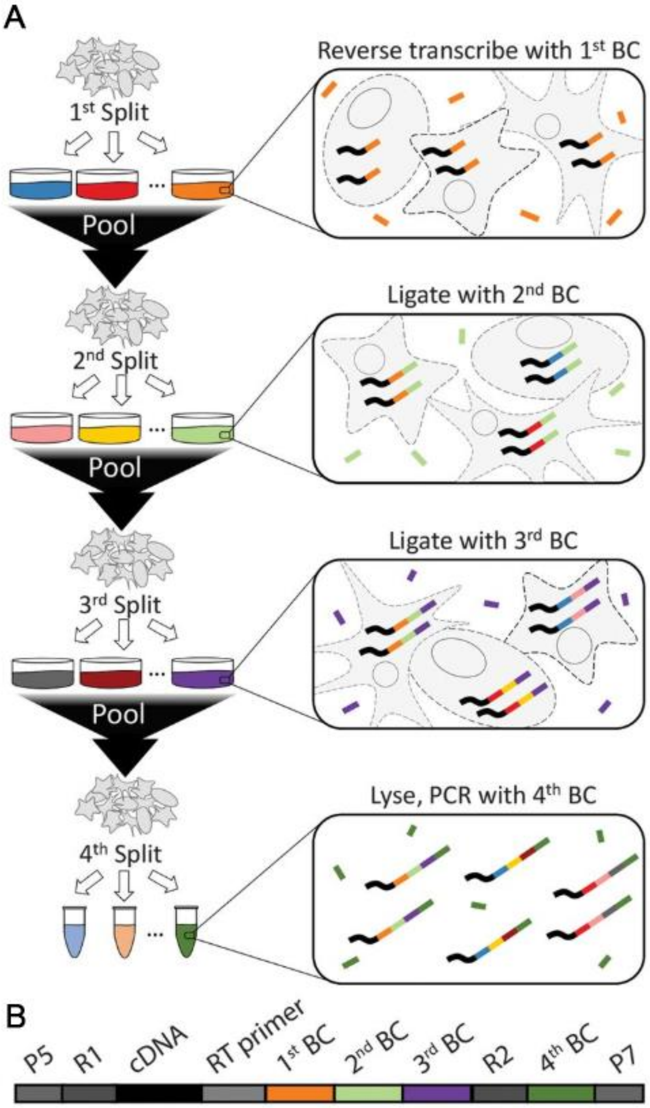


***Figure 1.6: Split-pool barcoding schematic used in SPLiT-seq.*** *A) Cells are randomly split into different reaction chambers which contain specific barcodes. After the barcodes are incorporated into the cellular RNA via Reverse Transcription (RT), the cells are pooled together and mixed. They are then randomly split again into a new batch of barcodes, which will be ligated next to the previous ones. At the fourth split, cells are lysed, and the final round of barcoding takes place via PCR. B) A schematic representation of the construct generated after library preparation (bar length not to scale). The construct contains P5 and P7 sequencing adapters, R1 and R2 PCR handles, the cDNA captured, and a sequence of four cell-specific barcodes (BCs). Figure adapted from Rosenberg et al., 2018.*

### 1.3.3 Single-nucleus sequencing technologies

All current scRNA sequencing approaches require that entities to be processed have an intact and geometrically uniform external membrane, i.e. the cell membrane. Ruptures in the cell membrane can cause the nucleic acids of the cell to diffuse out and possibly mix with those of other cells, essentially making demultiplexing information of origin impossible. In addition, all microfluidics and droplet-based approaches, as well as most of the other protocols involve steps that make assumptions about cellular geometry. Processes such as droplet encapsulation, cell sorting or filtering, as well as all microfluidic devices have been designed and optimised on the theoretical grounds of cellular uniformity and sphericity. This might explain why a considerable number of single-cell studies focus on profiling blood cells (Khan & Kaihara, 2019; Schafflick et al., 2020; Szabo et al., 2019; Y. Zhao et al., 2019); these cells have a very spherical geometry and can be easily isolated.

Taking the above into consideration, it becomes evident that solid tissues are much harder to sequence at a single cell level due to their tight, coherent structure. Cells are embedded in the extracellular matrix and need to be carefully dissociated to create a cell suspension while their cell membranes remain intact. The optimisation of the cell harvesting protocols and biases that can be introduced have been the main discussion point of many studies (Bonnycastle et al., 2020; Denisenko et al., 2019; O’Flanagan et al., 2019; S. Zhu et al., 2017). Frozen tissue poses a greater challenge since frozen cell membranes tend to fracture (Branton, 2016), making the isolation of viable single cells after thawing inefficient. Finally, one of the most informative tissues in neuroscience studies is also one of the most challenging ones to be processed: frozen brain tissue is of utmost importance to neuroscience research, however, due to the irregular and variable shape of the cell populations it consists of, coupled with the fact that it is usually frozen, it cannot be efficiently used as input material to any of the single-cell techniques developed to date. The variable post-mortem delay before tissue storage and lack of time-course samples in human disease complicate matters even more.

To address these limitations several single-nucleus sequencing protocols had been developed, which enabled the study of potentially informative archived frozen brain tissue and other biological material difficult to dissociate (Grindberg et al., 2013; Habib et al., 2016; Lacar et al., 2016; Lake et al., 2016). While these protocols could successfully provide single-nucleus transcriptomic information, they did rely on technologies that cannot be scaled up efficiently, such as nuclei sorting in 96- or 384-well plates, limiting their applicability for large numbers of cells or samples. A major breakthrough that could address these shortcomings was made in 2017 when Habib and colleagues combined existing knowledge from single-nucleus protocols with the pioneer Drop-seq technology (Habib et al., 2017). Their protocol, DroNc-seq, is a massively parallel single-nucleus RNA-seq method that combines the advantages of both worlds to provide high-throughput nuclei profiling at a low cost. The subsequent split-pool barcoding techniques including SPLiT-seq and sci-RNA-seq provide protocols that have been tested with both single cells and single nuclei, making them suitable for a larger range of studies.

Even though single-nucleus approaches are the only option for archived and frozen samples, their drawbacks should be considered before selecting them for a specific study. First, all single-cell sequencing approaches are able to capture only a small fraction of the available RNAs. This is especially true for single-nucleus approaches where the amount of RNA available is even lower, which means that the methods' sensitivity is reduced. Deep sequencing of similar cells and nuclei with Drop-seq and DroNc-seq detected on average 5,134 and 3,295 transcripts, respectively, indicating the lower sensitivity of single-nucleus sequencing (Habib et al., 2017). In addition, qualitative differences can be expected. Single nucleus approaches do not sequence mitochondrial transcripts from genes that do not reside in the cell's genome. Also, the nucleus contains a lower percentage of mature mRNAs than the cytoplasm, so a higher proportion of the data mapping to introns is expected. Indeed, during the same comparison of the two methods by Habib and colleagues, while the same percentage of cellular and nuclear reads map to the genome, only 9.1% of cellular reads map to introns while this percentage is 41.8% for nuclei. Nevertheless, it has been reported that the average expression profile of single nuclei correlates well with that of single cells for both Drop-seq/DroNc-seq and SPLiT-seq (Habib et al., 2017; Rosenberg et al., 2018).

### 1.3.4 Bioinformatics analysis of scRNA-seq data

RNA sequencing data has existed for years before the introduction of single-cell sequencing techniques. Given the widespread use of bulk RNA-seq, many analytical pipelines have been tested and gold standards started to emerge in the scientific community (Oshlack et al., 2010). While single-cell RNA sequencing still uses the same sequencing technologies as previous approaches, the data generated is much more intricate due to the complex structure of the sequenced constructs. Except for RNA sequence data, most constructs include additional nucleotide sequences to encode and preserve information about the cell and/or transcript of origin and can also incorporate protocol-specific barcodes. While existing computational workflows can be adapted for single-cell data, its unique computational challenges necessitate the development of novel analytical strategies that can fully exploit and interpret the additional information available (Stegle et al., 2015).

As more single-cell sequencing techniques became available, offering increasingly higher throughput, the necessity for efficient data analysis became stronger. The great potential of these approaches, coupled with increasing amounts of generated data has motivated computational biologists to develop novel analytical tools (Rostom et al., 2017). The immaturity of the field and the specific requirements of the constantly evolving protocols lead to an explosion in the number of available tools (1295 as of July 2022) (Zappia et al., 2018). This wide variety of experimental protocols and analytical methods makes the standardisation of computational workflows a challenging undertaking (Luecken & Theis, 2019). The following paragraphs give an overview of the general analytical steps that are commonly part of single-cell data analysis, without emphasising specific topics relevant to each methodology used.

Raw sequencing data needs to be converted to a count matrix before it can be further processed. This pre-processing step is very specific to the exact technology used. In summary, gene expression information, as well as information regarding the cell and/or molecule of origin, is extracted from the raw sequencing reads and combined in an x times y dimensional matrix, where x corresponds to the number of rows or features (transcripts) and y to the number of columns or barcodes of origin (ideally each barcode of origin should correspond to one cell, however, in practice, barcode collisions are expected). This matrix contains counts of molecules (if UMIs are used) or reads (if UMIs are not available) and forms the basis for subsequent analyses. Protocol-specific raw data processing pipelines such as split-seq-pipeline (SPLiT-seq) (Rosenberg et al., 2018), Drop-seq tools (Drop-seq and DroNc-seq) (Macosko et al., 2015) or Cell Ranger (10X) (Zheng et al., 2017) can be used to automate this step, offering basic Quality Control (QC) options, demultiplexing and genomic alignment.

For a successful analysis, it must be made sure that the data used originated from healthy and viable cells. After pre-processing, the data undergoes a thorough QC analysis to identify and discard potential outlier cells of poor quality that can skew the results. Some of the common QC metrics include the number of counts per cell, the number of genes per cell and the fraction of mitochondrial genes (Griffiths et al., 2018; Ilicic et al., 2016). Specific thresholds need to be defined and outliers that surpass them are discarded. Cells with a small number of counts and genes and a high percentage of mitochondrial genes might correspond to low-quality or dying cells where the cellular membrane is broken and RNA has diffused out, while the larger organelles, like mitochondria, remain. In contrast, a high number of counts or genes might signify a doublet, where more than one cell incorporated the same barcode. Even though there is a rational process behind the selection of these thresholds, the exact cut-off values will need to be empirically set and are determined by the experimental methods and underlying biology. For example, in an experiment with proliferating cells, an increase in transcriptomic reads would be expected for the actively dividing cells. In addition, an increase in mitochondrial genes might signify that the cell is involved in respiratory processes. Concluding, these covariates should not be considered in isolation and the thresholds set should be re-evaluated during the later stages of the analysis depending on the preliminary results and combining the knowledge of the underlying biology and current hypotheses tested.

Single-cell RNA-seq data is characterised by high levels of stochasticity. Inherent variability in each of the experimental steps will be captured in the expression levels but does not necessarily stem from biological differences. In contrast, some of the observed variability might arise solely due to technical noise and sampling effects (J. K. Kim et al., 2015; Kolodziejczyk et al., 2015; Stegle et al., 2015). Data normalisation is used to address this issue by appropriately scaling the data to obtain relative gene expression abundances between cells. Some of the most frequently used methods use CPM (counts per million) normalisation which originated from bulk expression analysis. Due to the heterogeneous nature of single-cell datasets, more complex normalisation approaches are usually more appropriate. Scran's pooling-based size factor estimation method uses linear regression over genes to estimate size factors (Lun, Bach, et al., 2016) and has been shown to perform better than other algorithms for batch correction and differential expression analysis (Büttner et al., 2019; Vallejos et al., 2017). Non-linear normalisation methods can account for more complex variation and have been shown to outperform global scaling methods in experiments with strong batch effects (Cole et al., 2019). To summarise, different normalisation methods perform better for different datasets and tools have been developed that can select the most appropriate one (Cole et al., 2019).

Normalised data is then usually log-transformed. This is important to reduce the skewness of the data, as most downstream tools assume normality, to change the distances between the expression values to represent log-fold changes, which is usually used to measure gene expression, and finally to mitigate the mean-variance relationship of the data (Brennecke et al., 2013).

The data will usually contain thousands of dimensions, even after thorough QC and filtering. Most of these are not informative about biology and contain unwanted noise. Thus, the next steps include multiple dimensionality reduction approaches to reduce the computational burden, remove unwanted noise and visualise the data. Feature selection is usually the first step, where the most informative genes are kept, while all the rest are discarded. The “informative” genes are usually the most variable ones across the dataset, also called Highly Variable Genes (HVGs) (Brennecke et al., 2013). After the selection of the HVGs, the dimensionality of the dataset can be further reduced using dedicated algorithms that embed the dataset in a low-dimensional space aiming to capture the underlying data structure. These techniques are often very effective, as single-cell data is inherently low-dimensional and the biological information can be described by much fewer dimensions than the number of genes (Heimberg et al., 2016). Data visualisation algorithms use two or three dimensions to visually represent the structure of the data and cannot be used for downstream analysis. These include Principal Component Analysis (PCA) (Hotelling, 1933), t-Stochastic Neighbour Embedding (t-SNE) (Van Der Maaten & Hinton, 2008), Uniform Manifold Approximation and Projection (UMAP) (Becht et al., 2018), force-directed graphs (Costa et al., 2018) and others. In contrast, data summarization algorithms use an arbitrary number of reduced dimensions to describe the data where the higher components are less important for describing the variability present. They can be used to reduce the data to its essential components and their output can be used in downstream workflows. The most commonly used methods include PCA and diffusion maps (Coifman et al., 2005).

Downstream analyses attempt to fit interpretable models to uncover biological insights, describe the biological systems and test hypotheses. These can be roughly divided into cell-level and gene-level analyses, while there is a substantial overlap of methods in both groups. Cell-level analyses focus on the characteristics of each cell, as it is described by the ensemble of its transcripts. The most commonly employed ones include clustering analysis and annotation (Kiselev et al., 2019) and trajectory inference/pseudotime analyses (Saelens et al., 2019). Grouping the cells in clusters is usually the first substantial result of the analysis pipeline. Being a classical unsupervised machine learning problem, many algorithms have been developed from different scientific fields. Systematic evaluation of different algorithms has shown that the Louvain algorithm (Blondel et al., 2008), the default method implemented in SCANPY and Seurat analysis packages, performs best for scRNA-seq data (Duò et al., 2018; Freytag et al., 2018). Common gene-level downstream analyses include differential expression, gene network (Ideker et al., 2002) and gene set enrichment analyses (Subramanian et al., 2005). Differential expression testing originates from bulk RNA-seq (Scholtens & von Heydebreck, 2005) and when used on single-cell data can account for cellular heterogeneity and perform comparisons within cell clusters of the same identity. This additional information increases the resolution of the analysis which can identify cell-identity-specific transcriptional perturbations.

Given the complexity of single-cell data and the richness of the information that can be extracted, single-cell transcriptomics data requires an ensemble of analytical tools to be manipulated. These independent tools are frequently aggregated in analysis platforms to facilitate the flow of information and the construction of efficient workflows. While Graphical User Interface platforms exist (Gardeux et al., 2017; Patel, 2018; Rue-Albrecht et al., 2018), these usually offer limited flexibility while web-based platforms are limited in their ability to scale due to computational infrastructure. Command-line platforms are much more prominent and have been developed mainly for the R and Python programming languages. Among these, Seurat (A. Butler et al., 2018) and Scater (McCarthy et al., 2017) are the most popular and comprehensive platforms based on R, while SCANPY (Wolf et al., 2018) is based on Python.

## 1.4 Single-cell transcriptomics in neurodegenerative diseases

Single-cell and single-nucleus methods have been particularly useful in profiling transcriptional perturbations in various neurodegenerative diseases. The increased resolution offered in comparison to bulk sequencing approaches has been pivotal in studying the Central Nervous System (CNS), which is known to be composed of very heterogeneous cell populations (Habib et al., 2017; Lake et al., 2016; Rosenberg et al., 2018; Zhong et al., 2018).

Most of the earlier studies focused mainly on immune cell populations, and more specifically microglia, as they are involved in the maintenance and elimination of synapses and can act as damage sensors in the CNS (Aguzzi et al., 2013). This choice is justified by a large body of evidence from previous studies, where immunological mechanisms have been implicated in the pathogenesis of neurodegenerative diseases (Gjoneska et al., 2015; Mosher & Wyss-Coray, 2014; Y. Wang et al., 2015; B. Zhang et al., 2013). The most studied diseases are Alzheimer's disease (AD) and multiple sclerosis (MS), while no single-cell study has been published on human or animal prion diseases to date.

In 2017, Mathys and colleagues used single-cell RNA to track microglia activation in a time-course experiment using the CK-p25 mouse model of severe neurodegeneration (Mathys et al., 2017). This model, while it does not contain any AD-associated mutations, has been shown to recapitulate many aspects of disease pathology and allows for precise triggering of neurodegeneration. The researchers profiled a total of 1,685 pre-sorted hippocampal cells expressing microglia markers from four time points (before neurodegeneration triggering, during early and late disease). Their main finding was that after neurodegeneration triggering, microglia populations have a quite different transcriptional profile and cluster separately from most cells isolated from the healthy brain. These activated microglia were organised in two different subclusters associated with disease progression, indicating the existence of an early- and a late-response phenotype. Later that year, Keren-Shaul and colleagues published a similar study where they identified a novel microglia type associated with neurodegenerative diseases, termed DAM (Disease-Associated Microglia) (Keren-Shaul et al., 2017). The researchers sorted and sequenced immune cells from brains of 5XFAD, an AD transgenic mouse model that expresses 5 human familial AD mutations. After cell clustering, they identified two microglia clusters that represent distinctive states observed in AD but not in the controls; these also expressed lower levels of several microglia homeostatic genes. They described that DAM activation follows two sequential stages, where the second includes induction of lipid metabolism and phagocytic pathways and is *Trem2*-dependent. Based on previous data, the authors hypothesise that this late phenotype could mitigate disease. Finally, they identify a similar DAM subpopulation in an ALS mouse model, generalising their study and suggesting that this newly identified microglia population might not be associated with a specific disease, but rather with general mechanisms involved in aggregated protein clearance.

The year 2019 saw many published studies exploring the immunological component of neurodegenerative and neuroinflammatory diseases. Masuda and colleagues combined single-cell transcriptomics, single-molecule fluorescence in situ hybridization (FISH) and immunohistochemistry to characterise microglial subclasses during development and disease (Masuda et al., 2019). Their mouse experiments indicate the presence of time- and location-dependent subtypes of microglia during homeostasis. This signature was enriched in transcripts such as *P2RY12*, *CX3CR1*, *TMEM119* and *SLC2A5*, which are known homeostatic genes. In contrast, neurodegenerative disease mouse models (cuprizone treatment and unilateral facial nerve axotomy) showed a division of microglia in distinct subtypes with different molecular hallmarks. They then extended their study by including 1,180 cortical microglia from human brain tissue without evidence of CNS pathology and 422 CD45+ cells from brain tissue of five multiple sclerosis patients. After clustering and removal of clusters having monocytic and lymphocytic profiles, the remaining seven microglial clusters were compared. Three clusters consisted entirely of healthy microglia and showed the highest levels of expression of homeostatic genes. One of the clusters consisted of microglia from both patients and controls and showed upregulation of chemokine and cytokine genes, which suggests that these microglia were pre-activated. One of the clusters was shown to be characterised by increased expression of *CTSD*, *APOC1*, *GPNMB*, *ANXA2* and *LGALS1*, and another showed increased expression of MHC II genes, suggesting an immunoregulatory role. Finally, the last cluster showed increased expression of *SPP1*, *PADI2* and *LPL*, which correlated with the signature of demyelination-associated microglia in mice.

Another broader study of MS was performed by Schirmer and colleagues (Schirmer et al., 2019). The researchers profiled all cell populations from 12 MS human brain tissue samples and 9 healthy controls using 10X snRNA-seq. Their experiment yielded 48,919 single-nucleus profiles, which were organised in 22 clusters. Interestingly, they observed a selective reduction of upper-layer excitatory neurons (ENs) in MS, while all other cell populations, including intermediate-layer and deep-layer ENs, remained unchanged. A trajectory analysis of L2 and L3 ENs identified upregulated Gene Ontology (GO) terms relative to oxidative stress, mitochondrial dysfunction, and cell death. Some long-noncoding RNAs were also found to be upregulated (*NORAD* and *BCYRN1*). Their findings suggest selective transcriptomic damage of upper-layer ENs in MS. Then, using spatial transcriptomics, they mapped glial gene expression in the cortical and subcortical lesion and non-lesion areas. Transcriptional perturbations suggesting an activated phenotype were identified in microglia, oligodendrocytes and astrocytes located in the rim areas of chronically active subcortical lesions. Upregulated genes were associated with cell stress, heat-shock response, iron accumulation MHC class I upregulation and protein degradation. Furthermore, distinct transcripts were identified for cortical astrocytes and subcortical lesion astrocytes, providing another example of spatial diversity in neurodegenerative diseases. Interestingly, their single-nucleus approach also identified phagocytosing cells based on the identification of transported myelin transcripts into their nucleus or perinuclear structures.

Using the same single-nucleus droplet-based technology, Mathys and colleagues profiled 80,660 nuclei from the prefrontal cortex of 48 individuals with varying degrees of

Alzheimer’s disease pathology, publishing their findings in the same year (Mathys et al., 2019). Their cell data was organised in 20 clusters, which were annotated to be excitatory and inhibitory neurons, astrocytes, oligodendrocytes, microglia, oligodendrocyte progenitor cells, endothelial cells and pericytes. Their DE analysis identified a strong signature of repression in excitatory and inhibitory neurons, while most transcripts in oligodendrocytes, astrocytes and microglia were upregulated. Their findings highlight the heterogeneity of cellular response to disease, a recurrent pattern common for all cited studies in this section. They then compared the single-nucleus data to bulk RNA-seq data of the same samples to underline the fact that information regarding cell-type-specific changes is not captured with bulk approaches, especially for DE genes with opposite directionality in different cell populations. A comparison of stratified data between early and late pathology indicated that transcriptomic perturbations occur before the appearance of severe pathological features. While early-pathology transcripts were shown to be cell-type specific, late-pathology ones were found to be commonly upregulated across cell types. These genes were associated with autophagy, apoptosis, and stress response, indicating a general perturbation of the proteostasis network. Subclustering of the cell populations showed that specific subpopulations were associated with disease pathology, indicating differential responses to disease among the same cell type, similar to the observations of both Masuda et al. and Schrimer et al. Finally, the observation of robust gender differences at the molecular level in AD patients lead the authors to hypothesize that transcriptional response to AD pathology might be sex-specific.

Another study on AD, which was published in the same year, identified the same recurrent pattern of heterogeneous cellular response to disease. Grubman and colleagues profiled a total of 13,214 single nuclei from the entorhinal cortex of 6 control and 6 AD brains using 10X technology (Grubman et al., 2019). The transcriptomes were clustered and annotated in six groups: microglia, astrocytes, neurons, oligodendrocyte progenitor cells (OPCs), oligodendrocytes, and endothelial cells, accounting for “hybrid” cells that expressed multiple markers and might represent intermediate cell states. The researchers then underline that their transcriptomic perturbations show high concordance of effect (>90%) with the previous study by Mathys et al. indicating replicability of single-nucleus RNA sequencing experiments. Further analysis showed that astrocytes, endothelial cells, and microglia exhibit coordinated gene expression differences, i.e., dysregulation is observed in clusters of genes, specific for each cell type; some gene clusters were found to be coordinated in multiple cell types, including genes associated with cell stress and topologically incorrect protein response. Interestingly, *APOE*, an important AD risk gene, was found to be repressed in astrocytes and oligodendrocyte progenitor cells and upregulated in specific microglial subpopulations, mirroring the results of the previous study. A subcluster-specific analysis revealed that AD and control cells tend to segregate in different subclusters, except in some clusters of neurons, suggesting disease-associated transcriptomic perturbations across almost all cell types. One of the novelties of the study was the integration of single-cell data with prior information from genome-wide association studies (GWAS). The researchers examined the cell-type specificity of expression of one thousand GWAS genes for AD and AD-related traits and identified microglial expression specificity for two of them (*RIN3* and *TBXAS1*) with functional roles in endocytosis and vasoconstriction. Furthermore, they predicted transcription factors that are driving the dynamic cell state transition towards AD. These included *AEBP1*, *SOX10*, *MYRF* and *NKX6-2*. Finally, they highlight transcription factor *TFEB*, which is shown to regulate ten GWAS targets in astrocytes; both the target and the factor were shown to be dysregulated in the same populations, establishing a functional link. Their analysed data was made accessible through a public web application (http://adsn.ddnetbio.com/) that allows easy exploration and sharing.

Del-Aguila et al. compared the transcriptional profiles of Mendelian AD versus sporadic in single-cell resolution, building upon their previous studies that utilised bulk RNA sequencing (Del-Aguila et al., 2019; Z. Li et al., 2018). They sampled post-mortem parietal lobe tissue to extract and profile nuclei from one individual with Mendelian AD (*PSEN1* p.A79V mutation) and two relatives with sporadic AD, using 10X technology. After performing extensive testing of different clustering and data integration techniques to reduce biases introduced by batch effects, they annotated six cell types, similarly to previous studies (neurons, astrocytes, oligodendrocytes, microglia, oligodendrocyte precursor cells, and endothelial cells). Using their high-resolution data, they sought to identify cell types similar to the disease-associated microglia (DAM) previously reported in mouse studies (Keren-Shaul et al., 2017). They detected 79 human homologs of the 500 known DAM markers, while only five of them were significantly associated with microglial cells in all samples (*EEF1A1*, *GLUL*, *KIAA1217*, *LDLRAD3*, and *SPP1*), leading them to conclude that the number of microglia sequenced was not enough to allow the identification of this signature. Finally, they also created a web application (http://ngi.pub/snuclRNA-seq/) to make their analysed data publicly explorable.

More recently, Mendiola and colleagues developed a novel sequencing strategy, termed ToxSeq, to characterise the transcriptional landscape of CNS innate immune cells that contribute to oxidative injury (Mendiola et al., 2020). Oxidative molecules or reactive oxygen species (ROS) have important biological regulatory roles, but dysregulation of their homeostatic mechanisms, a common feature linked to neurodegeneration, can lead to neurotoxicity. ToxSeq leverages single-cell sequencing technology coupled with cell staining and sorting to selectively profile oxidative stress-producing CNS innate immune cells. The authors generated the first oxidative stress innate immune cell atlas in neuroinflammatory disease, by applying ToxSeq to profile 8,701 CD11b+ cells from spinal cords of an experimental autoimmune encephalomyelitis (EAE) mouse model — commonly used to recapitulate the pathological hallmarks of MS — and healthy control mice. Clustering analysis identified 14 distinct CD11b+ clusters that can be divided into three larger groups: healthy and ROS-negative, EAE and ROS-negative and EAE and ROS-positive. Interestingly, ROS- cells consisted only of CNS-resident microglia, while ROS+ cells included microglia (approximately 15%) and peripheral immune cells (approximately 50%), mostly macrophages and monocytes. In addition, no ROS+ cells were identified in healthy spinal cord clusters. A differential gene expression analysis indicated heterogeneous disease response of the ROS+ cells, a common observation of similar studies. Gene ontology (GO) analyses identified a subcluster of activated microglia enriched with oxidative stress genes; these microglia showed increased activation of pathways relevant to oxidative stress, coagulation and antigen presentation and had the lowest expression of homeostatic markers. The authors then developed a fibrin-induced high-throughput drug screening assay to query 1,907 compounds that can potentially inhibit microglia activation without toxicity. 31 of these showed promising results and were included in follow-up studies to investigate their mechanism of action. In silico analyses identified acivicin as the most promising therapeutic molecule, a drug that inhibits the glutathione degrading enzyme GGT. A final experiment using three different demyelinating mouse models (relapsing-remitting, chronic and chronic progressive EAE, and LPS injection directly into the substantia nigra) indicated that acivicin treatment was successful at mitigating the negative effects of neurodegeneration. In summary, this important manuscript underlines the potential of novel transcriptomics approaches, when combined with complementary assays and powerful bioinformatics algorithms, to not only characterise disease mechanisms but also functionally dissect disease pathology and fundamentally contribute to rational drug design.

Single-cell studies have focused on immunological mechanisms and identified interesting microglia activation patterns. A common feature of all diseases studied is the heterogeneity of cellular response. Microglia have been shown to adopt distinct phenotypes in affected tissue; some of the cells assume a homeostatic role, which is thought to be beneficial for disease modulation, while others adopt a toxic phenotype that is implicated in neurodegeneration and inflammation. Transcriptional perturbations have, also, been shown to be cell-population specific in the initial stages of the disease, while common pathways were activated during the later stages, associated with stress response, autophagy, and apoptosis. While similarities are evident between other neurodegenerative diseases, we can only hypothesise that a similar pattern exists in prion diseases, since no single-cell data is currently available.

## 1.5 Towards a finer resolution in prion transcriptomics

As described in the previous section, the Alzheimer’s disease field was the first in neurodegenerative diseases to investigate specific cell populations with studies by Mathys et al. and Keren-Shaul et al. pushing the resolution of transcriptomics and profiling sorted microglia in 2017. More high-throughput and cell-type unbiased studies were soon to follow in 2019 and 2020, characterising Alzheimer’s disease, multiple sclerosis and experimental autoimmune encephalomyelitis and setting the stage for more targeted research. It is, thus, enigmatic why a thorough literature review suggests that no significant steps towards whole-transcriptome studies at a finer resolution have been made in the field of prion diseases, even though excellent models do exist (see section

4.1.1 for further information about prion disease mouse models). Some speculations are that it could be due to insufficient funding — single-cell experiments are very costly —, technical difficulties due to prion infectivity or unavailability of appropriate tissue samples. Whatever the reasons might be, the prion field is still lacking behind in the field of single-cell omics. However, there have been a couple of attempts to target specific cell types, either through targeted transcriptomics of a selected cell population or in a genome-wide and cell-type-specific manner.

Some of the cell-population-specific studies focused entirely on astrocytes. Even though they are the most abundant glial cells in the CNS and their physiological functions have been well characterised, their involvement in neurodegeneration has generally been understudied. There has been accumulating evidence that astrocytes have pivotal roles in chronic neurodegenerative diseases and acute trauma, while their neuroprotective versus neurotoxic potential is heavily debated (Liddelow & Barres, 2017; K. Li et al., 2019). Recent studies have suggested that activated astrocytes can adopt at least two opposing phenotypes, termed A1 and A2 in analogy to the M1 and M2 phenotype categories of macrophages (Liddelow et al., 2017). A1 astrocytes are associated with neural inflammation and are considered to contribute to neurodegeneration by producing neurotoxins such as INF-γ, C1q, and Lcn2, while A2 astrocytes are produced after ischemia and have neuroprotective action by releasing neurotrophic factors such as BDNF, VEGF, and bFGF. Liddelow et al. demonstrated that it is microglia that induces these astrocytic phenotypes and identified a gene panel that includes A1 and A2 markers and pan-reactive markers that are common for all activated astrocytes.

In Alzheimer's disease, astrocytic activation and dysfunction have been implicated with interference with amyloid-beta clearance, calcium excitotoxicity and GABA signalling, and the release of pro-inflammatory cytokines (Acosta et al., 2017; Rossi & Volterra, 2009; Salminen et al., 2008; Vincent et al., 2010). Reactive astrocytes have also been identified in Parkinson's disease. Evidence suggests astrocytic activation initiates the recruitment of microglia and is linked with neuroinflammation, while α-synuclein has been shown to accumulate intracellularly, disrupting astrocytic glutamate regulation and the reciprocal communication between neurons and astrocytes, which is shown to be of major importance to neuronal health (Barcia et al., 2012; Gu et al., 2010; Halliday & Stevens, 2011; Hirsch & Hunot, 2009). In ALS, astrocytes have been shown to contribute to motor neuron death with the degree of their reactivity correlating with neurodegeneration (K. Li et al., 2019; Pehar et al., 2017). Finally, astrocytes are recognised as key players in MS, modulating lesion formation and evolution, and the creation of the glial scar once the inflammation has subsided (Ponath et al., 2018).

While astrogliosis is one of the hallmarks of prion disease pathology, the characterisation of the reactivity state of astrocytes in prion mouse models and human prion diseases was only recently accomplished by a study published in 2019 (Hartmann et al., 2019). Hartmann and colleagues first used immunohistochemistry to demonstrate the abundance of A1 activated astrocytes in both RML-infected mouse brain samples and human brain samples from sCJD patients. They then used a triple-KO mouse model that fails to develop A1 astrocytes and identified a novel astrocytic polarisation profile in terminally sick RML-inoculated mice, termed C3$^{+}$-PrP$^{Sc}$-reactive-astrocytes, which is characterised by the expression of only some of the pan-reactive, A1-specific and A2-specific markers, suggesting a mixed astrocyte activation phenotype. Interestingly, the triple-KO mice experience an accelerated disease course with a decreased survival time, suggesting a protective role of A1 reactive astrocytes, which might confound, though, by an altered microglial response.

Building upon those findings Ugalde and colleagues investigated the correlation of A1-specific astrocyte markers with specific molecular subtypes of sCJD (Ugalde et al., 2020). For this study, the researchers quantified the expression of two A1 marker genes, *C3* and *GBP2*, in the frontal cortex of 35 sCJD patients and 8 healthy controls. They were able to confirm that the expression of both genes was elevated in disease, while the levels of *C3* expression stratified to codon 129 genotype with its expression found to be highest in homozygous methionine and lowest in homozygous valine patients. Regarding *GBP2*, they observed a positive correlation between the logarithm of its expression and disease duration. Overall, their findings highlight the interplay between a spectrum of astrocytic activation and patient-specific disease parameters.

Microglia are the resident immune cells of the CNS, they belong to the glial system and play important roles in maintaining brain homeostasis, neurodevelopment and learning and memory formation, while their impairment has been linked to severe pathological outcomes (S.-K. Chen et al., 2010; Ikegami et al., 2019; Paolicelli et al., 2011; Parkhurst et al., 2013). Microglia can act as sensors of brain pathology and can be activated by stimuli such as neurodegeneration, trauma or infection, assuming an array of phenotypes that can range from pro-inflammatory, characterised by the secretion of cytokines, chemokines, and reactive oxygen species, to anti-inflammatory, which can mediate beneficial effects and is associated with a release of neurotrophic and anti-inflammatory factors (Aguzzi et al., 2013; Cherry et al., 2014). The M1/M2 terminology has been used to describe the cytotoxic and neuroprotective phenotypes, respectively, even though it is becoming clear that these represent the extremes of a spectrum (Y. Tang & Le, 2016).

Neuroinflammation and microglial activation have been more extensively studied in the context of Alzheimer's disease, however, evidence suggests the existence of common activation pathways in neurodegenerative diseases such as Parkinson's disease, ALS, frontotemporal dementia, Huntington's disease and prion diseases (Aguzzi & Zhu, 2017; Heneka et al., 2014). Microglia-mediated neuroinflammation is an important component of PD with M1 activated microglia being identified in close proximity to dopaminergic neurons, while little is known regarding the M2 phenotype (Y. Tang & Le, 2016). Activation of microglia could be attributed to the accumulation of misfolded proteins, environmental factors or pathogens. In AD, studies have shown that microglia adopt mixed activation phenotypes and some subpopulations can be neuroprotective by degrading and reducing the burden of amyloid-beta plaques, while others release pro-inflammatory signals and show increased production of ROS (Meyer-Luehmann et al., 2008; Y. Tang & Le, 2016; D. G. Walker et al., 2006). Similar observations have been made regarding ALS, where microglia subpopulations have been shown to exhibit different gene expression signatures involving both protective and detrimental factors. Recent studies in ALS, AD, EAE and HD underline the importance of the temporal dimension, on top of the spatial, as it is becoming evident that microglia can undergo temporal transformations between disparate activation states (Ajami et al., 2018; B. E. Clarke & Patani, 2020; Mathys et al., 2017).

Microglial activation is a key component of prion diseases and can be easily recapitulated in mouse models upon prion infection (Aguzzi et al., 2013). Activated microglia have been identified in human patients and mouse models using immunohistochemistry since the early nineties, while its activation is observed before the onset of clinical signs and neuronal loss, indicating a driving force of neurodegeneration, instead of a secondary effect (Betmouni et al., 1996; Giese et al., 1998; Sasaki et al., 1993; Williams et al., 1994). A comprehensive study of microglial response by Vincenti and colleagues that used transcriptomics data to profile prion-infected mouse brains was published in 2015 (Vincenti et al., 2015). Their analysis of time-course data indicated that the upregulated genes during disease are expressed predominately by microglia, while isolated microglia from a prion disease mouse model intraperitoneally infected with 79A prions was characterised by a pro-inflammatory signature and an upregulation of genes associated with metabolism and respiratory stress. The following year, Alibhai et al. reported microglial response in prion-infected mouse brain and identified the presence of two distinct phenotypes, a homeostatic phenotype identified across all brain regions, and an innate immune response that was restricted only to sites of neurodegeneration (Alibhai et al., 2016). Overall, microglial response in prion diseases is a complex and dynamic process with activated microglia adopting diverse functions. Microglia respond to prion deposits during the early stages of the disease adopting a phagocytotic phenotype and facilitating $PrP^{Sc}$ removal, while the sustained prion accumulation soon overwhelms the protein recycling mechanisms of the cells, triggering neuronal damage and supporting a microglial switch to a proinflammatory phenotype (Aguzzi & Zhu, 2017).

The aforementioned studies focused on a specific cell type or used targeted approaches, instead of unbiased, whole transcriptome sequencing. The only genome-wide study targeting multiple cell populations was published in 2020 by Aguzzi’s group, which looked at alterations during prion disease progression in transgenic mice in a cell-type-specific manner using translating ribosome affinity purification (TRAP) and ribosome profiling (Scheckel et al., 2020). The researchers generated four transgenic mouse lines expressing tagged ribosomes regulated by Cre recombinase, which was under the control of the Camk2a, Pvalb, Gfap or Cx3cr1 promoters to induce expression specifically in excitatory CamKIIa neurons, inhibitory parvalbumin neurons, astrocytes and microglia, respectively. These mice were then inoculated with RML6 prions (passage 6 of RML strain mouse-adapted scrapie prions) or control brain homogenate and sacrificed at 6 time points: 2, 4, 8, 16, 24 weeks post-inoculation, and at the terminal stage of the disease. After validating the specificity of expression in each of the cell types, the researchers isolated the ribosomes and determined the translation rate of each transcript via ribosome profiling. A differential translation analysis comparing the two experimental groups at each time point highlighted that cell-type-specific changes become evident only at the later stages of the disease. Only 3 transcripts in total were found to be differentially translated during the first 4 time points, while more than 250 were identified at 24 weeks post-inoculation and more than two thousand at the terminal stage of the disease. Interestingly, the authors underline that most of the dysregulated transcripts pertained to astrocyte and microglia populations, while both excitatory and inhibitory neurons were associated with only a fraction of those dysregulated genes. Finally, it is discussed that these transcriptional changes are cell-type specific, with the larger fraction of transcripts being uniquely dysregulated in a cell type. This study is the only one to date that has approached prion transcriptomics in a cell-type-specific and genome-wide manner, although the resolution offered is not fine enough for it to be equated to single-cell transcriptomics. A more extensive discussion of these findings along with a comparison with our data will follow in section 4.2.4.

In summary, there have been attempts to study individual cell populations in the field of prion diseases. Most studies have focused on astrocytes and microglia and used immunohistochemistry or assayed known markers of activation using quantitative PCR. The results highlight the existence of a unifying theme that is recurrent in neurodegenerative diseases, which are characterised by dynamic spatiotemporal activation of astrocytes and microglia, while also underlining the complexity and breadth of pathophysiological cellular phenotypes that fall between a range defined by the two polar extremes of neuroprotection and neurotoxicity. However, none of the studies published has employed unbiased, whole transcriptome approaches like single-cell sequencing to transcriptionally profile all cell populations in human or mouse prion diseases. This is in contrast to other fields like AD, PD and ALS, where such studies have started to uncover interesting and disparate biological functions restricted to specific cell subtypes. It is high time the prion field caught up with innovations in transcriptomics, and this is exactly the aim of our research, which will be thoroughly discussed in the following section.

## 1.6 Hypotheses and aims

Despite substantial research aiming to elucidate prion disease pathogenesis, the underlying mechanisms of cellular toxicity and neurodegeneration are yet to be fully characterised. The transcriptional landscape of the prion-infected human brain, including changes in gene expression profiles related to tissue degeneration, has not been explored in-depth while confounding effects related to cellular heterogeneity have not been accounted for.

Our hypotheses are:

1. Cellular response to prion infection is heterogeneous, i.e., it involves distinct transcriptomic responses from different cell populations and subpopulations, some of which are associated with a homeostatic and others with a toxic phenotype.

2. Prion infection in different systems (cell lines, mouse, human) is associated with distinct but overlapping gene expression patterns. We expect to find commonly dysregulated pathways in mice and humans, even though they might not include the same genes.

3. Prion infection causes selective toxicity to specific cell subpopulations and leads to differences in their abundance.

To elucidate disease mechanisms, we aim to employ single-nucleus methodologies to transcriptionally profile prion-propagating cell lines and prion-infected mouse and human brains. For this study to be successful, preliminary work will have to be done to establish and validate our snRNA-seq protocols. In more detail, the most promising single-cell protocols will be selected by reviewing recent literature. Then, some of the most suitable ones for our use case will be thoroughly reviewed before committing to establishing them in our Institute. Initial experiments will allow us to compare the methods based on their robustness, safety, output, and suitability. We will, finally, select the most optimal approach and fine-tune it to perform best in our research environment. In parallel, we aim to establish a single-cell bioinformatics pipeline that will be essential to explore the generated data and extract meaningful conclusions.

Having established the required methodology, we will proceed to transcriptionally profile a prion-propagating cell line, which will serve as a reference point for future experiments. These experiments will also provide us with valuable experience and allow further optimisations. We will then need to validate our protocols using samples processed and stored in a similar way to our prion-infected material. This will be done by processing a control frozen mouse brain using our experimental and bioinformatics pipelines.

Confident about using our protocols with infectious and more valuable samples, we then aim to single-cell sequence the brain of an RML-prion-infected mouse model. To our knowledge, this will be the first time that the brain of a prion mouse model is transcriptionally characterised in single-cell resolution. These experiments will generate novel insights concerning RML prion disease in mice and will also allow subsequent comparison of cell-specific transcriptomic perturbations between mouse and human prion diseases.

Further experiments will involve case-control studies between sCJD patients and non-neurodegenerative disease controls. We are aiming to transcriptionally characterise human brain biopsies and post-mortem brain tissue using scRNA-seq technology to get a snapshot of the mechanisms involved in the late stages of sCJD. This will also be the first single-cell transcriptomics study of the prion-infected human brain.

Finally, having generated all the data needed, we are aiming to compare the transcriptomic profiles of mouse and human diseases to identify common gene expression patterns which might be involved in neurodegeneration in general.

In summary, this study has the following aims:

1. Review the literature and establish some of the most promising snRNA-seq methodologies in our Institute, validate their performance, and select the most optimal one to be used for subsequent experiments.

2. Transcriptionally profile prion-propagating cell lines using snRNA-seq approaches to identify heterogeneity in prion infection response *in vitro*.
3. Validate the snRNA-seq protocols using uninfected frozen brain tissue.
4. Perform a longitudinal case-control single-cell transcriptomics study of mouse prion disease using an RML-infected mouse model to characterise disease response heterogeneity *in vivo*.
5. Identify sCJD disease mechanisms and transcriptionally characterise human prion diseases by performing a case-control study of human brain tissue using sCJD brain biopsies and non-neurological control biopsies.
6. Identify late sCJD disease mechanisms of toxicity by performing a case-control study of human prion diseases using post-mortem brain tissue from sCJD patients and non-neurological controls.
7. Compare the transcriptomic profiles of mouse and human prion diseases during early and late stages to identify common gene expression patterns.

# 2 Materials and methods

## 2.1 Cell lines

PK1 and iPK1 cells were obtained from Emma Jones and cultured in Opti-MEM (Gibco; 31985-047) with 10% foetal bovine serum (FBS, Gibco; 41965-039) and 1% penicillin-streptomycin (Gibco; 15140-122) (complete Opti-MEM).

## 2.2 Cell culture

PK1 and iPK1 cells were cultured in Opti-MEM (Gibco; 31985-047) with 10% Foetal Bovine Serum (FBS, Gibco; 41965-039) and 1% penicillin-streptomycin (Gibco; 15140-122). HEK-293T cells (ATCC; CRL-3216) were cultured in Dulbecco's Modified Eagle Medium (Gibco; 41965-039) with 10% FBS and 1% penicillin-streptomycin. Cells were cultured in an incubator at 37ºC and 5% $CO_2$. Cells were passaged when 80% confluent by mechanical dissociation and plated after a 1:10 dilution.

## 2.3 Nuclei suspensions preparation

### 2.3.1 Nuclei extraction from tissue culture

10cm tissue culture dishes were removed from the incubator when they reached 70-80% confluence. The supernatant was removed and discarded, and cells were washed gently twice with 1 mL 1x PBS (Thermo Fisher Scientific; 10010023).2 mL of cold Nuclei EZ prep buffer (Sigma-Aldrich; NUC101) was added directly to the cells and cells were scraped with a plastic scraper. The resulting suspension was added to a glass 2 mL dounce tissue homogenizer (Sigma-Aldrich; D8938-1SET) and treated as frozen mouse brain tissue.

### 2.3.2 Nuclei extraction from frozen mouse brain

Each flash-frozen mouse brain was left to partially thaw. The olfactory bulb was removed, and a slice of the frontal lobe was cut and transferred to a glass 2 mL dounce tissue homogenizer on ice. All following steps were carried out on the ice and using ice-cold solutions. All centrifugation steps were performed at 500 g for 5 minutes at 4ºC using a pre-chilled centrifuge unless otherwise specified. 2 mL of Nuclei EZ prep was added, and tissue was homogenised using 20 strokes of the loose and 20 strokes of the tight pestle. The suspension was transferred to a 15 mL tube and 2 mL of Nuclei EZ prep was added. The suspension was incubated for 5 minutes and then centrifuged. The supernatant was discarded, 4 mL of Nuclei EZ prep was added, and the pellet was resuspended using a P1000 pipette. The suspension was incubated for 5 minutes and then centrifuged. The supernatant was discarded, and the pellet was resuspended in 4 mL Nuclei Suspension Buffer (NSB; 1x PBS, 0.01% BSA (Cambridge Bioscience; 227-10210) and 0.1% NxGen RNAse inhibitor (Lucigen; 30281-2)). The suspension was centrifuged, the supernatant discarded, and the pellet resuspended in 1 mL NSB. The suspension was filtered through a 35 um filter (Fisher Scientific; 10585801) and stored on ice.

If following the DroNc-seq protocol, the suspension was diluted to the final concentration. If following the SPLiT-seq protocol, the cells were first fixed and permeabilized and then diluted to the final concentration.

### 2.3.3 Nuclei extraction from frozen post-mortem human brain and human brain biopsies

For the post-mortem samples, each human brain was left to partially thaw and removed from the storage cassette. A small slice of the superior frontal gyrus (approximately 50-100 mg) was cut and transferred to a glass 2 mL dounce tissue homogenizer on ice. For the human biopsies, no structure was visible, and a small slice (approximately 50-100 mg) was transferred to a glass 2 mL dounce tissue homogeniser. All following steps were carried out on the ice and using ice-cold solutions. All centrifugation steps were performed at 500 g for 5 minutes at 4$^{o}$C using a pre-chilled centrifuge unless otherwise specified. 1.5 mL of Nuclei EZ prep was added, and tissue was homogenised using 20 strokes of the loose and 20 strokes of the tight pestle. The suspension was transferred to a 2 mL tube and 0.5 mL of Nuclei EZ prep was added. The suspension was incubated for 5 minutes and then centrifuged. The supernatant was discarded, 2 mL of Nuclei EZ prep were added, and the pellet was resuspended using a P1000 pipette. The suspension was incubated for 5 minutes and then centrifuged. The supernatant was discarded, 0.5 mL of wash buffer (1x PBS, 1% BSA (15260037; Gibco), 0.2 u/µL SUPERase In (AM2694; Invitrogen)) were added without resuspending, and the sample was incubated for 5 min to allow buffer interchange. Then 1.5 mL of wash buffer was added, and the sample was resuspended. The suspension was centrifuged, the supernatant was discarded, the pellet was resuspended in 500 µL wash buffer, and 0.5 mL of 50% OptiPrep Density Gradient Medium solution (D1556; Sigma-Aldrich) was added. The suspension was transferred on top of a 1 mL 29% OptiPrep cushion solution in a new tube and centrifuged at 10,000g for 30 min at 4ºC. The supernatant was discarded, and nuclei were resuspended in 750 µL Parse Nuclei Buffer (from the Parse Evercode WT kit) + 0.75% Bovine Albumin Fraction V (15260037; Gibco). Preparation then proceeded following step 7 of the Parse nuclei fixation protocol (page 15 of the protocol; section 7.5.3).

The following steps were performed using the Parse Evercode nuclei fixation kit (Parse Biosciences) according to the manufacturer's instructions.

### 2.3.4 Nuclei fixation and permeabilization for SPLiT-seq

The following solutions were prepared:

- 1.33% formalin (360 µL of 37% formaldehyde solution (Sigma-Aldrich; 252549) + 9.66 ml 1x PBS)
- 2 mL of 0.5X PBS + 5 µL SUPERase in (Thermo Fisher Scientific; AM2696) + 2.5 µL NxGen RNase inhibitor
- 500uL of 5% Triton X-100 (Sigma-Aldrich; T8787) + 2 µL of SUPERase In
- 1100uL of 100mM Tris pH 8.0 (Thermo Fisher Scientific; AM9855G) + 4 µL SUPERase In

3 mL of 1.33% formalin solution were added to the 1 mL of nuclei suspension. The suspension was incubated on ice for 10 minutes. 169 µL of 5% Triton-X was added to the fixed nuclei, the solution was mixed by pipetting and then incubated on ice for 3 minutes. Nuclei were centrifuged at 500g for 5 minutes at 4ºC, the supernatant was discarded, and the pellet was resuspended in 500 µL cold NSB. 500 µL of cold 100 mM Tris and 20 µL of 5% Triton-X were added. Nuclei were centrifuged again under the same conditions; the supernatant was discarded, and the pellet was resuspended in 400 µL of cold 0.5X PBS. Nuclei were filtered through a 35 um filter and counted.

### 2.3.5 Final dilution for DroNc-seq

Nuclei were counted using a Neubauer Improved C-Chip Disposable Haemocytometer (DHC-N01-50; Cambridge Bioscience) and diluted to a final concentration of 300 nuclei/μL using cold NSB. The nuclei were kept on ice until loaded into the syringe.

### 2.3.6 Final dilution for SPLiT-seq

Nuclei were counted using a Neubauer-Improved haemocytometer and diluted to a final concentration of 2000 nuclei/μL using cold 0.5x PBS supplemented with 0.2 units/μL SUPERase In RNAse inhibitor. The nuclei were stored frozen at -80$^{o}$C until library preparation.

### 2.3.7 Final dilution for Parse Evercode

Fixed nuclei were counted using a Neubauer-Improved haemocytometer and diluted to variable concentrations calculated using the sample loading table provided using cold nuclei suspension buffer provided with the Parse Evercode nuclei fixation kit (Parse Biosciences). The nuclei were stored frozen at -80$^{o}$C until library preparation.

## 2.4 DroNc-seq

For DroNc-seq sequencing, the original protocol was followed (A. Basu et al., 2017; Habib et al., 2017). The complete original protocol can be found in the supplementary materials, section 7.5.1, while a summary of the methodology including potential optimisations is provided below.

Barcoded beads (Chemgenes; Macosko-2011-10) were counted using a Fuchs-Rosenthal haemocytometer (Cambridge Bioscience; DHC-F01-50), washed and filtered as per protocol instructions, and stored at 4$^{o}$C. Before each experiment, an aliquot of 360,000 beads was spun down, the supernatant was removed and the beads were resuspended in 1.2 mL Drop-seq Lysis Buffer (DLB; 4 ml of nuclease-free H2O, 3 ml 20% Ficoll PM-400 (Sigma; F5415-50ML), 100 μL 20% Sarkosyl (2B Scientific; 40120977-1), 400 μL 0.5M EDTA (Thermo Fisher Scientific; AM9260G), 2 ml 1M Tris pH 7.5 (Thermo Fisher Scientific; 15567027), and 500 μL 1M DTT (Sigma-Aldrich; 646563-10X.5ML), DTT is added fresh before every experiment). The suspension was loaded in a 3 mL syringe (Fisher Scientific; 11303040) including a small stirring magnet (VP Scientific; VP772DP-N42-5-2).

Setup followed the DroNc-seq protocol. In summary, 7 mL of droplet generation oil (Bio-Rad Laboratories; 1864005) were loaded in a 10 mL syringe (Fisher Scientific; 15544835). 1.5 mL of cell suspension was loaded in a 3 mL syringe. Syringes were placed in syringe pumps (Linton instrumentation; KDS910) and infusion rates were set according to the protocol (beads and nuclei at 1.5 mL/h and oil at 16 mL/h). Needles (VWR International; 613-5377) and tubing (Scientific Commodities; BB31695-PE/2-100′ Roll) were affixed to the syringes and the microfluidic device (FlowJEM; DroNc-seq device) and the bead stirrer (VP Scientific, #710D2) was turned on. Nuclei, oil, and beads were flown for approximately 22 minutes and monitored for potential clogging of the device. Figure 2.1 is a photo of the working setup.

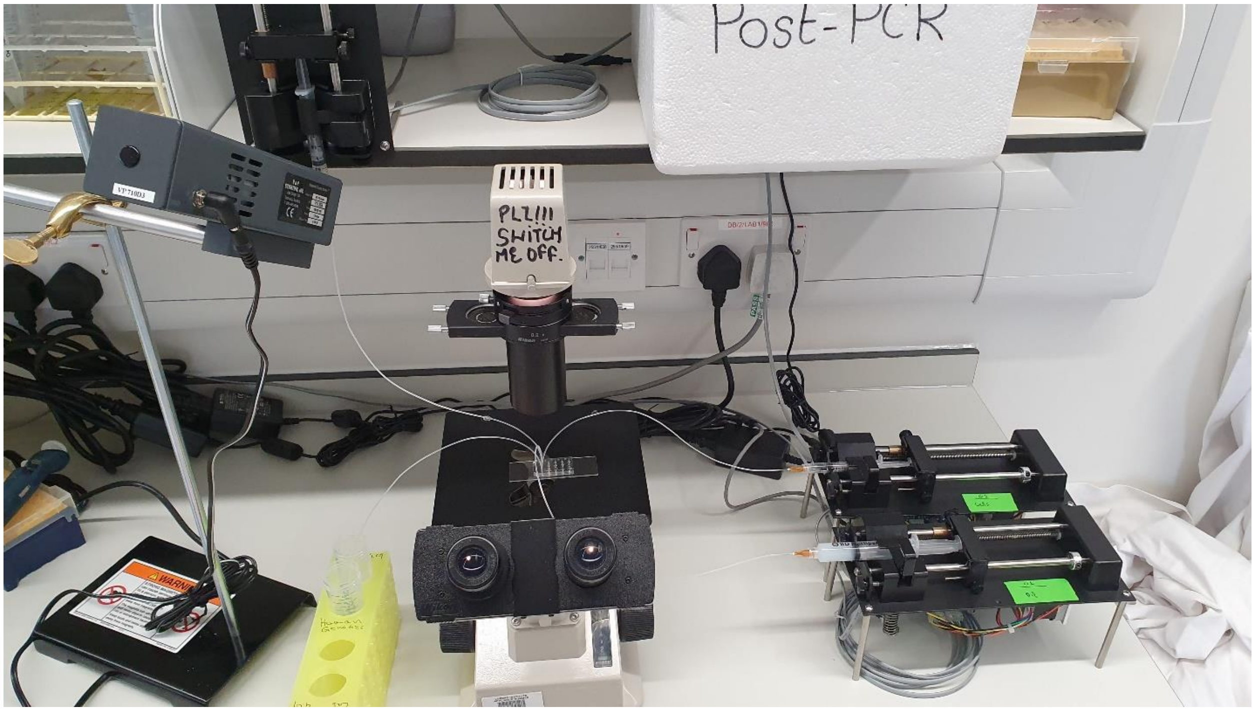


***Figure 2.1: Photo of the working DroNc-seq setup.*** *The three computer-controlled syringe pumps (top-left and right) were connected with tubing to the microfluidic device (centre) that was placed on a brightfield inverted microscope. A bead stirrer prevented the sedimentation of the barcoded beads. Flow rates were controlled by a computer (not shown), and the resulting emulsion was collected in a conical 15 mL tube. The device was constantly monitored for potential clogging.*

The resulting emulsion was collected in a 50 mL Falcon tube and incubated at room temperature for 45 minutes after collection stopped. Droplets were broken after introducing 1 ml of 1H,1H,2H,2H-Perfluorooctan-1-ol (Fisher Scientific; 11490701) and 30 mL 6x SSC (Thermo Fisher Scientific; AM9763) and shaking vigorously. Beads were isolated and washed as per protocol instructions. They were then resuspended in 200 µL reverse transcription mix (80 µL $H_2O$, 40 µL Maxima 5x RT Buffer, 40 µL 20% Ficoll PM-400 (Sigma; F5415-50ML), 20 µL 10 mM dNTP (Takara Bio; 639125), 5 µL NxGen RNase Inhibitor, 10 µL Maxima H-RT enzyme (Fisher; EP0753), and 5 µL 100 µM Template Switch Oligo, AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG (IDT, custom RNA oligo, HPLC purification). The beads were incubated on a rotating incubator for 30 minutes at room temperature and 1.5 hours at 42$^{o}$C.

Beads were washed and treated with exonuclease I (New England Biolabs; M0293L) as per protocol instructions. They were then washed, counted, and resuspended in a PCR mix (24.6 µL $H_2O$, 0.4 µL 100 µM SMART PCR primer, AAGCAGTGGTATCAACGCAGAGT (IDT, custom DNA oligo, standard desalting purification), and 25 µL 2x Kapa HiFi Hotstart Readymix (Kapa Biosystems; KK2602)) in different wells of a PCR plate, each containing 5,000 beads. Samples were amplified using the following PCR programme: 95°C for 3 min; then 4 cycles of 98°C for 20 sec, 65°C for 45 sec, 72°C for 3 min; then 12 cycles of 98°C for 20 sec, 67°C for 20 sec, 72°C for 3 min; and finally, 72°C for 5 min.

PCR products were cleaned with 0.6X Ampure XP beads (Beckman Coulter; A63881). Products were eluted in 15 µL $H_2O$ and a pool of 4 wells was used for library preparation.

The samples were quantified on a TapeStation 2200 (Agilent), using a gDNA tape (Agilent Technologies; 5067-5365) and 500-1000 pg of each was used for tagmentation using the Nextera XT sample prep kit, 96 samples (Illumina; FC-131-1096), and custom primer, AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC, (IDT, custom DNA oligo, HPLC purification), according to manufacturer's instructions. The resulting libraries were analysed on a TapeStation 2200 using a high sensitivity D1000 tape (Agilent Technologies; 5067-5582) and the amount of starting material was optimised so that the resulting tagmented library would have a size of 500-680 bp.

The resulting libraries were sequenced on an Illumina NextSeq 500 using a NextSeq 75 cycle High Output kit (Illumina; 20024911) according to the manufacturer's instructions. The settings used were: paired-end reads, read 1 length: 20 nt, read 2 length: 60 nt, Index 1 length: 8 nt, custom read 1 primer: GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC (IDT, custom DNA oligo, standard desalting).

## 2.5 SPLiT-seq

For SPLiT sequencing, the original protocol (version 3) was followed (Rosenberg et al., 2018). The complete original protocol can be found in the supplementary materials, section 7.5.2, while a summary of the methodology and optimisations is given below.

Barcode plates were ordered from IDT (custom oligos, standard desalting) and stock plates were prepared as per protocol instructions. A list of the barcode names and corresponding sequences is included in the supplementary materials, section 7.6.1.

For the reverse transcription, 4 µL of the first 24 wells of Stock plate 1 were transferred to a new PCR plate on ice. 8 µL of RT mix (per reaction: 4 µL Maxima 5x RT buffer, 0.124 µL NxGen RNAse inhibitor, 0.25 µL SUPERase In, 1 µL 10 mM Takara dNTPs, 2 µL Maxima H minus enzyme, 0.625 µL $H_2O$) was added to each well and then 8 µL of fixed nuclei suspension. The plate was placed in a thermocycler and PCR was carried out as per protocol instructions. All wells were then pooled together, Triton-X was added to a final concentration of 0.1% and the suspension was centrifuged for 3 minutes at 500g. The supernatant was discarded, and nuclei were resuspended in 2 mL 1x NEBuffer 3.1 (New England Biolabs; B7203S) + 20uL NxGen RNase Inhibitor.

For ligation round 1, the ligation mix (1337.5 µL water, 500 µL 10x T4 ligase buffer (included with ligase enzyme), 100 µL T4 DNA Ligase (New England Biolabs; M0202L), 100 µL BSA 10 mg/mL, 12.5 SUPERase In, 40 µL NxGen RNAse inhibitor) was added to the nuclei suspension and into a basin. 40 µL of the suspension were pipetted in each well of the Ligation round 1 barcode plate and the plate was incubated in a plate shaker at 37$^{o}$C for 30 minutes and 300 rpm rotation. Then 10 μL of the Ligation Round 1 blocking solution (316.8 μL 100 uM BC_0216, 300 μL 10x Ligase buffer, 583.2 μL water) were added to each well and the plate was incubated again under the same conditions.

For ligation round 2, the nuclei suspensions were pooled in a 15 mL Falcon and passed through a 40 um strainer to another Falcon. 100 μL T4 DNA ligase was added and the mix was transferred to a basin. 50 μL of the mix was added to each well of the Ligation Round 2 barcode plate. The plate was incubated as previously. Then 20 μL of the Ligation Round 2 blocking solution (369 μL 100 uM BC_0066, 800 μL 0.5 M EDTA, 2031 μL water) was added to each well. The wells were pooled in a 15 mL Falcon and passed through a 40 um strainer into another Falcon.

70 μL 10% Triton-X was added to the mix and nuclei were centrifuged for 5 minutes at 1000g. The supernatant was aspirated, and nuclei were washed with 4 mL wash buffer (4 mL 1x PBS, 40 μL 10% Triton-X, 10 μL SUPERase In) and centrifuged again under the same conditions. The supernatant was aspirated and nuclei were resuspended in 100 μL 1x PBS + 2 μL SUPERase In. Nuclei were counted using a Neubauer-Improved haemocytometer and the desired number of them was aliquoted in each 1.5 mL Eppendorf tube. PBS was used to fill each tube up to 50 μL. Each tube is referred to as a "sublibrary".

For nuclei lysis, 50 μL 2x lysis buffer (final concentrations: 20 mM Tris pH 8, 400 mM NaCl, 100 mM EDTA pH 8, 4.4% SDS (Thermo Fisher Scientific; AM9822)) were added to each tube followed by 10 μL 20 mg/mL Proteinase K. The mix was incubated at 55$^{o}$C for 2 hours with shaking at 300 rpm. Lysates were frozen at -80$^{o}$C and processed the following day.

5 μL 100 uM AEBSF (Abcam; ab141403) was used to stop the proteinase reaction. Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific; 65002) were washed and used to bind the barcoded transcripts as per protocol. Beads were resuspended in 200 μL Template Switch mix (88 μL water, 44 μL 5x Maxima buffer, 44 μL 20% Ficoll PM-400, 22 μL 10 mM Takara dNTPs, 5.5 μL NxGen RNAse inhibitor, 5.5 μL 100 μM Template Switch Oligo, 11 μL Maxima H minus enzyme) and incubated at a rotating incubator at room temperature for 30 minutes and 42$^{o}$C for 1.5 hours.

The sample was then washed and resuspended in PCR mix (121 µL 2x Kapa Hifi Master Mix, 9.68 µL 10 uM BC_0108, 9.68 µL 10 uM BC_0062, water up to 242 µL) and split equally in 4 different wells of a PCR plate. The following PCR programme was then used: 3 min at 95$^{o}$C, then 20 s at 98$^{o}$C, 45 s at 65$^{o}$C, 3 min at 72$^{o}$C for a total of 5 cycles, and then hold at 4$^{o}$C. Reactions were combined in a single tube, cleaned with 0.6x AMPure XP beads as per manufacturer's instructions, eluted in 20 µL water, mixed with 180 µL of the same PCR mix and split in 4 wells (50 µL per well). 2.5 µL EvaGreen (Biotium; #31000) was added to each well and amplification continued in a QuantStudio 12K Flex qPCR machine (Thermo Fisher Scientific) until the signal plateaued out of exponential amplification using the following programme: 3 min at 95$^{o}$C, then 20 s at 98$^{o}$C, 20 s at 67$^{o}$C, 3 min at 72$^{o}$C until signal plateaus out of exponential amplification, then 5 min at 72$^{o}$C, hold at 4$^{o}$C. Reactions were combined in a single tube, cleaned with 0.6x AMPure XP beads as per manufacturer's instructions, eluted in 10 µL water and analysed at TapeStation 2200 using a gDNA tape.

600 pg of each sample was used for tagmentation using the Nextera XT sample prep kit using custom primers one of BC_0076-BC_0083 and BC_0118, according to the manufacturer's instructions. The resulting libraries were analysed on a TapeStation 2200 using a high sensitivity D1000 tape or a high sensitivity D5000 tape.

The resulting libraries were sequenced on an Illumina NextSeq 500 using a NextSeq 150 cycle Mid Output kit (Illumina; 20024904) according to the manufacturer's instructions. The settings used were: paired-end reads, read 1 length: 66 nt, read 2 length: 94 nt, and index 1 length: 6 nt.

## 2.6 Evercode Whole Transcriptome

Evercode WT (whole transcriptome) is the proprietary and optimised protocol that evolved from SPLiT-seq. The methodology is closely related to that of SPLiT-seq, with a few differences. The Parse Evercode Whole Transcriptome kit (Parse Biosciences) contains all consumables and a detailed protocol that includes all steps from nuclei fixation up to sequencing, including catalogue numbers of reagents. All steps were carried out according to the protocol, which can be found in the supplementary materials, section 7.5.3.

Suspension preparation and split-pool barcoding were carried out in a BSL-3 laboratory. To move the sample to a BSL 2 laboratory, the following prion decontamination procedure was followed: at the end of the Parse Evercode WT user manual chapter 3.4, the resulting solution was incubated with 3 volumes of TRI-reagent at room temperature for 2 hours (R2050-1-50; Zymo Research). The mix was then transferred out of the BSL-3 facilities and nucleic acids were purified using the Direct-zol DNA/RNA miniprep kit (R2080; Zymo Research). The kit columns were substituted with Zymo-Spin IC Columns (C1004-50; Zymo Research) so that smaller elution volumes could be used, following the advice of Zymo customer support. 11 µL of the RNA fraction and 10 µL of the DNA fraction were eluted in the same tube. The resulting solution of 21 µL was used for the PCRs starting at section 3.5 of the Parse Evercode WT user manual.

## 2.7 Bulk RNA sequencing of iPK1 and PK1 cells

Cells were harvested when 80-90% confluent. Cells were washed with 1x PBS twice, dissociated by pipetting and pelleted. Pellet was resuspended in PBS and cells were counted using a Neubauer-Improved haemocytometer. 1 million cells were aliquoted in a separate tube and processed using the Direct-zol RNA Miniprep kit (Zymo Research; R2051), according to the manufacturer's instructions. RNA was analysed at TapeStation 2200 using an RNA tape (Agilent Technologies; 5067-5576) and then rRNA was removed using the RiboZero Gold kit (Illumina; MRZG12324) according to the manufacturer's instructions. RNA was then concentrated using RNA Clean and Concentrator (Zymo Research; R1013) and analysed again at TapeStation 2200. Two sequencing libraries were prepared using the TruSeq Stranded Total RNA library preparation kit (Illumina; 20020596) according to the manufacturer's instructions. Libraries were multiplexed using different indices and mixed in equal amounts before sequencing.

The final product was sequenced on an Illumina NextSeq 500 using a NextSeq 75 cycle High Output kit. The settings used were: paired-end reads, read 1 length: 43 nt, read 2 length: 43 nt, and index 1 length: 6 nt.

## 2.8 Data analysis

The version of R and all R packages used can be found in the supplementary materials, section 7.8. R scripts to reproduce the analysis can be found in section 7.7.

### 2.8.1 Sequencing quality control

Fastq files were subjected to quality control using a dockerised version of FastQC (Andrews, 2010) pulled from the repository biocontainers/fastqc:v0.11.9_cv8. The generated reports were manually examined for sequencing quality.

### 2.8.2 DroNc-seq data to count matrix

The fastq files were processed with open-source Drop-seq tools (https://github.com/broadinstitute/Drop-seq), following the original Drop-seq Alignment Cookbook found in the same GitHub repository. A copy of the document can be found in the supplementary materials, section 7.5.4. Transcriptomes GRCm38 (mm10) were used for aligning mouse data and GRCh38 (hg38) for human data. The count matrix generated was used as an input for the subsequent analyses.

### 2.8.3 SPLiT-seq data to count matrix

The fastq files were aligned to the human transcript using the STAR aligner (Dobin et al., 2013). The resulting sam files were converted to a binary format and were processed by the SPLiT-seq bioinformatics open-source tools (https://github.com/yjzhang/split-seq-pipeline) to generate a count matrix. Transcriptomes GRCm38 (mm10) and GRCm39 (mm39) were used for aligning mouse data and GRCh38 (hg38) for human data. The count matrix generated was used as an input for the subsequent analyses.

### 2.8.4 Parse Evercode data to count matrix

The fastq files were processed using the Parse Biosciences pipeline v0.9.6 to generate the count matrix. The pipeline is provided to registered users and requires authentication to be accessed, so no direct link is available. Its function and processes are similar to the SPLiT-seq open-source tools, and the final output is a count matrix that is used for further analyses.

### 2.8.5 Single-cell data analysis

#### *2.8.5.1 Analysis of pilot experiments*

Data analysis followed the best practices of the community as described in the Orchestrating Single-Cell Analysis with Bioconductor online book (Amezquita et al., 2020). A summary of the methodology is provided below.

These steps of the analysis were carried out using the R programming language and utilising Bioconductor packages (Gentleman et al., 2004). The count matrix was imported into R and used to create the object of class SingleCellExperiment. This object was manipulated for Quality Control, where outliers nuclei were filtered out based on the number of genes identified, and then R packages Scran (Lun, McCarthy, et al., 2016) and Scater (McCarthy et al., 2017) were used to normalise the gene counts and model the mean-variance relationship using a zero-inflated negative binomial distribution. Reduced dimensions were calculated for Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and t-Stochastic Neighbour Embedding (t-SNE) visualisations. All plots were drawn using ggplot2 (Gómez-Rubio, 2017). Finally, graph-based clustering was used to separate cell clusters and identify gene markers that drive these distinctions. These markers can be used for cell subpopulation and cell cycle annotation. We generated lists compatible with package scCATCH, which was used for automatic cell annotation (Shao et al., 2020).

### *2.8.5.2 Analysis of mouse and human experiments*

For the analysis of mouse and human data, a pipeline based on the Seurat v4 R package was employed, following the official vignettes and recommendations (A. Butler et al., 2018; Hao et al., 2021; Stuart et al., 2019). A summary of the methodology is provided below, while the analysis scripts can be found in the supplementary materials, section 7.5.

The count matrices generated were first used to create Seurat objects and relevant metadata was added. Then Ensembl IDs were converted to gene symbols using EnsDb version 104 for both the mouse and human data.

**Quality Control**

For quality control, the cells were filtered on the number of features to exclude low-quality cells and possible duplicates with a low threshold of 250 and a high of 2500. The percentage of mitochondrial genes was calculated and cells with more than 1% mitochondrial genes were discarded. A cell cycling score for the S and G2/M phases was assigned using known cell cycling genes (*MCM5, PCNA, TYMS, FEN1, MCM7, MCM4, RRM1, UNG, GINS2, MCM6, CDCA7, DTL, PRIM1, UHRF1, CENPU, HELLS, RFC2,* *POLR1B, NASP, RAD51AP1, GMNN, WDR76, SLBP, CCNE2, UBR7, POLD3, MSH2, ATAD2, RAD51, RRM2, CDC45, CDC6, EXO1, TIPIN, DSCC1, BLM, CASP8AP2, USP1, CLSPN, POLA1, CHAF1B, MRPL36, E2F8* as gene markers for the S phase, and *HMGB2, CDK1, NUSAP1, UBE2C, BIRC5, TPX2, TOP2A, NDC80, CKS2, NUF2, MKI67, CENPF, TACC3, PIMREG, SMC4, CCNB2, CKAP2L, CKAP2, AURKB, BUB1, KIF11, ANP32E, TUBB4B, GTSE1, KIF20B, HJURP, CDCA3, JPT1, CDC20, TTK, CDC25C, KIF2C, RANGAP1, NCAPD2, DLGAP5, CDCA2, CDCA8, ECT2, KIF23, HMMR, AURKA, PSRC1, ANLN, LBR, CKAP5, CENPE, CTCF, NEK2, G2E3, GAS2L3, CBX5, CENPA* as gene markers for the G2/M phase), and cell separation based on cell cycle was assessed by examining the PCA plots. The cell cycle was not regressed.

### Normalisation

The Seurat object was then split by experimental group (CD1, RML, PBS for the mouse experiment, sCJD and Control for the human experiment) and individual objects were normalised using SCTransform. The objects were then combined in one integrated object by first selecting the integration features and finding integration anchors. The integrated object was then annotated using label transfer from an annotated reference dataset.

### Annotation/Label transfer

For the annotation of the mouse data, the first step was to pre-process the reference data to be used for label transfer and cluster annotation. We used the published SPLiT-seq mouse data as a reference as it is well annotated and perfectly matches the sequencing methodology. Postnatal days 2 and 11 data obtained from mouse brain was downloaded from GEO (Sample GSM3017261) and filtered to include only anatomical regions that are found in the frontal lobe. We then used Seurat to normalise the datasets using SCTransform, select the integration features using the top 3000 variable features and prepare the integration anchors. The dataset was integrated, and a Principal Component Analysis was used to identify the first 50 PCs. The FindTransferAnchors and TransferData functions were used to identify data transfer anchors and transfer cell type metadata from the annotated reference to our datasets. The predicted cluster scores and mapping scores were visualised by generating histograms and clusters consisting of less than 100 cells were removed. The data was normalised again using SCTransform and PCs and UMAP coordinates were calculated.

The success of the reference data label transfer was assessed by plotting the expression of known marker genes in each cell type (*Aqp4, Slc1a2, Plpp3, Gja1* for astrocytes*; Mbp, Plp1* for oligodendrocytes; *Vcan, Mbp, Pdgfra* for oligodendrocyte precursor cells; *Rgs5, Flt1, Ly6c1, Pltp* for endothelial/smooth muscle cells; *Dock2, Dock8, Csf1r, P2ry12* for microglia/macrophages*; Dnah11* for ependymal cells; *Gria1, Snhg11* for neurons), and statistics such as the number of cells, the mean of features and the mean of counts in each cluster were calculated.

### Cell type proportions

To investigate changes in cell-type proportions two different approaches were followed. One approach was to calculate the percentage of each cell type (numbers of cells in specific cell type / total number of cells in time point) and plot the result using ggplot2. The other approach was to use scProportionTest, a small library in R that compares cell proportions between conditions using a Monte-Carlo permutation test and testing the null hypothesis that the difference in cell proportions for each cluster between the two conditions is a consequence of random sampling a subset of cells in each condition (https://github.com/rpolicastro/scProportionTest; (Miller et al., 2020)). To generate the null distribution, it pools the cells of both samples together and then randomly segregates the cells back to two conditions while maintaining sample sizes. It then calculates the proportional difference between the two conditions and compares it to the observed proportional difference for each cluster. This process is repeated 10,000 times and the p-value is calculated by taking the number of simulations where the proportional difference was as or more extreme than the observed one, over the total number of simulations.

### Differential gene expression using Seurat

Differential gene expression between the same two clusters across different conditions was performed using Seurat’s FindMarkers function. The statistical test used was the non-parametric Wilcoxon rank-sum test and the adjusted p-value was based on

Bonferroni correction using all features in the dataset. The differentially expressed genes were then filtered, keeping the ones that had an adjusted p-value of less than 0.05.

For the mouse dataset, to generate the differentially expressed gene lists of the RML versus CD1 groups, initially, a comparison between CD1 and PBS groups was done to identify genes that are shown to be dysregulated but could be due to technical noise or relevant to the inoculation with a brain homogenate and not prion specific. From the list of those genes, we selected genes that were identified in multiple clusters (more than 5) and excluded them from the RML vs CD1 comparison to reduce technical noise. This resulted in 7 excluded genes which were: *Calm1, Cdk8, Cmss1, Malat1, mt-Rnr1, mt-Rnr2,* and *Rn18s*.

For the human dataset, a full comparison of sCJD vs controls was done, without the exclusion of any genes.

### Differential gene expression using pseudobulk methods

To strengthen our findings, we also performed differential gene expression on aggregated, pseudobulk data using DESeq2 (Love et al., 2014) and glmGamPoi (Ahlmann-Eltze & Huber, 2021).

For the mouse dataset, we first subset the data to isolate a specific cell cluster from a specific time point. We then summed the gene counts for all cells of the cluster from each animal separately. That created bulk-sequencing-like data, where for a specific cell cluster we had information on the expression of features from each animal. We then used DESeq2 with the experimental design = ~ inocula, or glmGamPoi to fit a Gamma-Poisson model and compared the expression between the RML vs CD1 group using the Wald test, or a quasi-likelihood ratio test, respectively. Log2-fold change values were shrunk using the apeglm function (A. Zhu et al., 2019). p-values were corrected using the Benjamini and Hochberg method. The same process was then repeated for each cell cluster at each time point separately. The 7 genes previously identified as technical noise (*Calm1, Cdk8, Cmss1, Malat1, mt-Rnr1, mt-Rnr2,* and *Rn18s*) were also excluded from the final lists.

### Gene Ontology over-representation analysis (ORA) and Gene Set Enrichment Analysis (GSEA)

We used clusterProfiler in R for both ORA and GSEA (Wu et al., 2021). For the ORA we used the enrichGO function using all the available features in the dataset as the gene universe and the filtered differentially expressed genes as the query genes. The adjusted p-values were calculated using the Benjamini-Hochberg method. GO terms that were supported by less than 3 genes were filtered out. For the mouse dataset, we used the AH92582 annotation database, while for the human dataset we used the AH95744 annotation database.

For the GSEA we used the gseGO function with the same organism databases and set the minimal size of each gene set to 10, the maximal size of genes annotated for testing to 500, and the p-value cut-off to 0.05. The adjusted p-values were calculated using the Benjamini-Hochberg method.

### 2.8.6 Bulk RNA sequencing data analysis

Bulk RNA sequencing analysis followed the community best practices and recent workflow standards using Bioconductor packages (Love et al., 2015). In summary, fastq files were aligned to the mouse genome GRCm38 (mm10), with annotations provided, using tophat2 (D. Kim et al., 2013). The bam files were sorted and indexed using samtools (H. Li et al., 2009). GenomicAlignments (Lawrence et al., 2013) was used for read counting and the creation of the SummarizedExperiment object. DESeq2 (Love et al., 2014) was used to log-transform, normalise data and produce the normalised counts. Due to the availability of only two samples to be compared (PK1 and iPK1 cells), no statistical tests were deemed suitable. Thus, the ratio of PK1 over iPK1 normalised counts was calculated and a list of the top 2000 genes showing the highest differential expression, either upregulation or downregulation, was generated.

### 2.8.7 Bulk RNA-seq and SPLiT-seq data correlation

For the comparison of single-cell and bulk sequencing data to be possible, we generated pseudo-bulk data from the single-cell experiment by summing the expression of each gene across all cells. The resulting data frame was used to generate a SummarizedExperiment object. Bulk sequencing and pseudo-bulk sequencing SummarizedExperiments were integrated, and data were log-transformed and normalised. The normalised counts of the top 2000 differentially expressed genes (as selected previously) were extracted and plotted. Finally, the Pearson correlation coefficient between normalised bulk counts and normalised single-cell counts was calculated.

## 2.9 RNA extraction from single nuclei suspensions

Suspensions of 500,000-800,000 nuclei in total were mixed with 3 volumes of RTI Reagent (included in R2063; Zymo Research) and then total RNA was extracted using the Direct-zol RNA Microprep kit (R2063; Zymo Research) according to manufacturer's instructions, including the DNase I treatment step. RNA was eluted in 20 µL of $H_2O$, visualised for quality control on a 2200 TapeStation (Agilent) using High Sensitivity RNA Screen Tapes (5067-5579; Agilent), quantified using the Qubit RNA High Sensitivity Assay (Q32852; Invitrogen), and stored at -80$^{o}$C until further processing.

RNA extraction was performed with the help of Emmanuelle Vire.

## 2.10 Reverse transcription

200 ng of RNA were processed using the QuantiTect Reverse Transcription Kit (205313; Qiagen) in two separate reactions of 100 ng RNA each, according to the manufacturer's instructions. The resulting cDNA of the two reactions (40 µL in total) was pooled together, diluted to approximately 250 µL using $H_2O$, and stored at -20$^{o}$C.

Reverse transcription was performed by Emmanuelle Vire.

## 2.11 Real-time PCR

Real-time PCR was performed using a protocol based on the Fast SYBR Green reagent and following the manufacturer's instructions. Briefly, a master mix containing 10 µL of Fast SYBR Green Master Mix, 2 µL of 10X forward and reverse primer mix (the exact concentration of the primers is proprietary. See section 7.6.2 for a list of all primers used), and 6 µL of $H_2O$ per reaction was prepared and 18 µL were distributed to each well of a MicroAmp Fast Optical 96-Well Reaction Plate (4346906; Applied Biosystems). 2 µL of cDNA template was added, the plate was sealed using MicroAmp Optical Adhesive Films (4311971; Applied Biosystems), vortexed briefly and spun down. Reagents and plates were kept on ice. Real-time PCR was performed on a QuantStudio 3 Real-Time PCR

System (A28567; Applied Biosystem) operated at Fast mode using the following cycling conditions: 20 sec at 95$^{o}$C; 40 cycles of 3 sec at 95$^{o}$C, 30 sec at 60$^{o}$C.

For data analysis, the Sequence Detection System software was used to automatically determine the threshold cycles for the amplification curves ($C_T$), and the relative quantification method (comparative or $\Delta\Delta C_T$ method) was used to measure gene expression of samples of the RML group relative to samples of the CD1 group. Data analysis was performed following the guidelines of published literature (S. C. Taylor et al., 2019). First, the mean of technical replicates was calculated, removing outlier samples. Then, the average Ct for the 20 dpi CD1 group was calculated and used to calculate the relative difference between the control group and the mean per individual sample ($\Delta C_T$). The relative quantities were calculated from the $\Delta C_T$, assuming a reaction efficiency of 100% using the formula: RQ = $2^\wedge\Delta C_T$. For each inoculum/time point combination, a normalization factor is determined from the geometric mean of the 2 endogenous controls, *Tubb4a* and *Sdha*, selected for their high levels of expression and low variability in the snRNA-seq data and constant expression in the real-time PCR data (Supplementary Figure 1). The relative normalised expression is then calculated per sample by dividing the relative quantity by the normalisation factor.

Real-time PCRs were performed by Emmanuelle Vire and Tom Trainer.

## 2.12 Single-cell transcriptomics of murine prion disease

### 2.12.1 Mouse experiment 324

4–6-week-old female FVB inbred mice were ordered from Envigo (FVB/NHan®Hsd; Order code: 862) and left to be acclimatised for one week. The animals were then chipped and inoculated when around 6-8 weeks old.

Inoculations on anaesthetised were conducted intracerebrally in the right parietal lobe with 30 µL of one of the following preparations:

- For the RML group (inoculum code I21742): 1% RML prion-infected brain homogenate prepared from 10% I17700 RML stock.
- For the CD1 group (inoculum code I21744): 1% uninfected CD1 brain homogenate prepared from 10% I14040 uninfected CD1 stock.

- For the PBS group (inoculum code I56): 1x sterile DPBS (Gibco; 14190-086)

The number of mice in each group and time point are given below where rows represent the 3 different groups (CD1, RML and PBS-inoculated mice) and columns represent the 5 time points. The disease end-stage is defined as the day when scrapie sickness is confirmed.

| | **20 dpi** | **40 dpi** | **80 dpi** | **120 dpi** | **End-stage** |
|---|---|---|---|---|---|
| **CD1** | 15 | 15 | 15 | 15 | 15 |
| **RML** | 15 | 15 | 15 | 15 | 15 |
| **PBS** | 5 | 5 | 5 | 5 | 5 |

Mice were monitored daily for neurological signs of the disease and were culled by $CO_2$ exposure either on schedule for the 20, 40, 80 and 120 dpi time points or at scrapie sickness confirmation for the end-stage according to the animal research guidelines. Early signs include erect ears, rigid tail, piloerection, ungroomed appearance, slightly hunched posture, and clasping of hind limbs when lifted. Scrapie was confirmed when signs of ataxia, generalized tremor, loss of righting reflex, or limb paralysis were observed (O'Shea et al., 2008).

When culling, the brain was removed and the left hemisphere was stored in 10% formal saline for further histopathological analysis, while the right was snap-frozen and stored at -80$^{o}$C until further processing. Blood was collected and split into two aliquots. One of those was stored in PAXgene Blood RNA Tubes (BD biosciences; 762165) and frozen at -80$^{o}$C to be used in future whole-blood transcriptomics studies, and the other aliquot was centrifuged, and plasma was collected and stored at -80$^{o}$C.

All animal work was performed by staff at the animal facility including Nick Kaye and Craig Fitzhugh under approval and license granted by the UK Home Office (Animals (Scientific Procedures) Act 1986), which conformed to UCL institutional and Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines. Experimental design adhered to the principles of the 3Rs - Replacement, Reduction and Refinement. Animal ordering was mediated by Lucy Draper.

### 2.12.2 Immunohistochemistry for prion-related neuropathology

Immunohistochemistry was performed as previously described with modifications (Wadsworth et al., 2021). Briefly, mouse brains were fixed in 10% buffered formal saline and paraffin wax embedded. Serial sections of 5 um were taken and deparaffinised. The sections were then processed to investigate PrP deposition on a Ventana Discovery XT automated IHC staining machine (Roche Tissue Diagnostics) using protocols developed on a Ventana Benchmark staining machine (Wadsworth et al., 2017). Sections were treated with cell conditioning solution (Discovery CC1; Roche Tissue Diagnostics) at 95°C for 60 minutes or with a medium concentration of protease (Protease 1; Roche Tissue Diagnostics) for 4 minutes. For PrP deposition anti-PrP monoclonal antibodies ICSM35 were used in conjunction with biotinylated polyclonal rabbit anti-mouse immunoglobulin secondary antibodies (Dako; Agilent) and Ventana proprietary detection reagents utilizing 3,3′-diaminobenzidine tetrahydrochloride as the chromogen (DAB Map Detection Kit; Roche Tissue Diagnostics).

For haematoxylin and eosin (H&E) staining conventional methods on a Gemini AS Automated Slide Stainer (Thermo Fisher Scientific) were used. Positive controls for the staining technique were used throughout. All slides were digitally scanned on a Hamamatsu NanoZoomer 360 instrument, and images were captured from the NDP.serve3 software (NanoZoomer Digital Pathology) and composed with Adobe Photoshop.

Immunohistochemistry was performed by Tamsin Nazari, Florin Pintilli, Conor Preston, Fabio Argentina, and Jackie Linehan and analysed by Prof. Sebastian Brandner.

### 2.12.3 Brain homogenisation

2 mL screw-cap tubes with a conical bottom (Alpha Laboratories; CP5932) were filled with ribolysing beads (Fisher Scientific; 15515809) to cover the bevelled bottom of the tube and weighted. One frozen right mouse brain hemisphere was transferred into each tube and tubes were weighed again to calculate the mass of the brain. An appropriate volume of PBS was then added (Gibco; 14190-086) to prepare a 20% w/v homogenate (x4 the brain mass, assuming that brain tissue density is close to 1). The tubes were then tightly screwed, and tissue was homogenised in a Precellys Evolution homogeniser

(Bertin Instruments; P000062-PEVO0-A) operated at 6500 rpm for 45 seconds. The tubes were then left at 4$^{o}$C for 1 h to reduce frothing. 500 µL of homogenate was pipetted out and transferred to a new tube, where it was diluted to 10% w/v using PBS. Homogenates were stored at -80$^{o}$C.

All sample handling procedures took place within a class 1 microbiological safety cabinet.

### 2.12.4 Scrapie cell assay

The scrapie cell assay was performed as previously described, in an automated manner (Klöhn et al., 2003). Briefly, the cell lines were seeded at 1.8*10$^{4}$ cells / well in a 96-well plate, 24 hours before infection with 10% w/v RML brain homogenate at the following dilutions: 3*10$^{-6}$, 10$^{-6}$, 3*10$^{-7}$, 10$^{-7}$, 3*10$^{-8}$, 10$^{-8}$. The cells were then split using an automated liquid handling robot (Beckman Coulter; Biomek FX) every three to four days and assays after the third and fourth passages. 25,000 cells were plated on ELISpot IP Filter Plates (PVDF membrane, 0.45 um, Merck; MSIPN4550) and fixed at 50$^{o}$C for 1 h before treatment with 1 ug/ml proteinase K (Roche; 3115828001) in lysis buffer (50 mM Tris HCl pH 8, 150 mM NaCl, 0.5% w/v sodium deoxycholate, 0.5% v/v Triton X-100) at 40$^{o}$C for 1 h. Plates were washed and treated with 3M guanidine thiocyanate (Melford; G54000) for decontamination and antigen retrieval before blocking with SuperBlock blocking buffer (Thermo; 37545). Staining was performed using an anti-PrP antibody (clone ICSM18; D-Gen Ltd; Table 2) followed by detection with alkaline-phosphatase-linked anti-IgG1 antiserum (Southern Biotech; 1070-04). Spots were visualised with alkaline phosphatase conjugate substrate (Bio-Rad; 170-6432) and PK-resistant infected cells were counted using the Bioreader 5000-Eβ (BioSys Karben, Germany). All assays were performed by Christian Schmidt, George Thirlway, and Parvin Ahmed.

### 2.12.5 RNAscope

mRNA was detected as red punctae in coronal FFPE mouse brain sections counterstained with haematoxylin using RNAscope® 2.5 VS target probes (Advanced Cellular Diagnostics; ACD) against each transcript (*Prnp*, Cat No. 476619; *Gfap*, Cat No. 313249; *C3*, Cat No. 417849) and an RNAscope® VS universal AP Reagent kit (ACD, Cat No. 323250). Probes targeting *Ppib* (ACD, Cat No. 313919) and *DapB* (ACD, Cat No. 312039) mRNA were used as positive and negative controls, respectively. Staining of tissue and RNA detection was automated using a Discovery Ultra IHC/ISH staining platform (Roche Diagnostics). Briefly, slides were deparaffinised and underwent treatment with target retrieval buffers and a protease solution to free RNA from protein complexes before being incubated with target probes and sequential rounds of the signal amplifying oligonucleotides (all reagents provided by Advanced Cell Diagnostics).

Whole Slide Images (WSIs) of each section at 40x magnification were obtained using a NanoZoomer S360 (Hamamatsu Photonics K.K.). The cortex, hippocampus, thalamus, cerebral nuclei, cerebellum, and brainstem were manually annotated in QuPath v0.3.2 (Bankhead et al., 2017). The positive pixel detection tool in QuPath was used to generate RNAscope® positive percentage values by region from which relative changes in transcript levels were inferred.

RNAscope and analysis were performed by Tom Murphy, Tamsin Nazari, and Emmanuelle Vire.

## 2.13 Single-cell transcriptomics of human prion disease

### 2.13.1 Human samples

We selected 10 individuals from archived tissue collected by the National Prion Clinic and stored in our Unit. Our selection criteria included the final diagnosis, which was sporadic CJD, availability of frozen frontal cortex samples, storage of samples in histopathology cassettes that enable us to identify the different anatomical regions of the frontal cortex more easily, and codon 129 methionine homozygous *Prnp* genotype. For our control group, we included frontal cortex samples from individuals with low-level AD pathology or pathological ageing provided by the Queen Square Brain Bank. These samples (N = 10) were matched for sex (male = 4 sCJD and 5 controls; female = 6 sCJD and 5 controls) but not age (mean age for sCJD = 70.4 years, SD = 8.6; mean age for controls = 83.7 years, SD = 8.6) More information regarding the clinicopathological variables of the selected patients can be found in Supplementary Table 1.

We were also able to source 3 non-dominant frontal lobe biopsy samples from sCJD patients. These extremely rare samples have been collected over 20 years by the National Prion Clinic because the differential diagnosis of CJD sometimes requires excluding neuroinflammatory conditions like primary cerebral vasculitis. Occasionally, this can only be determined through histological examination of brain tissue in life. These samples offer two main advantages: they are well preserved, and there is no post-mortem delay since tissue archiving is fast, usually less than 30 minutes after sample collection. The control samples for this group included frontal lobe biopsies from non-neurodegenerative disease controls with mixed clinical diagnoses and only non-specific minor histological changes (pathological non-diagnostic samples), provided by BRAIN UK. These samples (N = 3) were sampled similarly to the biopsies and individuals were matched for sex (male = 2 sCJD and 2 controls; female = 1 sCJD and 1 control) while the age was matched only partially (mean age for sCJD = 56.6, SD = 13.3; mean age for controls = 60.6, SD = 3.3). More information regarding the clinicopathological variables can be found in Supplementary Table 1.

### 2.13.2 DNA extraction from prion-infected frozen brain tissue

50-100 mg of brain tissue were transferred to a 2 mL screw-cap tube (Alpha Laboratories; CP5932). 450 µL ATL lysis buffer (QIAGEN; 939016) and 50 µL proteinase K 20 mg/mL (Invitrogen; AM2548) were added, and tubes were left in a Thermomixer Comfort heating block (Eppendorf) overnight at 50ºC with mixing at 800 rpm. The next day, 500 µL of TRIS-equilibrated phenol (Sigma-Aldrich; P4557) were added and mixed by inversion. The tubes were centrifuged at 16,000 g for 5 min at room temperature before transferring the upper aqueous phase to a fresh tube and discarding the lower organic phase. The addition of phenol, centrifugation and keeping of the aqueous phase was repeated. 500 µL of a 1:1 mix of TRIS-equilibrated phenol and chloroform mixture were added and mixed by inversion. After centrifugation, the aqueous phase was transferred to a fresh tube and 500 µL chloroform was added. After centrifugation, the aqueous phase was transferred to a fresh tube and removed from BSL-3 facilities to BSL-2 facilities. 500 µL of 100% cold ethanol was added to induce DNA precipitation. The supernatant was then aspirated and discarded without disturbing the DNA pellet, which was left to dry for a couple of minutes and was then resuspended in water.

### 2.13.3 *PRNP* codon 129 genotyping

TaqMan® SNP Genotyping Assays with appropriate probes were used for *PRNP* codon 129 genotyping according to the manufacturer's instructions. Briefly, 5 μL of TaqMan™ Genotyping Master Mix (Applied Biosystems; 4371353), 0.5 μL of assay probes (Thermo Fisher Scientific; 4351379; assay ID: C___2969398_10), 1 μL DNA and 3.5 μL water were added in each well of a MicroAmp™ Fast Optical 96-Well Reaction Plate (Applied Biosystems; 4346906). The plate was sealed, vortexed and spun down and then placed in a QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific) where the following PCR program was run: 1 cycle of 10 min at 95°C, 40 cycles of 15 s at 95°C and 1 min at 60°C. End-point fluorescence was detected, and the allelic discrimination plots were used to identify the sample genotype. MM, MV, VV, and negative controls were included in each run.

# 3 Experimental setup and pilot experiments

## 3.1 Introduction

### 3.1.1 Prion-propagating cell lines

Cell lines are invaluable experimental models to study prion propagation and biology as they allow experimentation under carefully controlled conditions, providing a cost-effective solution compared to animal studies, with the caveat of a less physiological system. Some of the earliest reports of attempts to propagate prions in culture are from Clarke and Haig. The authors established a cell line from a prion-infected mouse showing clinical signs after inoculation with the Chandler prion strain (M. C. Clarke & Haig, 1970). These cells were then passaged up to forty-one times to dilute the original inoculum and samples were titrated using mouse bioassays that showed high prion titres. This paper also mentions earlier work by Gustafson and Kanitz that observed irregular nuclei in cell cultures of prion-infected sheep and mouse brain preparations, published in Slow, Latent, and Temperate Virus Infections of the U.S. Department of Health, Education, and Welfare in 1965, however, the full text of the original manuscript is not available.

Fast forward 17 years of relative inactivity in the field and the Chesebro and Prusiner groups independently used mouse neuroblastoma cells to successfully propagate murine prions (D. A. Butler et al., 1988; R. E. Race et al., 1987, 1988). The researchers followed a different approach than previous studies and instead of establishing new cell lines from a prion-infected mouse brain, they infected existing cell lines by exposing them to infectious brain homogenate. They then had to clone individual cells and characterise the new subclones to ascertain the stability of infection. The one clone that could sustain prion infectivity was expanded and formed the stock that would extensively be used for prion research during the following decades (Solassol et al., 2003). These cells were named ScN2a (Scrapie N2a cells, where N2a is the Neuro 2A mouse neural crest-derived cell line).

The general approach of infecting cell lines with prions and subcloning to identify susceptible subpopulations can be used to test the propagation of different prion strains in different cell lines. For example, it has been shown that rat cells could be infected by mouse ME7 and 139A prions, and N2a cells overexpressing PrP with RML, 22L and 139A prions (Nishida et al., 2000; Rubenstein et al., 1992). Hamster cells have also been shown to be susceptible to prion infection, as well as cells derived from elk and deer (Bian et al., 2010; Raymond et al., 2006; Taraboulos et al., 1990). By expressing PrP from the appropriate species through transfection, a rabbit kidney epithelial cell line — RK13 — can support the propagation of sheep, elk, goat, mouse and bank vole prions (Bian et al., 2010; Courageot et al., 2008; Dassanayake et al., 2016; H.-J. Kim et al., 2012; Vilette et al., 2001). Unfortunately, similar approaches have failed to generate any cell lines that can propagate human prions (Krance et al., 2020). The few exceptions where human prion propagation has been successful concern models of terminally differentiated cells, which necessitate *de novo* infection for each new experiment, limiting the reproducibility of the system and its suitability for phenotypic drug screening (Groveman et al., 2019; Hannaoui et al., 2014; Krejciova et al., 2017).

In addition to the study of the species barriers and transmissibility of prion strains, prion-propagating cell lines have been extensively used to elucidate the molecular mechanisms and cellular events that are implicated in the formation of disease-associated PrP and disease progression, to study the pathophysiology of prion disease and to facilitate the discovery of novel therapeutics. Cultured cell lines are an especially valuable tool in the quest for discovering novel anti-prion compounds as they are easy to manipulate and cost-efficient compared to in vivo models, they recapitulate the key molecular events in prion disease, are amenable to high-throughput screening and can be used to design reproducible experiments that bypass ethical concerns associated with the use of animals and human tissue (Krance et al., 2020). While phenotypic screening in mouse cells has identified a multitude of small anti-prion molecules that are effective against mouse prions in vivo, further studies showed unsatisfactory results when those were tested in humanised mice infected with human prions (Berry et al., 2013; Giles et al., 2015, 2016; Kawasaki et al., 2007). These studies also indicate that the emergence of drug-resistant prions further complicates human prion disease therapeutics.

Our study used N2aPK1 cells — also referred to as PK1 cells from now on — a highly prion-sensitive subclone of N2a cells that were derived during research done by Peter Klöhn et al. (Klöhn et al., 2003). These cells were a product of three rounds of subcloning and susceptibility screening (N2a > N2a/Gary > N2aPD88 > N2aPK1), while the scientists demonstrated a more than x1000 increase in prion sensitivity compared to the original N2a cells used. PK1 cells are also the cell line used for the Scrapie Cell Assay, an in vitro cell-based prion infectivity assay that will be discussed in more detail in section 4.1.2.

### 3.1.2 Chapter summary

We selected two high-throughput single-nucleus RNA-seq protocols that can be used with frozen tissue samples, namely DroNc-seq and SPLiT-seq, and followed the authors' protocols to set up the equipment in our Biosafety Level 2 laboratories. To validate the functionality of our setup, we performed the recommended species-mixing experiments. These experiments should be performed with every new setup and aim to assess the correct operation of both the equipment and the protocols used.

The core principle of massive parallel single-cell and single-nucleus protocols is that they can provide single-cell resolution by introducing unique barcodes to the transcriptomes of each cell or nucleus. However, a small probability remains that two or more nuclei will have the same barcode, either due to the stochastic nature of the technique or due to protocol execution mistakes. This probability is usually referred to as the "doublet rate" for DroNc-seq or the "barcode collision rate" for SPLiT-seq. The species-mixing experiments provide a framework to test this rate for both protocols by sequencing a mix of human and mouse cells at the same time. Both species have excellent transcriptomic annotation that allows demultiplexing the data and identifying the number of transcripts that originated from a mouse or a human cell for every unique nucleus barcode. While the terms "nucleus" and "nucleus barcode" are usually used interchangeably, it is important to underline here that a unique nucleus barcode might not necessarily be associated with transcripts from only one unique nucleus, due to doublets/collisions and possible incorporation of ambient RNA.

## 3.2 Results

### 3.2.1 Experimental setup validation and species-mixing experiments

#### *3.2.1.1 Using DroNc-seq*

Human HEK293T and murine PK1 cells were cultured, and their nuclei were extracted and counted. A 50/50 mix of human/murine nuclei suspension was prepared and loaded to the DroNc-seq system as described in the methodology. The emulsion was collected for 22 minutes. During this experiment, the microfluidic device clogged once and had to be replaced. Previous tests also led to device clogging, highlighting a potential drawback of the method.  After droplet lysis, the beads were washed and counted, and 110,000 beads were finally recovered. Of these, 20,000 beads were used for library preparation and sequencing. This library was multiplexed with 2 more DroNc-seq libraries, all used in equal amounts.

Sequencing generated approximately 46,5 million reads associated with this experiment, as expected, which were processed with the Drop-seq pipeline and aligned to a combined human/mouse annotated transcriptome. Our experiments identified only 102 nuclei barcodes that are associated with more than 4000 reads each. While the threshold values are chosen arbitrarily, they are usually more stringent in the official protocol than our analysis, demonstrating the low output of this specific experiment. Indeed, while 20,000 beads were used for library preparation, only 102 barcodes were kept after filtering, while further relaxation of the filtering criteria introduced an unacceptably high number of barcodes containing little information. In our experiment, only a small number of cell barcodes was associated with a relatively high fraction of reads, while most of the cell barcodes only contained a very small fraction of the reads (Figure 3.1).

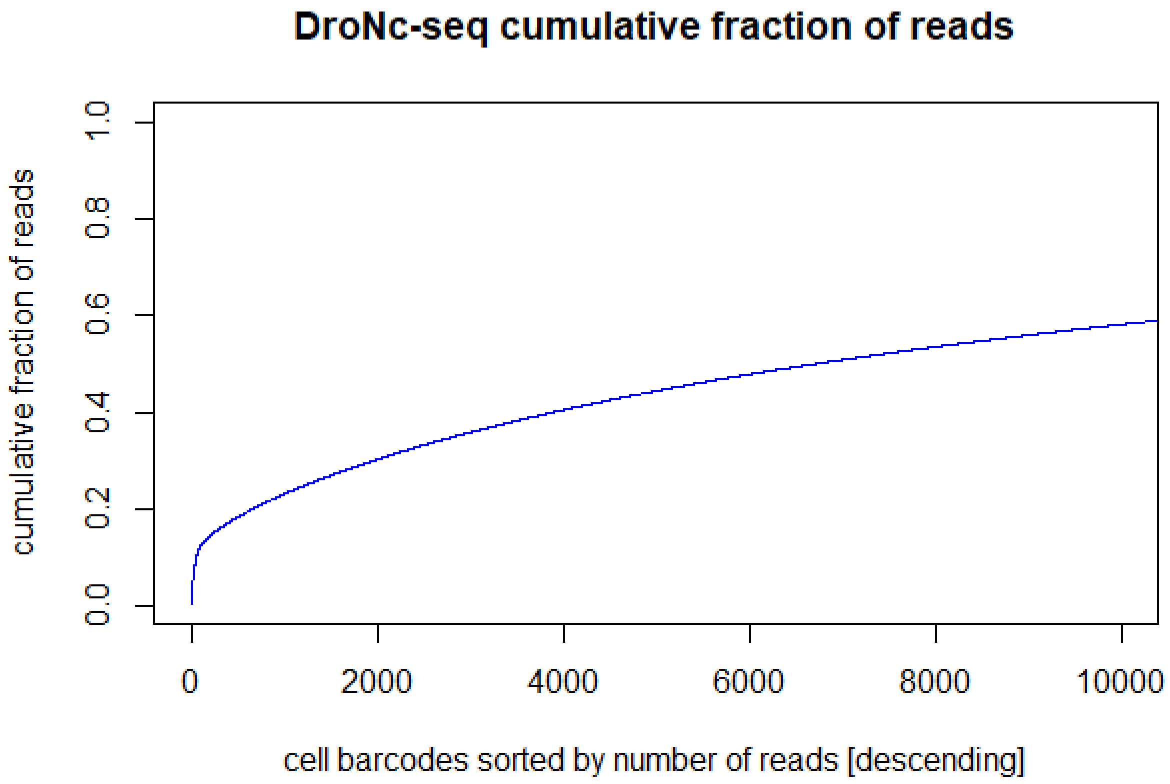


***Figure 3.1: DroNc-seq cell barcodes were characterised by a gradually increasing cumulative fraction of reads.*** *The figure shows the top 10,000 barcode sequences containing the most information versus the cumulative fraction of reads they are associated with. For most of the barcodes, the curve has a very gradual slope, indicating that each barcode was only associated with a very low number of reads. There are a few barcodes that are associated with more reads and contain the most information. These are characterised by a steeper increase of the curve.*

Recognising a substantial loss of information, we evaluated the efficiency of the technique by calculating the number of genes per nucleus identified for human and mouse cells using only basic filtering. This number was very low for DroNc-seq, where the median number of genes identified was 14 for human and 27 for mouse cells.

After demultiplexing, reads were assigned to unique barcodes of origin, barcodes with fewer than 4000 reads were filtered out and the remainders' identity was calculated using the transcriptomic annotations. Plotting the number of reads and *in silico* calculated species of origin demonstrates our setup's ability to provide single-nucleus resolution data (Figure 3.2). Our approach identified 27 human nuclei (26.5%), 70 murine nuclei (68.6%), and 5 nuclei of mixed origin (4.9%). These doublet rates are in agreement with the method's expected doublet rate, which has been calculated to be approximately 5% (Habib et al., 2017). In addition, a closer examination of Figure 3.2 suggests that this ambiguity arises mostly for nuclei with a very low number of reads, suggesting that these barcodes might correspond to empty droplets carrying high amounts of ambient RNA from both species. In contrast, 2 barcodes with a high number of reads could correspond to droplets containing both human and mouse cells. Finally, human or mouse barcodes with a very high number of reads might also correspond to droplets containing more than one nucleus of the same species.

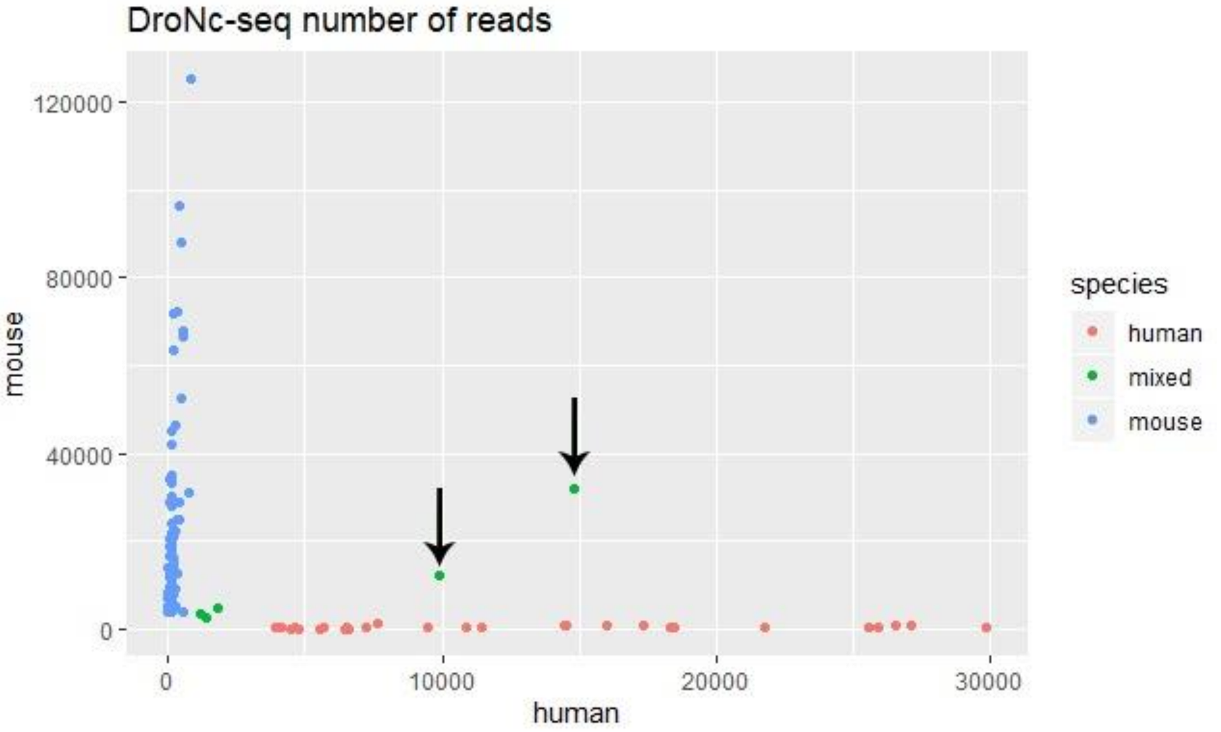


***Figure 3.2: DroNc-seq species-mixing experiment discriminates between human and mouse cells.*** *Each point represents a unique cell barcode. Most of the barcodes are only associated with one species, either mouse (blue) or human (red), while some are associated with reads mapping to both mouse and human transcriptomes (green). Most of these mixed barcodes have a small number of reads and might correspond to empty droplets containing ambient RNA, while two of them (indicated by arrows) have a high number of both mouse and human transcripts and could correspond to co-encapsulations of both human and mouse nuclei in a single droplet. Barcodes with a high number of reads from the same species might also correspond to multiplets where the co-encapsulated nuclei originated from the same species.*

#### *3.2.1.2 Using SPLiT-seq*

Human HEK293T and murine PK1 cells were cultured, and their nuclei extracted and counted. A 50/50 mix of human/murine nuclei suspension was prepared and diluted to the starting concentration of SPLiT-seq samples. A total of 40,000 nuclei were used as input to the reverse transcription round (10 wells, 4,000 nuclei each). The library generated was mixed with two more libraries in equal amounts (approximately 2000 nuclei each) and the resulting multiplexed library was sequenced.

Sequencing generated approximately 48.5 million reads associated with this experiment. These were pre-processed using the SPLiT-seq-pipeline. After barcode demultiplexing and filtering, we recovered 1494 cell barcodes passing the quality control thresholds. Their reads were then aligned to a combination of both human and mouse annotated transcriptomes and their species of origin were calculated (Figure 3.3). We recovered 453 human cells (23.7%), 1077 murine cells (72.1%) and 63 cells of mixed origin (4.2%). Most of the barcodes associated with both human and mouse transcripts were found to have a small number of Unique Molecular Identifiers (UMIs), while some of the barcodes associated with a single species and having very high UMI counts might also represent barcode collisions of nuclei from the same species. The barcode collision rate was in accordance with statistics calculated by the authors and our calculations were based on sample load (Rosenberg et al., 2018). This data suggests that SPLiT-seq is capable of discriminating between the human and mouse cells, and thus, can provide single-nucleus resolution data.

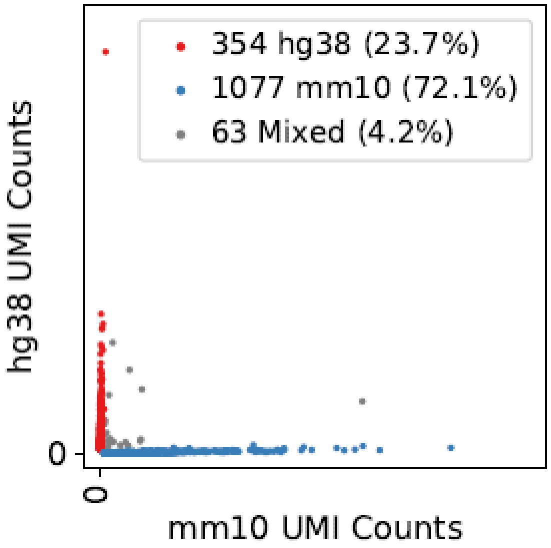


***Figure 3.3: SPLiT-seq species-mixing experiment discriminates between human and mouse cells.*** *Each point represents a unique barcode. Axes represent numbers of Unique Molecular Identifier (UMI) counts. Most of the human (red) and mouse (blue) barcodes are specific to only one species, while a few barcodes are associated with transcripts mapping to both human and mouse transcriptomes (grey). Most of the ambiguous barcodes have a low UMI count. Some of the barcodes associated with a single species but have a very high UMI count might be caused by a barcode collision of nuclei from the same species.*

We then evaluated the efficiency of the method by calculating the median number of genes identified in human and mouse cells, using only basic filtering. We calculated a median of 145 and 196 identified genes per nucleus for human and mouse cells, respectively. While these numbers are still lower than the statistics published by the authors of the technique where 677 genes per nucleus are mentioned, this can be due to differences in the sequencing depth. In our case, nuclei were sequenced at a depth resulting in approximately 250 UMIs per nucleus, while the original manuscript had a much higher sequencing depth resulting in approximately 1000 UMIs per nucleus. More importantly, the comparison between the efficiencies of the two protocols and practical considerations led us to the decision of using SPLiT-seq for the following experiments (see discussion, section 3.3.1).

### 3.2.2 Correlation of SPLiT-seq and bulk RNA sequencing data

Bulk RNA sequencing has long been the gold standard method of transcriptomics. Newer single-cell methods are expected to uncover hidden cell-type-specific expression patterns, but their novelty comes with the cost of more limited method validation. To assess the concordance between our experimental methodology and a more traditional approach, we compared our single-cell data with data generated using an extensively validated bulk RNA sequencing protocol.

We generated bulk and single-nucleus RNA-seq data from the same two cell lines, PK1 and iPK1 cells, using Illumina's whole-transcriptome TrueSeq Stranded Total RNA solution and SPLiT-seq. We then converted our single-nucleus data to pseudo-bulk by summing the expression of each gene across all cells. Finally, we integrated the pseudo-bulk and bulk datasets and calculated the concordance of expression of the top 2000 differentially expressed genes between the two cell lines (Figure 3.4). Our results suggest high concordance between single-nucleus data generated by SPLiT-seq and bulk RNA-seq data (Spearman correlation coefficients for PK1 and iPK1 cells were 0.728 and 0.766, respectively). Overall, our data recapitulate the findings of previous studies (Collin et al., 2019; Macosko et al., 2015) and provide additional evidence that validates our methodology.

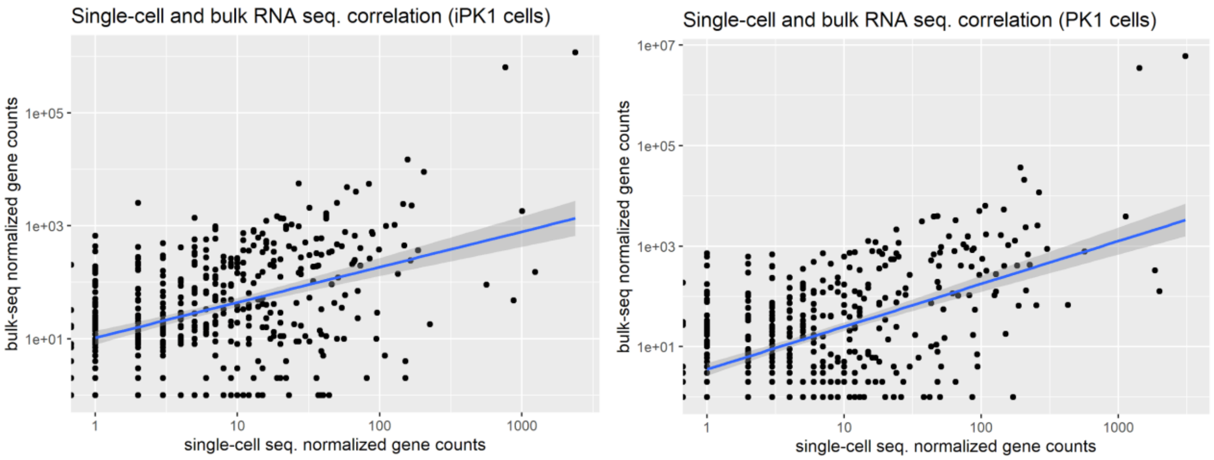


***Figure 3.4: SPLiT-seq data show a high correlation with bulk RNA-seq data**. PK1 and iPK1 cell lines were sequenced using bulk RNA-seq and single-nucleus SPLiT-seq. The single-nucleus data was converted to pseudo-bulk by calculating the sum of expression of each gene across all cells. The two datasets were integrated and the normalized number of counts from bulk sequencing (y-axis) and single-cell sequencing (x-axis) was plotted. The Spearman correlation coefficients for iPK1 (left) and PK1 (right) cells were 0.766 and 0.728, respectively. A linear model was fitted to visualise the relationship between the two datasets (blue line). The dark grey area around the line corresponds to a 0.95 confidence interval.*

### 3.2.3 PK1 and iPK1 cell lines transcriptomics

Before moving on to complex brain tissue, we wanted to evaluate and validate our methodology using a prion-susceptible neuroblastoma cell line. The PK1 cell line is a result of serial sub-cloning of mouse neuroblastoma N2a cells and can propagate RML prions in vitro when inoculated with RML-infected mouse brain homogenate. The infected cells can chronically sustain prion infection and are referred to as iPK1 (chronically infected-PK1 cells).

We prepared two nuclei suspensions from PK1 and iPK1 cells and processed them using SPLiT-seq to generate two multiplexed libraries. We used the same plate for barcoding both cell lines to decrease the impact of possible batch effects while using different reverse transcription barcodes for each cell line, to enable their identification during the in-silico analysis. After sequencing and data pre-processing, we identified 2188 PK1 and 1662 iPK1 nuclei barcodes with a median of 160 and 232 genes per nucleus, respectively. We performed QC filtering to remove cells with less than 100 genes or UMIs and recovered 1450 PK1 and 1339 iPK1 high-quality cells with a median of 247.5 and 303 genes per nucleus, respectively (Figure 3.5).

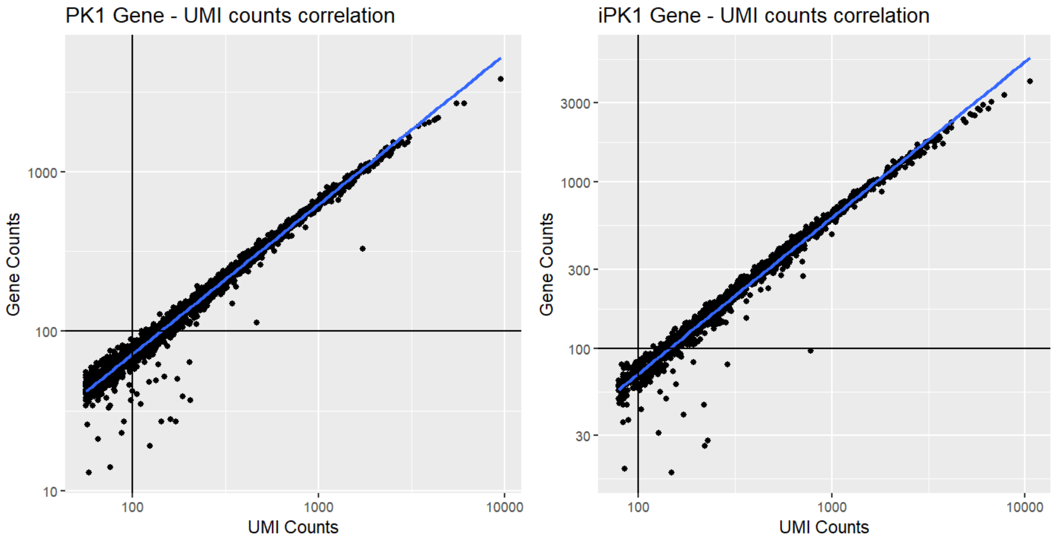


***Figure 3.5: Filtering of low-quality cells.*** *Cells with less than 100 UMI or gene counts were filtered out. Horizontal and vertical black lines define the QC thresholds. The number of identified genes increases with a higher number of UMIs, as expected.*

We then proceeded to reduce the dimensions of the dataset using PCA, and then visualise it by plotting the first 2 PCs (Figure 3.6), and via t-SNE plots with a range of perplexity values calculated on the first 50 PCs (Figure 3.7). Both plots suggest that the data is very homogenous. Even though there is some separation between the two populations, it would be impossible to separate them without the a priori knowledge of their barcodes that was used to overlay them with different colours. Importantly, the first and second principal components can only explain 3 and 2 per cent of the data variability, a very low number in comparison to scRNA-seq of complex tissues with multiple cell types. Overall, this homogeneity is expected from a cell line and highlights the fact that more sensitive techniques such as Smart-seq2 (Picelli et al., 2014) may need to be used with such populations to be able to identify minute differences in gene expression levels.

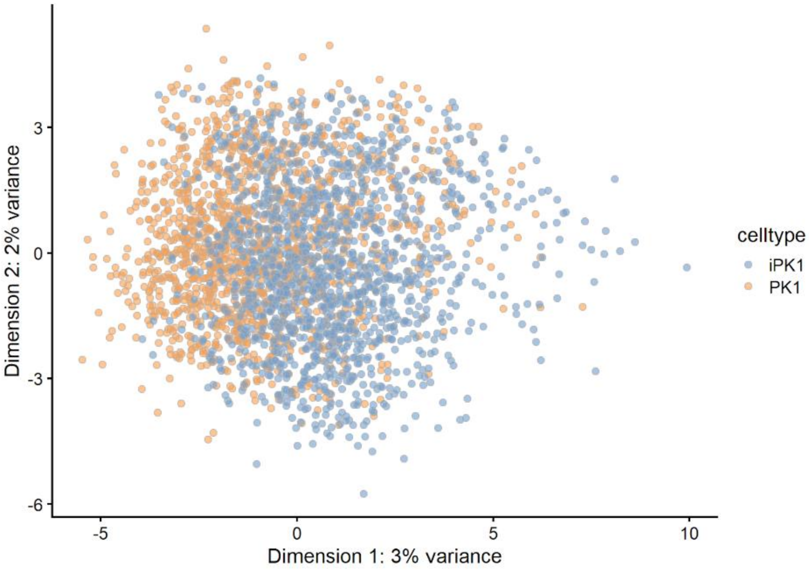


***Figure 3.6: Plot of the first 2 Principal Components.*** *Cells are coloured based on their cell line of origin. Both components explain a low percentage of the variance and there is a substantial overlap between the two cell populations, which suggests low heterogeneity of the data.*

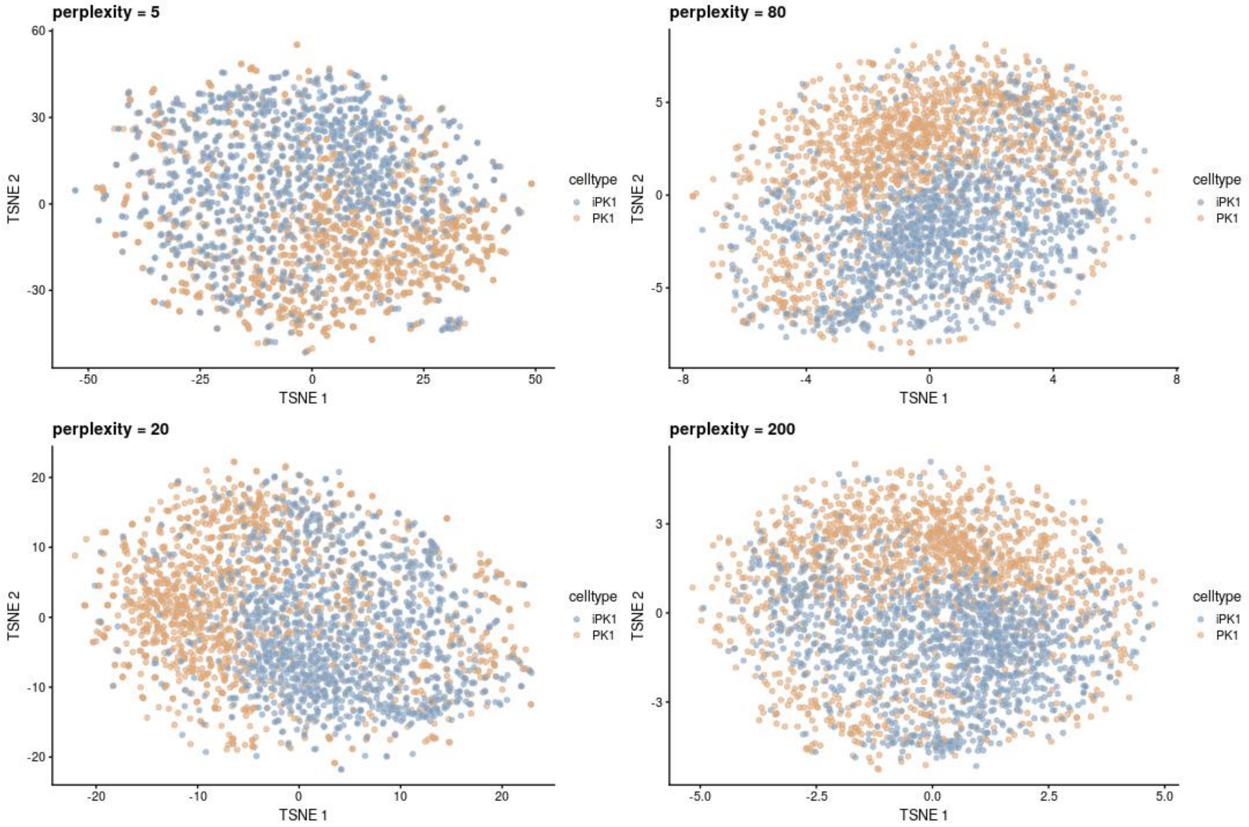


***Figure 3.7: t-SNE plots of PK1 and iPK1 cells using a range of different perplexity values.*** *The visualisation suggests that the plot is robust to the choice of perplexity value. Even though the two cell lines seem to separate to an extent, there is substantial overlap, making a clear separation of the two populations impossible.*

Recognising the homogeneity of the dataset, we proceed with testing different clustering algorithms. We used a graph-based clustering approach and evaluated the two most commonly used algorithms: the number weighting scheme with the walktrap community detection algorithm (recommended default of the Scran package) and the Jaccard weighting scheme with the Louvain community detection algorithm (recommended default of the Seurat package). We tried three different values for the number of k neighbours, 10, 20 and 30. We evaluated the cluster separation plots (data not shown) to select the one that showed the best performance. While due to the homogeneity of the data none of the approaches led to a good cluster separation, as expected, we selected the number/walktrap algorithm with 20 neighbours that performed best and overlaid the cluster information on the t-SNE plot to visualise cluster relationships (Figure 3.8). By comparing the two plots of cluster information and cell identity, we notice that clusters 2 and 4 mostly comprise infected cells, while cluster 3 comprises non-infected cells. Cluster 1 includes both infected and non-infected cells.

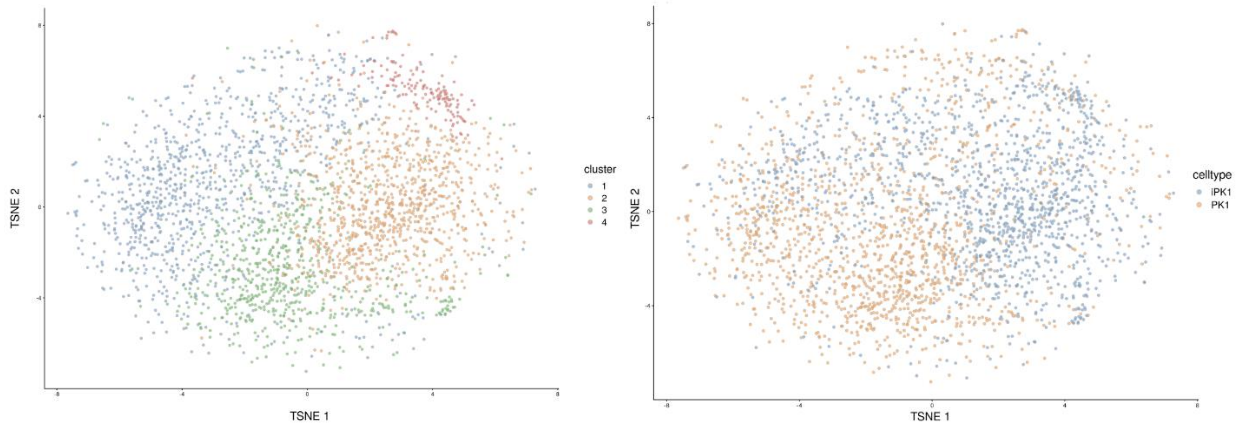


***Figure 3.8: t-SNE visualisation of PK1 and iPK1 cells overlaid with cluster information.*** *The algorithm selected (number weighting scheme, walktrap community detection, 20 neighbours) identified 4 clusters of cells (left). Overlaying the same plot with the a priori information about cell line identity (right) allows us to estimate the cell types most associated with each cluster. Clusters 2 and 4 comprise mostly infected cells, while cluster 3 of non-infected cells. Cluster 1 includes both infected and non-infected cells.*

To identify functional differences between the clusters, we extracted the upregulated marker genes, i.e. genes that show differential expression between clusters, drive cluster separation and are characteristic for each cluster. A one-sided pairwise t-test was used to compare gene expression between each pair of clusters. Only upregulated genes that were differentially expressed with a log-fold change of more than 1 between the current group and any other group were included in the final list. Genes were then ranked according to their p-values for each cluster separately and the final list contained the top 5 genes (ranked by significance) from each pairwise comparison. This approach identified 11 marker genes for cluster 1 (*Cdk8, Gm42418, Lars2, Malat1, Gm26917, Comt, Xist, Pde1c, Gm20388, Gm48641, Cep112*), 10 marker genes for cluster 2 (*Cmss1, Gm15564, Cdk8, Gm42418, Lars2, Malat1, Gm26917, Map1b, Xist, Comt*), 7 for cluster 3 (*Malat1, Gm26917, Unc5c, Xist, Map1b, Gm48641, Comt*), and 7 for cluster 4 (*Gm42418, mt-Rnr1, Lars2, mt-Rnr2, Cmss1, Cd44, Gm15564*). We attribute the substantial overlap between the markers of all clusters to the homogeneity of the dataset; because no major differences between gene expression of each cluster exist, the algorithms include genes with even small fluctuations in their expression. We focused on clusters 2 and 4, which are of particular interest because they comprise mostly infected cells. Interestingly, cluster 2 shows increased expression of a set of genes (*Cmss1, Gm15564, Gm42418 and Lars2*) in comparison to non-infected cell clusters 1 and 3, while the same set of genes is found to be even more upregulated in cluster 4 (Figure 3.9). *Gm15564* and *Gm42418* are predicted long non-coding RNAs (lncRNAs), with no known function. *Cmss1* encodes the Cms1 ribosomal small subunit homolog and *Lars2* an Aminoacyl-tRNA synthetase. Cluster 4 is also characterised by upregulation of *mt-Rnr1* and *mt-Rnr2*, the mitochondrially encoded 12S and 16S rRNAs. These transcripts are not normally found in the nucleus, suggesting that some mitochondria might have remained in our nuclei preparation.

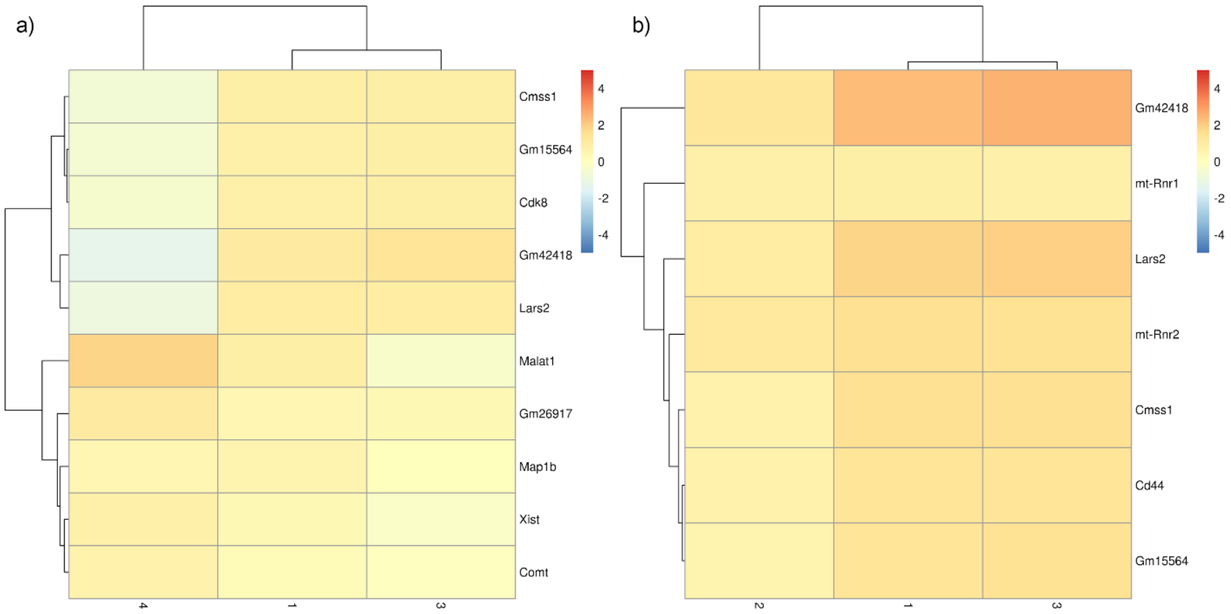


***Figure 3.9: Expression of marker genes of clusters 2 (a) and 4 (b) compared to all other clusters.*** *Clusters 2 and 4 comprise mostly infected cells. Cluster 2 shows increased expression of Cmss1, Gm15564, Cdk8, Gm42418 and Lars2 compared to clusters 1 and 3 which mostly comprise non-infected cells. In addition, cluster 4 shows an even higher expression of these transcripts.*

Due to the identification of many upregulated lncRNAs and mitochondrial transcripts, we decided to quantify identified transcript biotypes of the dataset. Our analysis suggested that these transcripts only account for a small percentage of the total number of transcripts identified (Figure 3.10). The most abundant transcripts were protein-coding genes, as expected, while mitochondrial transcripts account for less than 1% of the data. This evidence suggests that due to the homogeneity of the data, the introduction of even a small number of transcripts that deviate from this uniformity can skew the differential expression results.

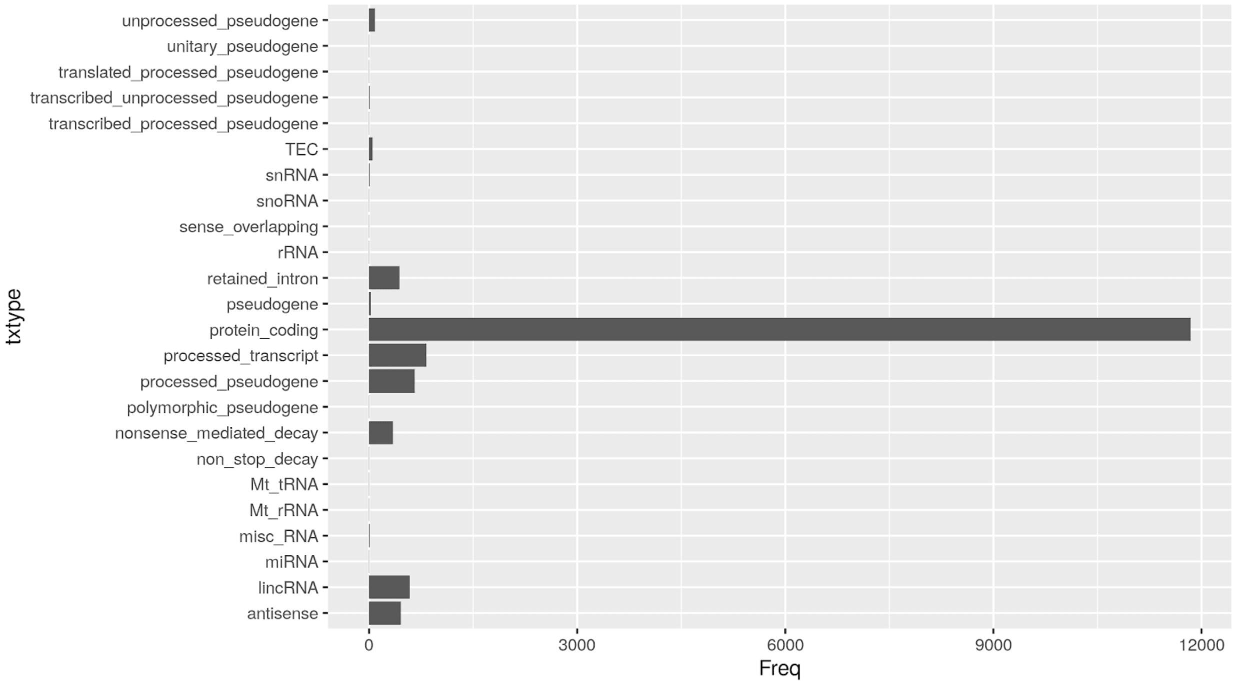


***Figure 3.10: Most of the transcripts identified are from protein-coding genes.*** *The frequency of each transcript type was quantified, highlighting that protein-coding genes account for most of the data, while other transcripts are found in small percentages. Mitochondrial transcripts account for less than 1% of the data and their bar is not visible in this graph.*

Overall, our exploratory analysis suggests that no conclusions can be drawn concerning differences in the transcriptomic profiles of these two cell lines when using single-nucleus approaches. The homogeneity that characterises each cell line does not allow for meaningful data clustering or marker gene detection. In addition, very subtle transcriptional differences, if they exist, will inadvertently be lost when using single-cell approaches, as their sensitivity is much lower than bulk RNA sequencing methods.

### 3.2.4 SPLiT-seq validation on frozen mouse brain

Single-cell approaches are inherently most suitable for profiling heterogeneous populations, where transcriptional differences are substantial and can be identified easily, even with less sensitive methods. We decided to validate our protocol using frozen mouse brain tissue, a sample that will be used for future experiments and closely resembles frozen human brain tissue, which will also be sequenced.

We prepared a nuclei suspension from a healthy mouse frontal lobe and barcoded it using SPLiT-seq to prepare a single library, which was subsequently sequenced. We identified 13,635 cells, having a median of 603 genes per cell, much higher than in previous experiments. More importantly, this high-quality data meant that our QC filtering only removed 1 cell which had less than 100 genes identified. The data was normalised and log-transformed. The dimensions of the data were then reduced by PCA and the first 8 PCs were kept, as they explained most of the variability of the data. We visualised the data by plotting the first 2 PCs (Figure 3.11) highlighting interesting data variability, as expected.

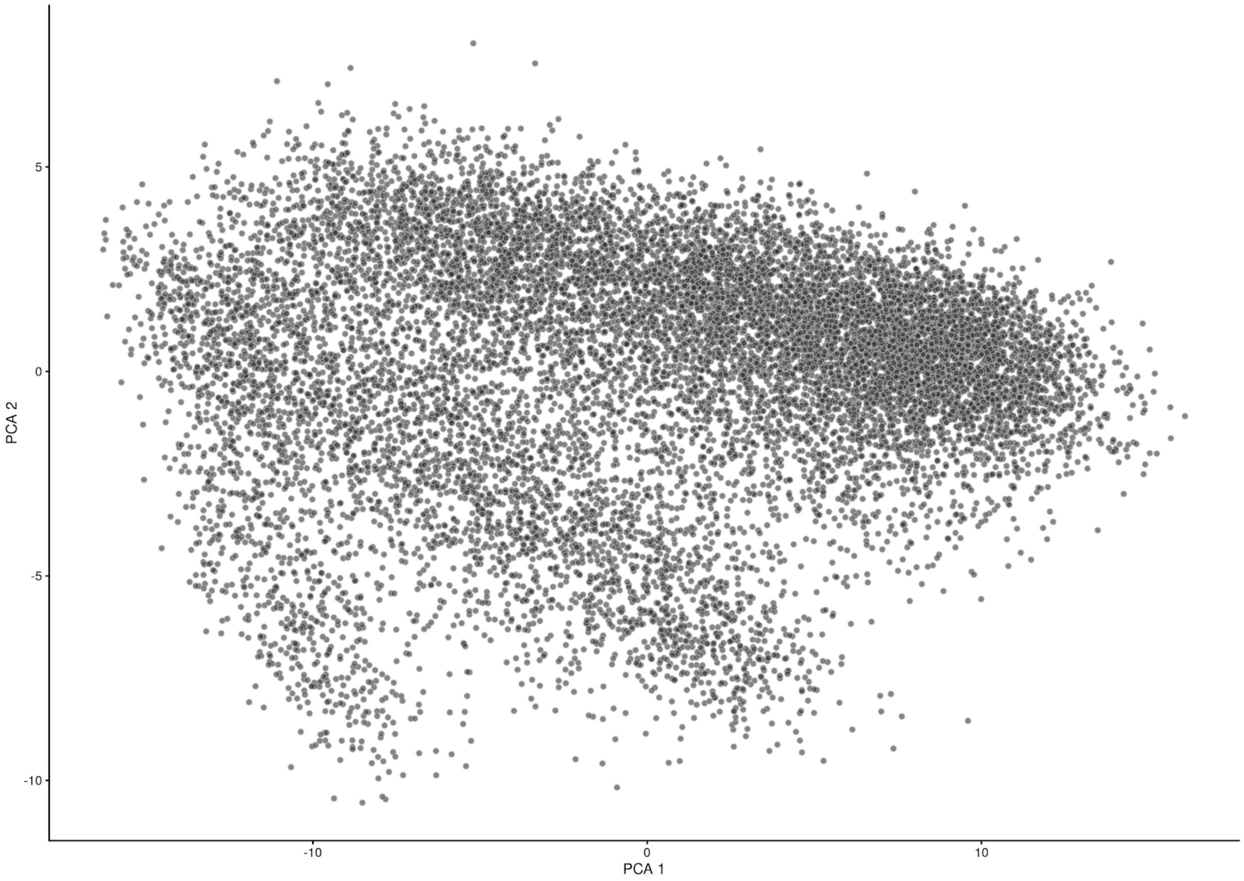


***Figure 3.11: A PCA plot suggests data variability in both the first and second principal components and separates the cells into three large clusters.***

We then clustered the data using graph-based clustering and the Jaccard weighting scheme and Louvain community detection algorithms and overlaid the cluster information on top of a t-SNE plot (Figure 3.12). We identified 17 cell clusters, which were also clearly separated in the visualisation.

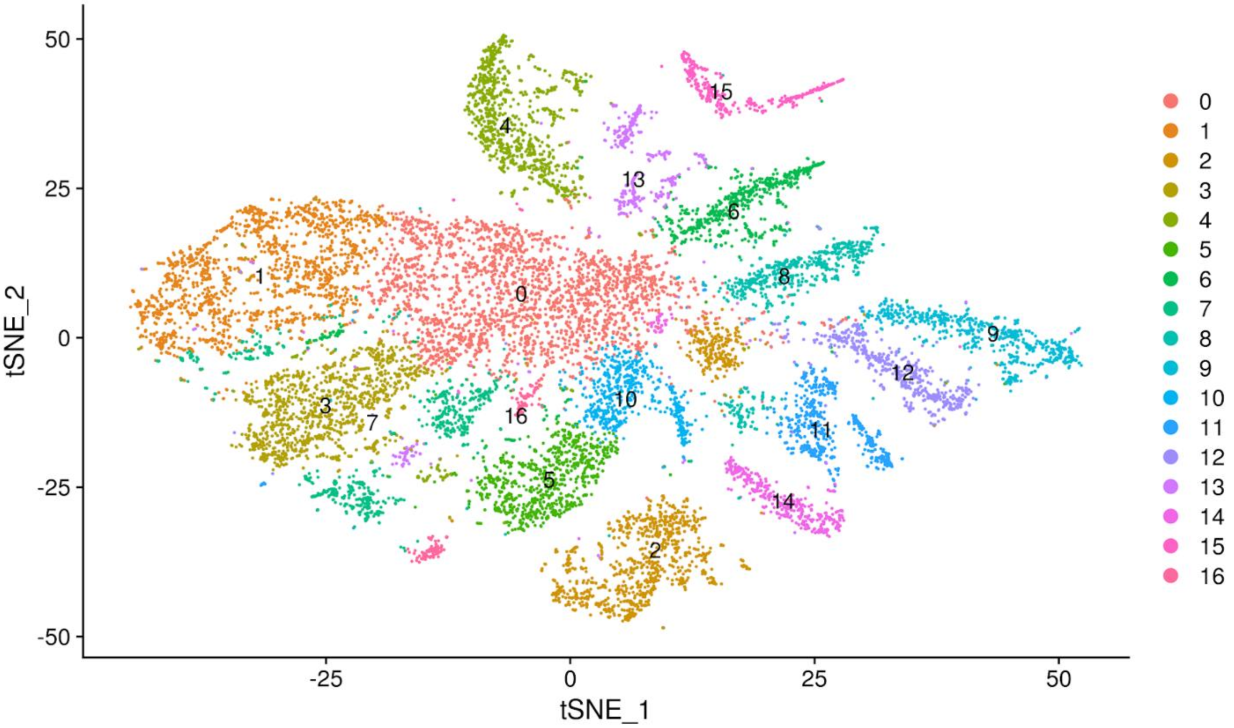


***Figure 3.12: Cluster information is overlaid on a t-SNE plot of the dataset, showing a clear separation of the 17 clusters identified (0 to 16).***

To functionally characterise the clusters, we identified their marker genes. Only upregulated genes expressed in at least 25% of the cells of each cluster and showing differential expression of more than 0.25 log-fold were considered. To examine their specificity, we drew a heatmap of the top 5 marker genes of each cluster and their expression in all clusters (Figure 3.13). It is evident that these marker genes, whose names are shown on the left of the heatmap, are very specific to each cluster, allowing us to confidently use them for cluster annotation.

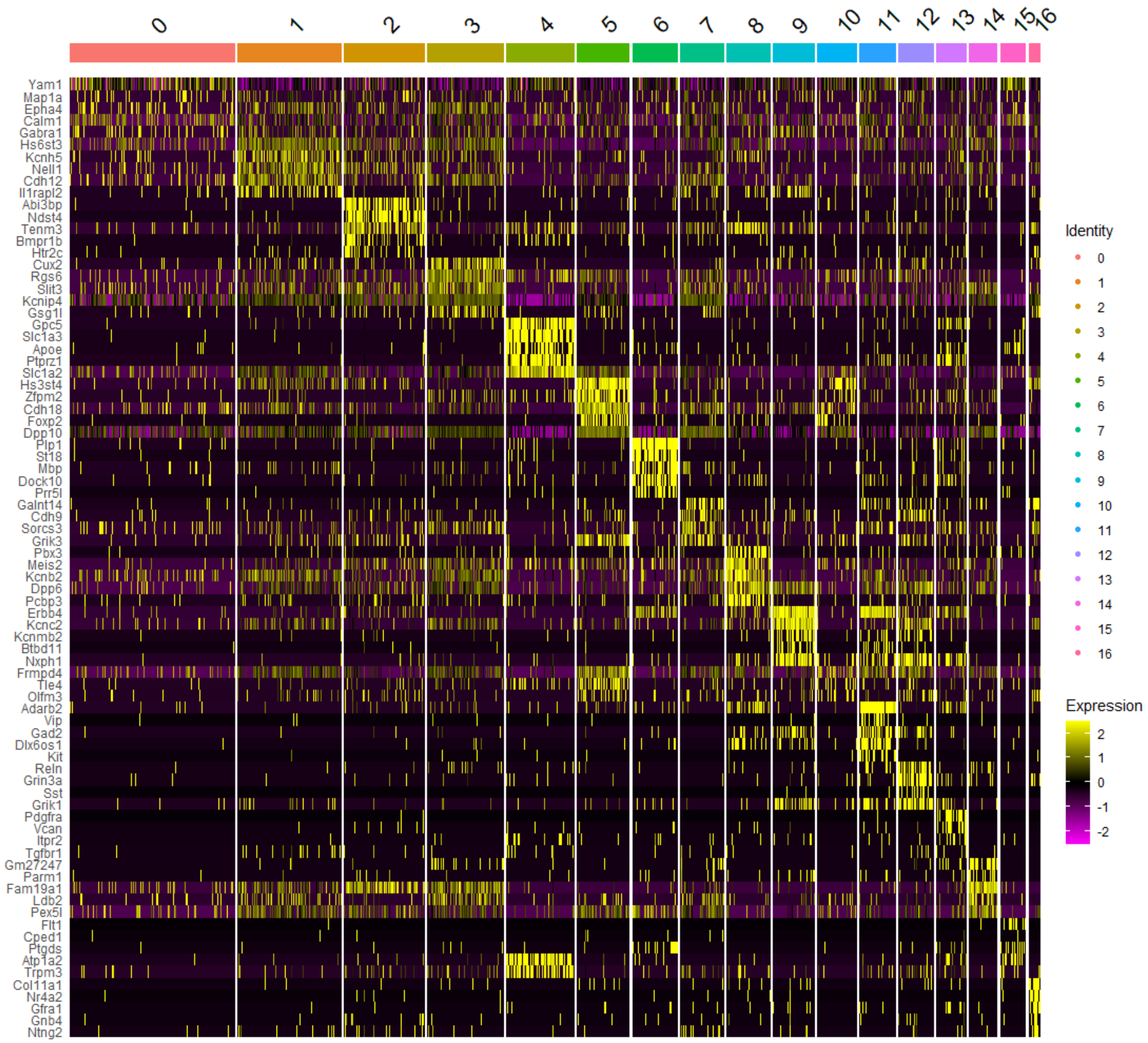


***Figure 3.13: Heatmap of the top 5 more differentially expressed gene markers of each cluster validates their specificity.*** *Each cluster is uniquely characterised by the set of the marker genes identified.*

We used the marker gene list generated previously to identify cell types using the scCATCH package. Automatic annotation identified 7 cell clusters: type IC spiral ganglion neuron, neuron, quiescent neural stem cell, oligodendrocyte, type II spiral ganglion neuron, oligodendrocyte precursor cell, and endothelial cell (Figure 3.14). Of these, the broad label ‘neurons’ spans across multiple different clusters, most probably corresponding to a multitude of neuronal subtypes. Even though automatic annotation is time-efficient and can provide an overview of the identified cell populations, manual annotation will be required to increase the resolution of cell populations identified and correctly characterise their subtypes. Overall, our data is in concordance with single-nucleus studies reviewed in the introduction and suggest that our protocol functions correctly when used with frozen mouse brain tissue.

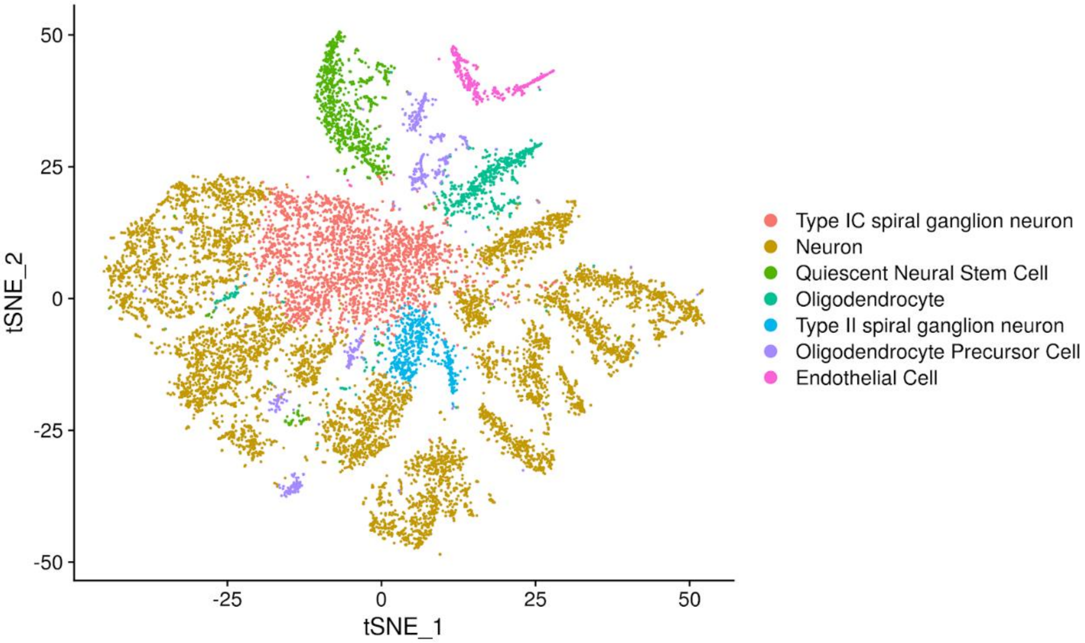


***Figure 3.14: Automatic annotation using scCATCH identified 7 cell populations.*** *We used the automatic annotation tool scCATCH, which compares the marker genes of each cluster with curated databases of cell populations and their markers. While the method does not have enough resolution to identify neuronal subpopulations, which are separate in the t-SNE plot, it allows for quick annotation before manual curation of the data.*

## 3.3 Discussion

### 3.3.1 Comparison between DroNc-seq and SPLiT-seq

Single-cell technologies have revolutionised the field of transcriptomics and in a short time have become the tools of choice for many cutting-edge studies. Some of the early research outputs of a very immature field have already led to novel insights, while the promise of increased resolution has attracted the necessary attention leading to a boom of published novel methods. While all newer approaches claim substantial improvements over older techniques, the actual user experience can differ. In addition, the suitability of each methodology will differ based on the research context and should be critically evaluated considering the nature of the input material, cost, desired output, and possible constraints relevant to the research environment. Here we aimed to select methodologies that can be used to profile prion-infected human brain samples in single-cell resolution and evaluate their practicality and performance in our specific context. Our main consideration was the nature of our samples, infected with prions which are lethal human pathogens and being stored frozen, and led us to select two single-nucleus approaches that are compatible with frozen brain tissue and can be implemented safely at a Biosafety Level (BSL) 3 laboratory, namely DroNc-seq and SPLiT-seq.

DroNc-seq is a droplet-based high-throughput snRNA-seq protocol that emerged as a complementary approach to Drop-seq for processing frozen brain tissue. At the core of the method is a microfluidic device that allows the encapsulation of nuclei and barcoded beads in nanolitre droplets. Due to the schematics and protocols being open source, we were able to obtain all required equipment, including prefabricated microfluidic devices and barcoded beads. We followed the official instructions to build the system in our BSL 2 laboratories for testing purposes. Our preliminary tests included a species-mixing experiment aimed to validate the correct operation of our hardware, its doublet rate and throughput, and the bioinformatics pipeline. While we were able to generate single-nucleus resolution data that could successfully discriminate between mouse and human cells, our overall data throughput was not satisfactory compared to other published studies reviewed earlier (Habib et al., 2017; Mathys et al., 2019). We speculate that further protocol optimisations would be required to maximise data output. In addition, we noticed a variable performance of the technique, even under the same experimental conditions and high sensitivity to minute details, indicating lower robustness than claimed by the authors. For example, our microfluidic devices would exhibit frequent clogging due to the high rate of bead flow, which would necessitate their replacement during the experiment. Avoidance of dust was also found to be detrimental to the correct operation of the devices, as well as the gentle handling of carrier oil. We also noticed that our number of beads recovered would also be affected by the exact equipment used, for example, the use of specific centrifuges would usually lead to lower percentages of loss. Overall, we would be reluctant to use this method with precious human brain biopsy samples, where otherwise trivial human error could potentially mean the loss of invaluable material. In addition, some safety concerns were also raised during testing, involving the use of needles to affix tubing onto the syringes and the high probability of spillages.

The alternative technique tested, SPLiT-seq, is based on the principles of combinatorial indexing and can be used, with adaptations, to profile both single cells and nuclei. This method did not require any special equipment and had a more gradual learning curve, as it only required basic liquid handling and common molecular biology techniques, such as PCRs and ligations. This reduced complexity also makes the protocols easier to troubleshoot and adapt to be used in a BSL-3 environment. Our preliminary species-mixing experiments validated the single-cell resolution of the protocol and its low barcode collision rate. Further optimisations led to the definitive version of the protocol, as described in the methods. Overall, we found the method to be robust and performing as expected with relative ease, while our data output started as low, but increased to levels comparable to the original manuscripts during our last experiment involving mouse brain tissue (Rosenberg et al., 2018).

Based on our first observations, we have ultimately decided to proceed with our study using the SPLiT-seq protocol, due to its higher reliability, ease of use and safety. We speculate that the protocol can be adjusted to conform to BSL-3 laboratory requirements with relative ease, allowing us to process infected human brain tissue.

### 3.3.2 Transcriptomic alterations of prion infection in PK1 cells

While our ultimate aims are to profile prion-infected mouse and human brain tissues, we decided to assess our protocols using a mouse cell line that can propagate prions *in vitro*. PK1 cells can be chronically infected with RML prions after inoculation with infected mouse brain homogenate. While heterogeneity has already been described in this particular cell line (Marbiah et al., 2014), we aimed to identify potential cell subpopulations using a single-nucleus sequencing approach.

Our experiment focused on comparing the transcriptomic profiles of chronically infected and uninfected cell lines using SPLiT-seq. The generated dataset had a low median number of genes identified resulting in lower-than-usual sensitivity. While we were not able to discriminate between infected and uninfected cells without having a priori information, our analysis identified clusters of cells exhibiting transcriptomic differences.

By comparing clusters consisting mainly of infected cells to ones consisting mainly of uninfected cells we identified increased expression of a set of genes: *Cmss1, Gm15564, Gm42418* and *Lars2*. Two of them encode lncRNAs with no known function (*Gm15564* and *Gm42418*), while the other two are related to protein translation (*Cmss1*, *Lars2*). This suggests that either an increase in translation might be associated with cell infection, or our clustering algorithm separated these clusters as more transcriptionally active. Overall, we could not uncover substantial heterogeneity either between the two cell lines or cell subpopulations in the same cell line. In addition, existing methodologies do not allow us to concurrently titrate the infectivity levels of each cell, meaning that infected cells might not be uniformly infected, which could explain the lack of difference between the two populations. Our results are not unexpected given the low sensitivity of high-throughput single-cell methods in general. Indeed, this challenge becomes more pronounced for single nuclei sequencing techniques, where the amount of input RNA is even lower. We argue that our single-nucleus methodology is not sensitive enough to identify minute transcriptomic changes that could be relevant to prion propagation in an otherwise homogeneous cell line, especially when the median number of genes identified is low.

### 3.3.3 Validation of SPLiT-seq protocol using frozen mouse brain

We then proceeded to validate our protocol by processing an inherently heterogeneous sample of frozen mouse frontal lobe, which would also allow us to further optimise our protocol for input material that closely resembles the frozen human brain. Our latest, optimised protocol performed exceptionally well, generating much richer data than previously, which mirrored the sensitivity described in the original manuscript (Rosenberg et al., 2018). The plethora of different cell populations allowed us to assess bioinformatics pipelines based both on the Bioconductor ecosystem and the Seurat toolbox. Overall, we identified 17 clusters of cells, prior to any manual curation, and uncovered sets of gene markers that uniquely characterise each of them. We then used these markers to identify 7 broad cell populations: type IC spiral ganglion neurons, neurons, quiescent neural stem cells, oligodendrocytes, type II spiral ganglion neurons, oligodendrocyte precursor cells and endothelial cells.

# 4 Single-cell transcriptomics of murine prion disease

## 4.1 Introduction

### 4.1.1 Mouse models of prion disease

The long incubation times and fatal nature of prion diseases necessitate the use of animal models for their study. While some of the TSEs have animals as their primary host, most of those - like cattle, sheep and deer - are not suitable for controlled studies due to their long lifespan, large size, high cost of maintenance and technical difficulties in their scientific manipulation. In addition, primates can still be valuable models, especially for studying human prion diseases, and have been used in numerous studies (Comoy et al., 2013, 2015, 2017; B. Race et al., 2018). However, in addition to the inconveniences of maintaining other large animals, there is also the added controversy of using primates for scientific investigation.

Mice and hamsters are the two most widely used animal models for studying prion diseases (Watts & Prusiner, 2014). Both animals are small in size, easy to maintain, have short generation times and are easy to manipulate. Mice have gradually replaced hamster models due to the extensive study of their genome and the plethora of available molecular biology techniques that can be used for their genetic manipulation (Sebastian Brandner & Jaunmuktane, 2017).

Animal models are usually employed to recapitulate a specific aspect of the disease or to test a hypothesis and are not expected to faithfully recreate human disease, but facilitate the elucidation of biological questions. Nonetheless, wild-type and transgenic mice have been extensively used in prion research and some scientists have argued that the term “model” is inappropriate, as prion-inoculated mice do develop *bona fide* prion disease and recapitulate all biochemical and neuropathological hallmarks of human and animal disease (Watts & Prusiner, 2014), making them invaluable for testing new therapeutic interventions. This is in contrast to animal models used in other neurodegenerative diseases, such as Alzheimer’s disease, where most of the mouse models include autosomal-dominant mutations that mimic the familial and not the sporadic type of disease or do not model both amyloid-beta aggregation and tau dysfunction (King, 2018), or Parkinson’s disease, where none of the available models can perfectly mimic the neuropathology (α-synuclein aggregates, dopaminergic neurodegeneration) and recreate the clinical syndrome (Konnova & Swanberg, 2018). Remarkably, murine prion disease models have been shown to replicate aspects of the transcriptomic response to human neurodegenerative diseases (Burns et al., 2015).

Early experiments in the prion field used mouse models for transmission and adaptation studies to investigate prion strains and the species barrier by serially propagating sheep scrapie to wild-type mice. The introduction of transgenic mouse models allowed the design of more intricate experiments to dissect the species barriers with the use of mice expressing hamster PrP being a milestone that demonstrated the importance of the PrP amino acid sequence to incubation time, neuropathology and scrapie susceptibility and allowed the circumvention of the species barrier for the first time (Scott et al., 1989). Other milestones were the first attempt to model an Inherited Prion disease, GSS, by expressing the murine equivalent of the human P102L mutation in 1990 (Hsiao et al., 1990), and the generation of the first *Prnp* knock-out mouse in 1992 (Büeler et al., 1992).

Wild-type inbred mice also offer the advantage of a tightly controlled and consistent genetic background and were the first to be used for prion research. The most common strains used are C57Bl/6L, C57Bl/6N, C57BL/10, FVB, and 129/Ola (Sebastian Brandner & Jaunmuktane, 2017). While most of the research was focused on adapting scrapie prions to mice and enabling further propagation, some studies attempted to propagate and transmit human prions as well (A F Hill et al., 1997; Kitamoto et al., 1989). The long incubation periods and low attack rates underlined the existence of a species barrier and highlighted the importance of PrP homology, paving the way for the generation of humanised transgenic mice, i.e. mice expressing the human PrP homolog.

*Prnp* knock-out mice have also been extensively used in prion research, both directly, for studying the function of the cellular prion protein, and indirectly, by enabling the creation of transgenic mice devoid of murine PrP expression, which causes interference with the transgenes. Importantly, PrP null mice demonstrated the importance of the host PrP for the propagation of prions and the development of neuropathology and clinical scrapie disease (Sailer et al., 1994). PrP null mice are generated by removing large regions of the *Prnp* open reading frame, which halts the expression of $PrP^{C}$. While most of the models have minor phenotypes, this genetic manipulation has led to confounding and unexpected degeneration in a Japanese model, which was later attributed to the overproduction of the *Dpl* gene (R. C. Moore et al., 1999; Sakaguchi et al., 1996). Nevertheless, a direct connection to human neurogenerative disease has not been established (S Mead et al., 2000). Another confounding factor has been the use of embryonic stem cells from the 129Ola mouse strain while crossing with non-129 backgrounds. To address these shortcomings, a definitive study published in 2016 generated co-isogenic *Prnp* null mice on a pure C57BL/6J background that could not demonstrate any previously described phenotype, except a chronic demyelinating peripheral neuropathy, underlying the involvement of the cellular PrP in myelination and stressing the importance of the meticulous engineering of mouse models in general (Nuvolone et al., 2016).

The development of *Prnp* knock-out mice enabled the generation of humanised mouse models by intercrossing PrP null mice with transgenic mice expressing human or chimeric PrP. These models allowed researchers to overcome the species barrier and study human prion isolates while replicating the disease both biochemically and neuropathologically. Further developments introduced transgenic PrP overexpressing lines, which offered shorter incubation times, making them invaluable for disease modelling and drug discovery, albeit with the caveat that they mirror the human disease less faithfully (Sebastian Brandner & Jaunmuktane, 2017).

For this study, we opted to use wild-type mice, specifically the FVB/N strain. The selection of wild-type mice was of great importance as these mice express PrP at physiological levels and develop *bona fide* prion disease after inoculation with RML prions (Sandberg et al., 2011). The FVB/N inbred mouse strain originates from outbred NIH General-purpose Swiss mice established in 1935. Two strains were later selected for resistance to the action of histamine following a *Bordetella pertussis* vaccination. A subgroup of sensitive mice in the eighth generation was found to carry the Fv-1$^b$ sensitivity allele to the B strain of Friend leukaemia virus. These homozygous mice were then inbred and designated as the FVB strain (Taketo et al., 1991).

FVB/N mice have been routinely used for the study of prion diseases directly, and for the generation of FVB-congenic mouse lines (Asante et al., 2002, 2015; Sebastian Brandner & Jaunmuktane, 2017). It is a well-characterised mouse model that has, more recently, been used to dissect the mechanistic phases of prion propagation and toxicity (Sandberg et al., 2011). In this milestone publication, Sandberg et al. demonstrated that prion propagation in mouse brain proceeds in two mechanistically distinct phases: the first is an exponential phase, which is not rate-limited by PrP concentration and has no clinical symptoms, and the second is a plateau phase, which determines the time to clinical onset in an inversely proportional manner to PrP concentration. FVB/N wild-type mice with physiological PrP expression levels exhibited a mean incubation period of 137 days. In their follow-up paper, the authors extensively studied the kinetics of prion infection and toxicity and the neuropathology of RML-inoculated FVB/N mice, generating a large amount of valuable data that can later be integrated with our newly generated snRNA-seq data to draw meaningful conclusions (Sandberg et al., 2014).

### 4.1.2 Quantifying prion infectivity – The Scrapie Cell Assay

Quantifying prion infectivity is often essential for prion research as it is necessary for assaying the efficiency of purification procedures or the efficacy of treatment; however, it can be challenging as it requires a suitable biological system that can effectively propagate prions. Early observations that some prions can be propagated in mice led to the development of the first end-point titration approaches which were used to estimate the infectivity of biological material by assessing the survival of prion-inoculated mice (Chandler, 1963). These methods were time-consuming, tedious, and expensive, requiring around 12 months and 60 mice to quantify the infectivity of a single sample. The long incubation times meant that research would have to be effectively stalled for months until results were obtained and used to plan future experiments. In addition, the considerable number of animals required and the associated cost for their housing and maintenance made running multiple experiments in parallel impractical or impossible.

One of the first optimisations of the animal-based assays was the introduction of the incubation time interval assay where measurements of the intervals between inoculation and disease onset and inoculation and death are correlated with the titre of the infectious scrapie agent (Prusiner et al., 1982). The combination of these assays with inocula that could produce scrapie in the Syrian hamsters in only around 70 days after intracerebral inoculation meant a substantial reduction of the time required to quantify infectious samples, from 12 months to just over 2 (Kimberlin & Walker, 1977). Importantly, time interval assays require only a fraction of the number of animals, reducing cost and increasing the number of experiments that can be run in parallel (Prusiner, 1998).

The extensive research and long history of animal bioassays make them the gold standard for quantifying prion infectivity and incubation time to this day. However, as the use of animals is a necessity, they remain relatively time-consuming and expensive, even after further optimisation. The appeal of an *in vitro* system that could partially replace animal-based studies led to the development of alternative cell-based and cell-free methodologies. These include the protein misfolding cyclic amplification (PMCA) (Saborio et al., 2001), the real-time quaking-induced conversion (RT-QuIC) (Atarashi et al., 2008), and the scrapie cell assay (SCA) (Klöhn et al., 2003; Mahal et al., 2008).

PMCA and RT-QuIC are cell-free methods that involve the incubation of an infectious seed that contains $PrP^{Sc}$ with an appropriate template (brain homogenate that contains $PrP^{C}$ / recombinant $PrP^{C}$) under conversion-enabling conditions. Even though these techniques can be used to amplify infectious material (PMCA) and are helpful for the rapid diagnosis of clinical samples (RT-QuIC), a key disadvantage is that only indirect measurement of infectivity is possible. To address these limitations, the Weissmann group developed a cell-based infectivity assay termed standard scrapie cell assay (SSCA) or scrapie cell assay (SCA) in 2003 with research led by Peter-Christian Klöhn (Klöhn et al., 2003). The researchers first subcloned and then isolated highly susceptible neuroblastoma N2a cells, termed N2aPK1 cells, which were then exposed to infectious material for 3 days, grown to confluence and split 1:10 three times to remove any remaining starting material. The number of $PrP^{Sc}$-containing cells was then quantified using automated approaches and used as a proxy to estimate the infectious titre of the original sample. The authors claim that this method provides sensitivity comparable to the gold standard animal bioassays while reducing the assay time to a few days and the cost to only a fraction of the original assays, while the added benefit of easy automation allows experiment parallelisation of unprecedented scale. For applications where maximum sensitivity is desired, an end-point titration format of the scrapie cell assay can be used, the scrapie cell assay in end point format (SCEPA).

Although the SCA revolutionised the field of prion infectivity titration, there remain several challenges pertaining to the sensitivity to different prion strains, the genetic instability of the N2a cells and the introduction of false positives when using steel wires. The original assay used N2aPK1 cells, which are highly susceptible to RML prion infection but show variable levels of sensitivity to other murine-adapted prion strains. To extend the usability of the SCA to more prion strains, the Weissmann group assembled four cell lines (N2a-PK1, N2a-R33, LD9 and CAD5), subclones of the N2a, CAD5 and L292 cells, and quantified their responses to four murine-adapted prion strains (RML, 22L, 301C, and Me7) (Mahal et al., 2007). The authors point out the heterogenous response of sibling subclones, which highlighted the instability of these cell lines, especially when they underwent several serial passages. The genetic instability and variable susceptibility of the N2a cell lines were further validated one year later by Chasseigneaux et al. (Chasseigneaux et al., 2008), while a specific genetic signature of prion susceptibility could not be identified. While the SCA is suitable for murine-adapted prions, there have been successful attempts to quantify ovine scrapie (RK-13 cells that express ovine $PrP^{C}$) (Arellano-Anaya et al., 2011; Courageot et al., 2008; Neale et al., 2010) and Chronic Wasting Disease (CWD) (Elk21 cells) (Bian et al., 2010). Unfortunately, no cell systems have been developed to date for the propagation of bovine and human prions. Finally, the use of steel wires with the SCA has led to the identification of a positive signal in control groups that contained only normal brain homogenate (Edgeworth et al., 2010). The authors argued that this unexpected false-positive result could be due to a catalytic conversion of normal cellular prion protein to prions on the surface of the wires, or because of an increase in the concentration of prions that were already present in previously unidentifiable amounts. Overall, this issue underlines the importance of carefully controlled experiments to minimise false positives that could emerge due to background noise.

### 4.1.3 Chapter summary

In this chapter, we applied the SPLiT-seq methodology to transcriptionally profile murine prion disease. We designed a time-course experiment of RML- and control brain homogenate-inoculated FVB mice to track the temporal cellular response to prion disease. We generated a rich dataset of more than 200,000 high-quality transcriptomes, which we then analysed to identify perturbed transcripts and gene networks. We included additional modalities to the dataset, including prion infectivity measurements and immunohistochemical observations. Finally, we validated some of our findings using real-time quantitative PCR and RNA in situ hybridisation.

## 4.2 Results

### 4.2.1 Tissue collection

We designed a time-course experiment to study RML prion disease in mice under tightly controlled experimental settings and at a single-cell level. The experimental design was based on previous observations and studies published by groups of our Institute and other external research groups (Sandberg et al., 2011, 2014; Scheckel et al., 2020). The RML-inoculated FVB mouse model was characterised in depth by Sandberg's studies, providing the necessary information to allow us to select 5 different time points when samples would be collected (Figure 4.1a). After careful assessment of the prion infectivity curves published in the same study, we identified the following time points to be of interest as they represent the different stages of prion accumulation and could, thus, be more suitable for integrating our transcriptomics data with prion infectivity: 20 dpi and 40 dpi fall in the beginning of the exponential phase, temporary closer to inoculation and could be used to investigate early disease mechanisms and signatures of vulnerability; 80 dpi stands at the end of the exponential phase and the beginning of the plateau phase and could provide information to explore the mechanistic shift in prion replication; 120 dpi is representative of the plateau phase, before the clinical onset of disease and could be important in elucidating early mechanisms of neurotoxicity; finally, disease end-stage is defined as the start of clinical signs when scrapie sickness is confirmed, and — although it does not coincide with the actual terminal stage of disease due to ethical concerns — can provide valuable information regarding mechanisms of toxicity and cell death (Figure 4.1b).

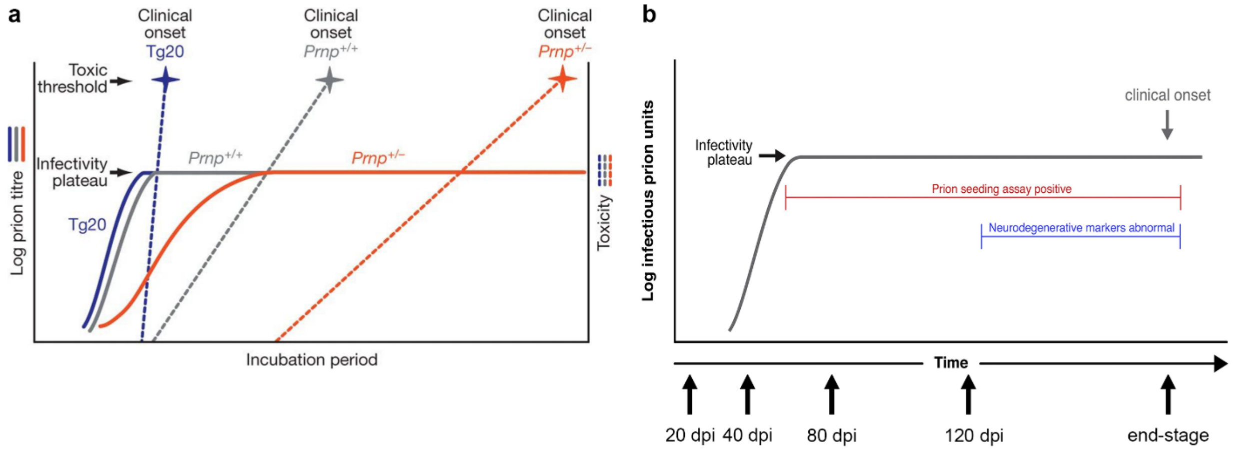


***Figure 4.1: Experimental design of mouse transcriptomics study. (a)*** *Two phases of prion propagation in vivo, as reported by Sandberg's study. We are interested in the kinetics of the Prnp$^{+/+}$ model, designated with the grey line since we are using wild-type mice with two copies of the Prnp gene. Adapted from Sandberg et al. 2011* ***(b)*** *Distribution of the 5 time points assessed in our study, compared to prion infectivity titre. 20 dpi and 40 dpi time points are located at the beginning of the exponential phase, 80 dpi time point is located at the beginning of the plateau phase, 120 dpi time point is located in the plateau phase, before the appearance of clinical signs, and end-stage is located at the clinical onset of the disease. Adapted from Mok & Mead, 2020.*

Scheckel's study underlined the importance of time-course data, as transcriptomic changes are suggested to be dynamic in the temporal dimension. The study was also appropriately controlled, using uninfected brain homogenate for the inoculation of the control groups. Since having a tightly controlled study would be of paramount importance for the interpretation of future results, we decided to include two control groups in our experiment. One of the groups was inoculated with uninfected CD1 brain homogenate to control for the RML-inoculated group so that the only variable assessed would be the presence of clinical stage brain, presumably containing prions and non-propagating toxic materials (Sandberg et al. 2014). The RML brain homogenate is also produced in a CD1 background and diluted with CD1 brain homogenate to minimise genetic heterogeneity. The additional control group was inoculated with PBS only and comparing it to CD1-inoculated controls could allow us to identify technical noise and transcriptional changes caused by the intracerebral introduction of a foreign brain homogenate in a live animal. Importantly, to minimise external variability all animals were inoculated at approximately the same age (flexibility of up to 2 weeks was allowed for technical reasons), animals at the same time point were inoculated with the same volume of inoculum on the same day for all 3 groups, and when a diseased animal was culled, a control animal was also culled on the same day.

By the end of the experiment, 4 mice were either found dead or culled due to health concerns (1 from the end-stage CD1 group, 3 from the end-stage RML group), and one mouse of the end-stage RML group did not develop scrapie after 213 dpi and was culled. These samples were excluded from further analyses (animal IDs: 829979, 829990, 829991, 829992). Animals from the end-stage RML group developed scrapie symptoms at a mean of 168 dpi (SD = 5).

Brain, blood, and plasma were collected and stored appropriately. The left-brain hemisphere was formalin-fixed and processed for immunohistochemistry, while the right was snap-frozen and stored at -80ºC until used for the preparation of brain homogenate for the scrapie cell assay or processed for single-nucleus sequencing. Blood and plasma were stored for future studies.

### 4.2.2 Pathology and immunohistochemistry

To ensure that our inoculation experiment was successfully concluded, we performed immunohistochemistry analyses on fixed brains from each cull. These brains were stained with the anti-PrP antibody ICSM35 and visualised under a microscope to assess abnormal PrP deposition and spongiform changes (Figure 4.2). Localised diffuse synaptic deposition of abnormal prion protein was evident at 40dpi in the cortex, hippocampus, and thalamus, which became more evident in the cortex, thalamus, midbrain, and brainstem at 80dpi, and increased throughout the course of the disease. Mild spongiosis in the hippocampus and thalamus was evident as early as 80 dpi and became more pronounced as the disease progressed.

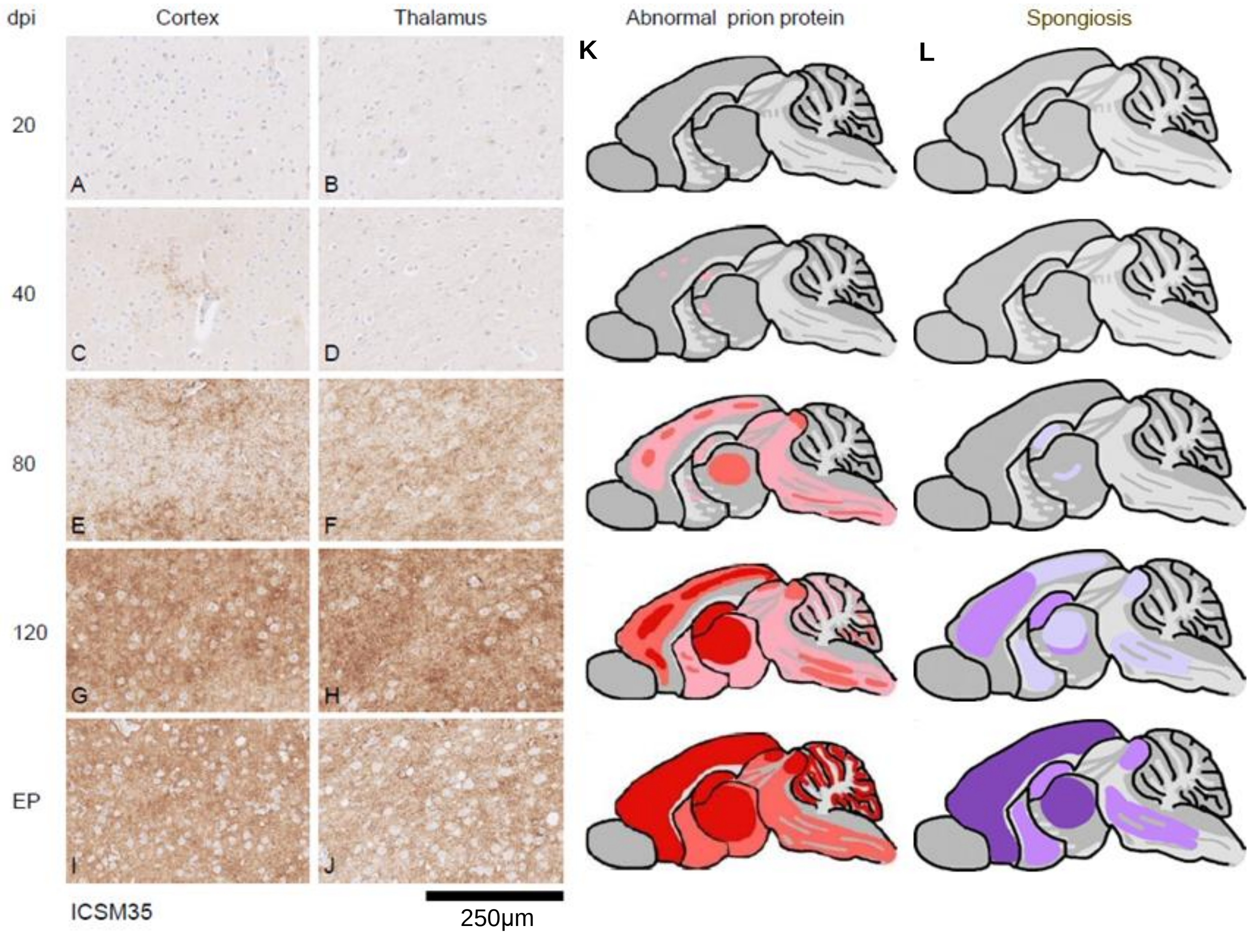


***Figure 4.2: Time course of abnormal prion protein accumulation in the brain of FVB mice inoculated with RML.*** *FVB mice were intracerebrally inoculated with RML prions and groups of mice were culled at 4 defined time points (20-120 dpi) and the onset of clinical prion disease (EP). Formalin-fixed brains from each time point were analysed for abnormal PrP deposition and spongiosis. The bright-field microscope images (A-J) show the abnormal PrP deposition in the cortex and thalamus using the ICSM35 antibody for staining. (**K**) PrP deposition became evident at 80 dpi (images E, F) and became more pronounced as the disease progressed. The schematic is an overview of the distribution of prion protein deposits, where graded red shades reflect the intensity of prion protein deposits. (**L**) Mild spongiosis in the hippocampus and thalamus was evident in 9/10 animals at 80 dpi and became more pronounced in 10/10 animals as the disease progressed. The schematic is an overview of the distribution of vacuoles, where graded purple shades reflect the intensity of spongiosis. The hippocampus has not been assessed for neuronal loss. EP: disease endpoint.*

Regarding the presence of residual inoculum in the early time points, a proportion of the brains (4/15) at time point 20 dpi, inoculated with RML, showed immunopositive material located at the fringe between hippocampus and corpus callosum. This material appears as small, solid, and densely immunoreactive. Occasionally, there are processes, presumably from astrocytes of the hippocampus, which also show weak immunolabelling. The interpretation of this finding is that the deposits represent residual inoculum, and we interpret the presence of immunoreactive material in astrocytes processes as an early uptake.

A similar finding in 2/15 animals of the 40 dpi RML group is observed, but in addition, there is also fine granular immunopositive material, more in keeping with incipient, de novo production of the abnormal prion protein. In the subsequent time points (80, 120 dpi, and endpoint) no such “residual inoculum” is identified. There is widespread de novo deposition of abnormal prion protein as expected in these time culls. No immunoreactive material is seen in the 2 control groups (normal brain homogenate and PBS).

### 4.2.3 Prion infectivity titration

We used the automated scrapie cell assay to titrate the prion infectivity of the mouse brain samples. 3 right-brain samples from each time point of the CD1 and RML groups, 30 samples in total, were separately homogenised and used to infect susceptible PK1 cells, as previously described. The $PrP^{Sc}$ spot count was quantified after the 2nd and 3rd splits of the bioassay cells. All samples except one were assayed on the same SCA experiment due to space constraints. The last sample (RML group, end-stage time point) was assayed in a separate experiment. One CD1 control sample showed low levels of infectivity, which was suspected to be an artefact due to cross-contamination and was assayed 3 more times in a separate experiment, which all gave negative results confirming our suspicions; the original, low-infectivity value has been replaced with zero. One sample showed a low level of infectivity at 20 dpi, two samples showed higher levels of infectivity at 40 dpi, while infectivity plateaued at 80 dpi with all three samples showing high infectivity levels, which remained until the end-stage (Figure 4.3). A closer inspection of the plot suggests that the exponential phase spans the 20-80 dpi time frame, when the plateau phase begins, validating the effective selection of the 5 time points and agreeing with the findings of previous studies. Supplementary Table 2 contains the raw and log-transformed infectious units of all samples assayed at the end of the 3rd split.

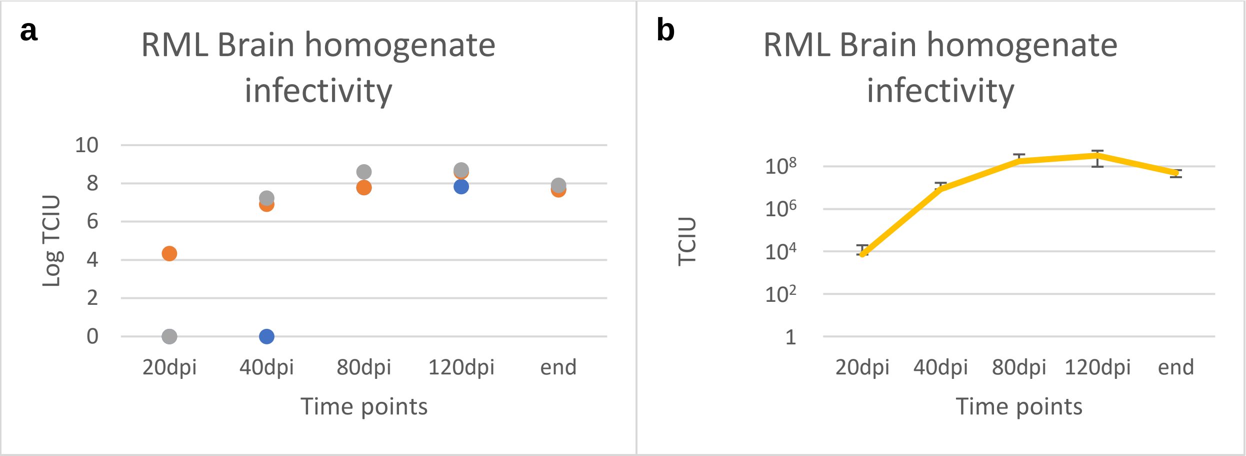


***Figure 4.3: The scrapie cell assay validates the effective selection of the 5 time points for the mouse transcriptomics study. (a)*** *Log-transformed infectious units from 3 individual samples are plotted on a linear y-axis. Each colour represents a biological replicate (**b)** The mean of infectious units from the same samples (n = 3) is plotted on a logarithmic y-axis. Error bars indicate the standard deviation. TCIU: tissue culture infectious units, referred to as infectious units in the text.*

### 4.2.4 Single-nucleus RNA sequencing

#### Nuclei extraction, library preparation, and sequencing

We performed single-nucleus RNA sequencing on the mouse brain tissue using the SPLiT-seq protocol, which showed favourable results as discussed in the previous chapter. As a first step, the frontal lobes of the mouse brains were dissected and processed to prepare nuclei suspensions, which were then fixed and stored. All suspensions were examined under the microscope for quality assurance. We noticed the presence of more cellular debris in some suspensions, but the amount was not quantified. Due to earlier test runs, the suspensions of the 20 dpi time point had been thawed and frozen twice, while for all other time points they had only thawed once when used for the library preparation. We then attempted to quantify the infectivity using the SCA as previously described (section 4.2.3), however, the fixed nuclei suspensions showed no infectivity, hindering the integration of infectivity data from each sample with the transcriptomic information.

When all samples for a specific time point were ready, a single plate including all 3 groups of samples was prepared and processed through the SPLiT-seq library preparation protocol. Supplementary Figure 2 shows a representative image of the layout of a loaded

SPLiT-seq 96-well plate. At the end of the split-pool barcoding rounds, the number of nuclei recovered was quantified and the yield was calculated to be as expected for all time points, except from the 20 dpi, which had lower starting sample concentrations (Table 2). The rest of the protocol was then followed with the next possible quality control steps after cDNA amplification and after library tagmentation. Supplementary Figure 3 shows representative TapeStation traces after the aforementioned steps. Parts of the protocol had been repeated until all libraries generated TapeStation traces that had the expected size distribution.

| | Starting sample concentration (nuclei / µL) | Final yield (nuclei) | Percentage recovery | Number of sublibraries prepared | Number of nuclei per sublibrary |
|---|---|---|---|---|---|
| **20 dpi** | 1200 | 33000 | 6.37% | 3 | 9500-10000 |
| **40 dpi** | 2000 | 100300 | 13% | 6 | 15000 |
| **80 dpi** | 2000 | 82600 | 10.70% | 6 | 15000 |
| **120 dpi** | 2000 | 142800 | 18.60% | 6 | 15000 |
| **end-stage** | 2000 | 113050 | 14.70% | 6 | 15000 |

***Table 2: Final yield and single nuclei recovery after split-pool barcoding rounds.*** *For the 20 dpi, nuclei suspensions were prepared at a lower concentration, leading to a lower yield, a smaller number of sublibraries prepared, and a lower number of nuclei per sublibrary. For all other time points, the yield was higher, and 6 libraries were prepared, all of 15,000 nuclei.*

Three of the final, tagmented libraries from each time point were then pooled and sequenced during a total of 5 sequencing runs, one for each time point, yielding an expected number of nuclei sequenced around 45 thousand per time point, or 225 thousand in total. Sequencing generated approximately 2,085 million reads in total (Table 3). The quality of sequencing was then assessed by running FastQC on the resulting fastq files and examining the summary statistics of the reports (External Supplementary File 1). After making sure that all libraries were adequately sequenced and high quality, the fastq files were processed using the SPLiT-seq bioinformatics pipeline to demultiplex the biological samples and generate the count matrices, which were loaded into Seurat for further analysis.

| Time point | Library number | Sequencing reads (in millions) |
|---|---|---|
| **20 dpi** | 1 | 139.5 |
| **20 dpi** | 2 | 110.5 |

| | | |
|---|---|---|
| **20 dpi** | 3 | 128.3 |
| **40 dpi** | 1 | 213.1 |
| **40 dpi** | 2 | 178.3 |
| **40 dpi** | 3 | 187.4 |
| **80 dpi** | 1 | 118 |
| **80 dpi** | 2 | 117.5 |
| **80 dpi** | 3 | 96.7 |
| **120 dpi** | 1 | 124.4 |
| **120 dpi** | 2 | 120.2 |
| **120 dpi** | 3 | 142.9 |
| **end-stage** | 1 | 129.1 |
| **end-stage** | 2 | 140.8 |
| **end-stage** | 3 | 138.1 |
| | **Total** | **2084.8** |

***Table 3: Sequencing generated a total of 2085 million reads across all libraries.***

### Quality control in Seurat

We filtered cells based on their feature count (features are equivalent to genes in the context of the analysis, and the two terms will be used interchangeably) and the percentage of mitochondrial genes expressed. Cells with a low feature count are less informative and might represent background noise, while cells with a very high feature count might correspond to multiplets. Cut-off values were set to between 250 and 2500 features based on published literature (Mathys et al., 2019; Rosenberg et al., 2018) and previous tests. Regarding mitochondrial transcripts, for single-cell methods they are indicative of mitochondrial rupture and cell death, however since we are using a single-nucleus method where mitochondria should have been removed during suspension preparation, a high percentage would indicate a serious failure of the technique. Thus, the acceptable threshold was intentionally set very low, to less than 1%.

More than half a million cells were identified prior to quality control, while the number was reduced to 210,710 when the filters were applied (Figure 4.4 and Table 4). The 20 dpi time point showed the highest difference between the unfiltered and filtered data, and this was attributed to the way the cell identification algorithm of the split-seq toolkit works and will be further discussed later. Importantly, by setting our manual filtering criteria any discrepancies between the different libraries can be removed, resulting in 210 thousand high-quality cell transcriptomes. All time points have been successfully sequenced resulting in 36 to 51 thousand high-quality transcriptomes each after filtering (M = 1050 features per nucleus, SD = 565). Samples from both RML and CD1 groups had a similar number of identified cells, while samples from the PBS group comprised fewer cells (Figure 4.5). This was expected and part of the experimental design since all PBS samples included cells from 2 wells of a 96-well plate, while RML and CD1 groups included cells from 2 or 3 wells of the plate (visualised in the layout of a SPLiT-seq plate in Supplementary Figure 2). There was one outlier sample (from mouse #828719) which only contributed 61 cells in the 20 dpi dataset after filtering. This sample was located in the last two wells of the 96-well plate and the small number of cells could be attributed to the exhaustion of the reverse transcription master mix. All subsequent analyses were performed both including and excluding the spurious sample and produced comparable results, so we decided not to filter it out. Supplementary Figure 4 includes more detailed violin plots of the number of features for each biological sample before and after filtering, as well as correlation plots between the numbers of counts and features. Supplementary Table 3 includes detailed numbers of cells identified from each biological replicate.

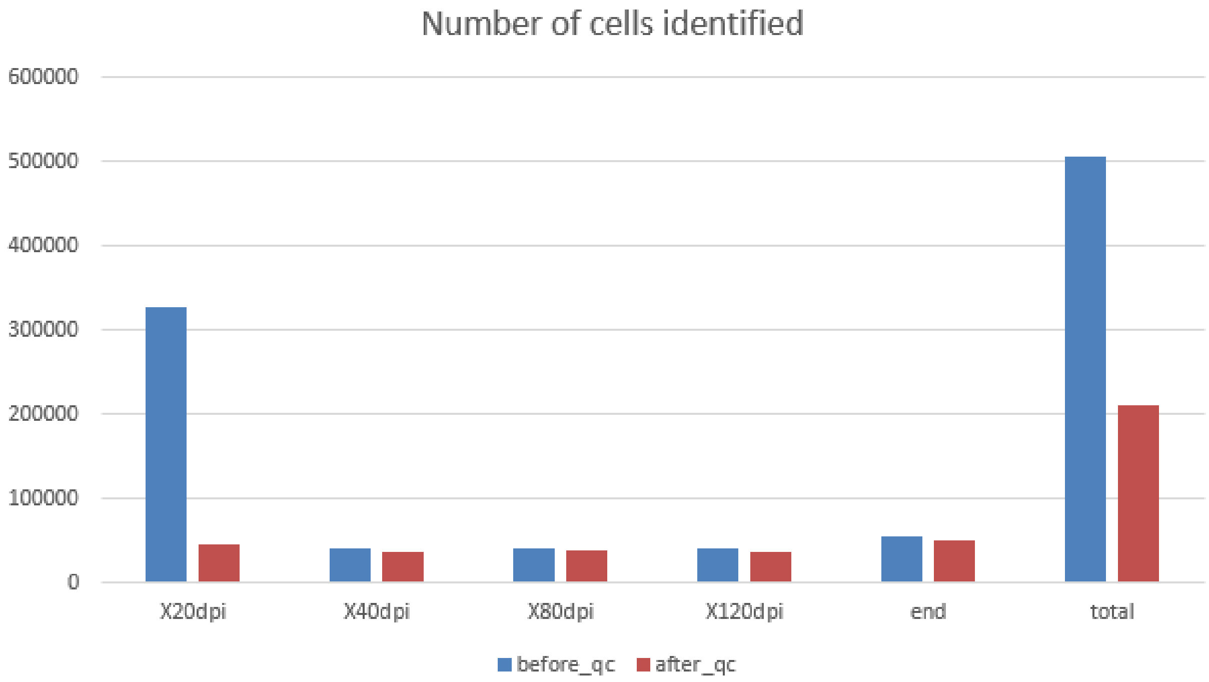


***Figure 4.4: The analysis identified more than half a million cells before quality control and approximately 200 thousand cells after filtering.*** *While filtering removed only a small fraction of cells for*

*most time points, there was a substantial reduction when the 20 dpi dataset was filtered. This was attributed to the automatic setting of a less stringent threshold of the cell identification algorithm and was easily rectified when applying our custom filtering criteria.*

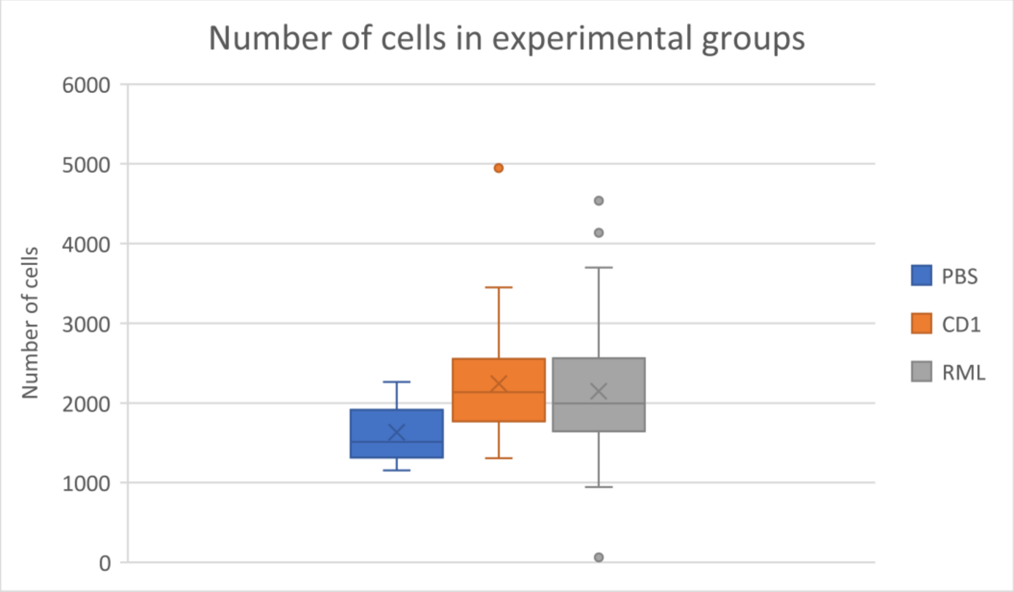


***Figure 4.5: Samples from both RML and CD1 groups had a similar number of identified cells, while samples from the PBS group comprised fewer cells, as expected from the experimental design.*** *The box and whisker charts show the distribution of the number of cells identified per biological sample into quartiles, highlighting the mean (x symbol), median (horizontal line), and outliers (coloured dots). Sample 828719 is the RML outlier with the lowest number of identified cells.*

| | 20 dpi | 40 dpi | 80 dpi | 120 dpi | End-stage | Total |
|---|---|---|---|---|---|---|
| **Before QC** | 327017 | 41756 | 41750 | 40461 | 55986 | 506970 |
| **After QC** | 46582 | 37355 | 39158 | 36392 | 51223 | 210710 |

***Table 4: The analysis identified 210 thousand high-quality transcriptomes across all time points.***

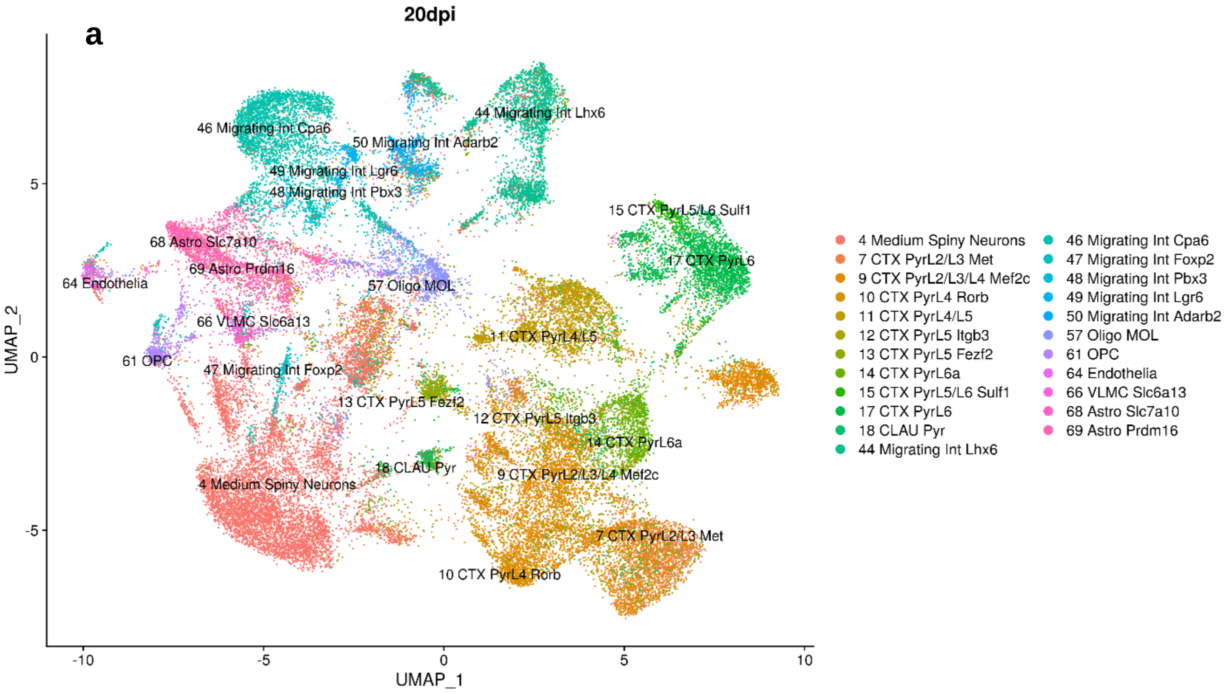
Complete removal of mitochondrial and rRNA genes, as well as regressing them were also attempted, however, neither had any impact on the conclusions of downstream analyses and we decided to proceed with the fewest data modifications possible, including all genes. In addition, no bias was evident in the distributions of transcript lengths or chromosomes of the identified features (Supplementary Figure 5).

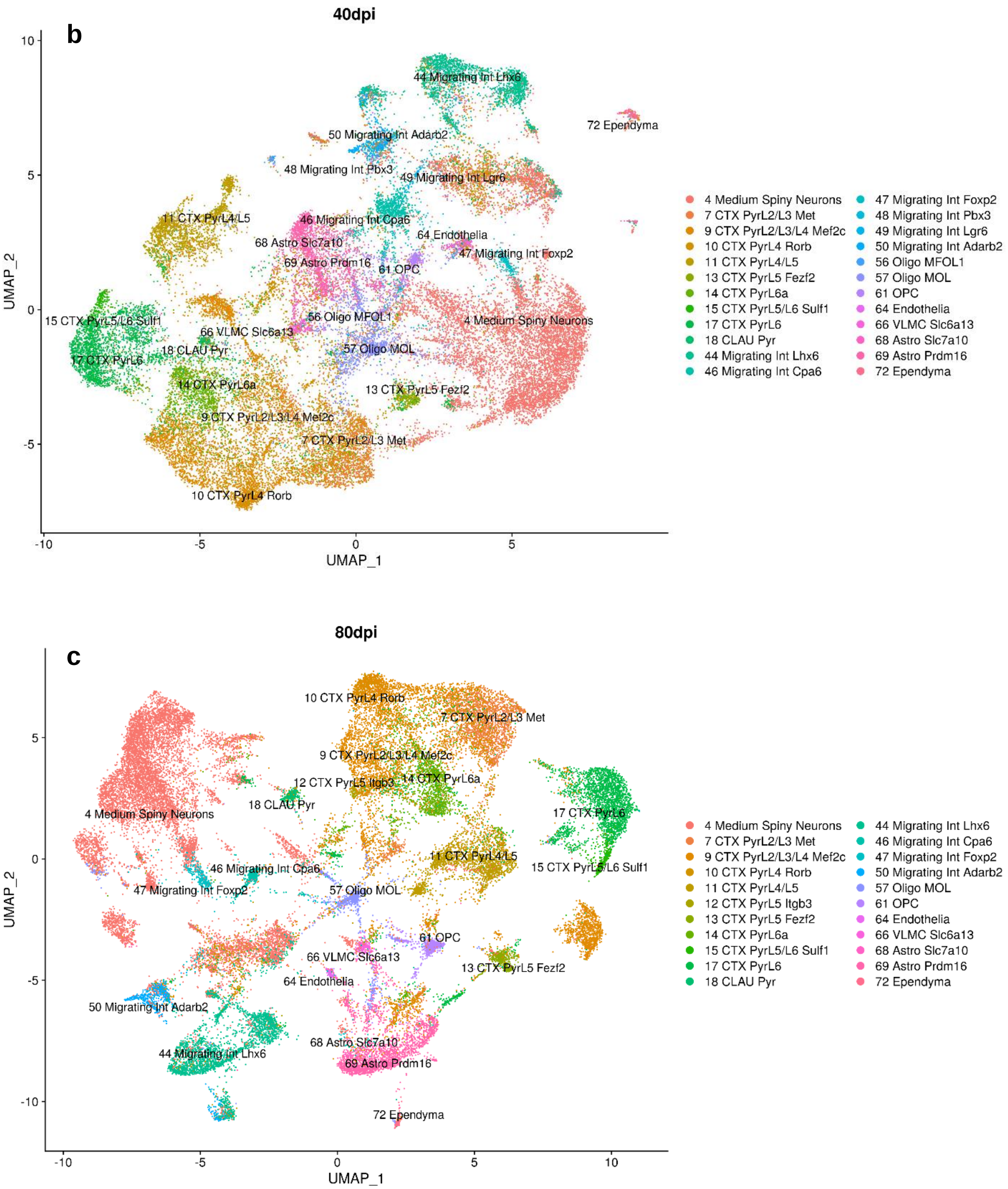
The effects of the cell cycle were also assessed. It is known that differences in the cell cycle stages can introduce variability in the data and drive cluster separation. We assessed the effect of the cell cycle in our dataset by scoring each cell individually and annotating the most probable phase using known marker genes. PCA plots computed specifically on cell cycle features for each time point suggested that cell cycle effects are modest (Supplementary Figure 6). In addition, variability introduced by cell cycle differences is not detrimental to downstream analyses and is relevant to the underlying biology. Since the cell cycle effects were not very pronounced, we decided not to regress out the cell cycle genes and preserve this additional information.

### Clustering and cluster annotation

Since our dataset consisted of multiple time points that were individually sequenced and processed separately, we needed to perform clustering and annotation in a way that allows reproducibility across all time points and minimizes subjective decisions, which — even though are important for judging the results of analyses and making sure that algorithms operate as expected — often introduce variability and hinder comparisons across separate experiments. We decided to adopt a relatively novel approach of using an annotated dataset as a reference and transferring the annotation labels on the query datasets. This technique, introduced in the v3 version of Seurat, offers a data-driven approach to clustering and annotation (Stuart et al., 2019).

For the label transfer to be accurate, a suitable annotated reference dataset needs to be used. We used the single-nucleus dataset from the original SPLiT-seq manuscript, which was generated using the same protocol from the control mouse brain, after removing clusters of cells not present in the frontal lobe to increase specificity. The resulting annotated datasets are multi-dimensional hindering the conception of the underlying biology. After removing clusters of less than 100 cells, we performed dimensionality reduction using the UMAP algorithm to visualise cluster identities in a two-dimensional space (Figure 4.6). Cell identities are coloured on top of the visualisation and are not considered when calculating UMAP coordinates. Most cells that cluster together are also part of the same annotated cluster, confirming the success of the label transfer approach. A careful examination of the plots suggests that most of the cells identified are assigned to clusters of cortical neurons, while large clusters of medium spiny neurons are present, followed by clusters of migrating interneurons. Regarding the glial cells, astrocytes are the most abundant, while oligodendrocytes are also identified in large numbers. Smaller clusters of oligodendrocyte precursor cells are also visible. Microglia clusters could be identified in only the last two time points (120 dpi, end-stage), while for the other datasets they have not passed the filtering criteria of consisting of more than 100 cells and have been excluded. Other small clusters of cells are also visible in some datasets, including endothelial cells, ependymal cells, and vascular and leptomeningeal cells.




40dpi
b
44 Migrating Int Lhx6
72 Ependyma
50 Migrating Int Adarb2
48 Migrating Int Pbx3
49 Migrating Int Lgr6
11 CTX PyrL4/L5
46 Migrating Int Cpa6
64 Endothelia
68 Astro Slc7a10
47 Migrating Int Foxp2
69 Astro Prdm16
61 OPC
15 CTX PyrL5/L6 Sulf1
56 Oligo MFOL1
4 Medium Spiny Neurons
66 VLMC Slc6a13
18 CLAU Pyr
57 Oligo MOL
17 CTX PyrL6
14 CTX PyrL6a
13 CTX PyrL5 Fezf2
9 CTX PyrL2/L3/L4 Mef2c
7 CTX PyrL2/L3 Met
10 CTX PyrL4 Rorb
10
5
0
-5
UMAP_2
-10
-5
0
5
UMAP_1
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
48 Migrating Int Pbx3
49 Migrating Int Lgr6
50 Migrating Int Adarb2
56 Oligo MFOL1
57 Oligo MOL
61 OPC
64 Endothelia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma
80dpi
c
10 CTX PyrL4 Rorb
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
12 CTX PyrL5 Itgb3
14 CTX PyrL6a
18 CLAU Pyr
4 Medium Spiny Neurons
17 CTX PyrL6
11 CTX PyrL4/L5
46 Migrating Int Cpa6
15 CTX PyrL5/L6 Sulf1
47 Migrating Int Foxp2
57 Oligo MOL
61 OPC
66 VLMC Slc6a13
13 CTX PyrL5 Fezf2
64 Endothelia
50 Migrating Int Adarb2
68 Astro Slc7a10
44 Migrating Int Lhx6
69 Astro Prdm16
72 Ependyma
5
0
-5
-10
UMAP_2
-10
-5
0
5
10
UMAP_1
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
12 CTX PyrL5 Itgb3
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
50 Migrating Int Adarb2
57 Oligo MOL
61 OPC
64 Endothelia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma

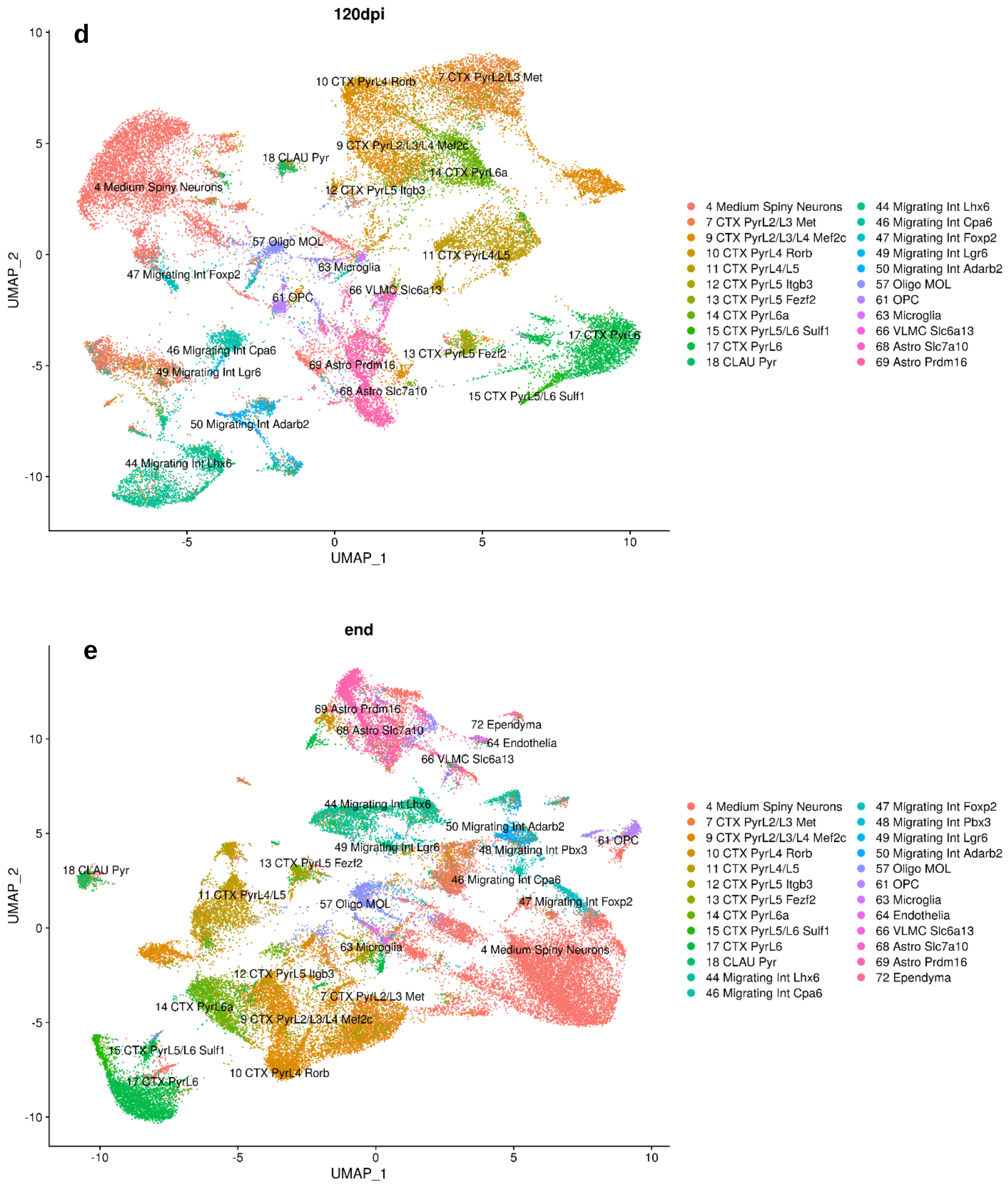
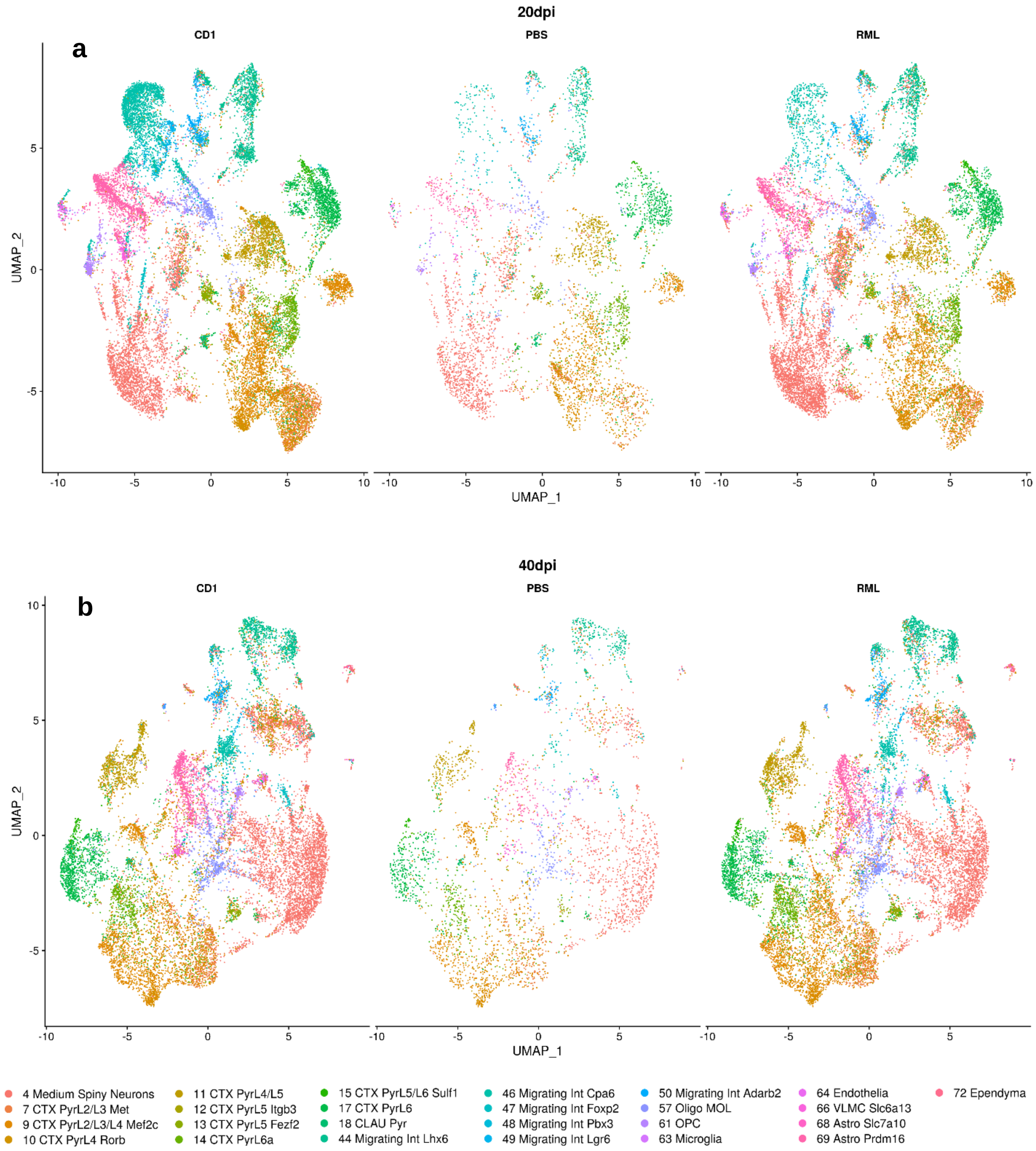
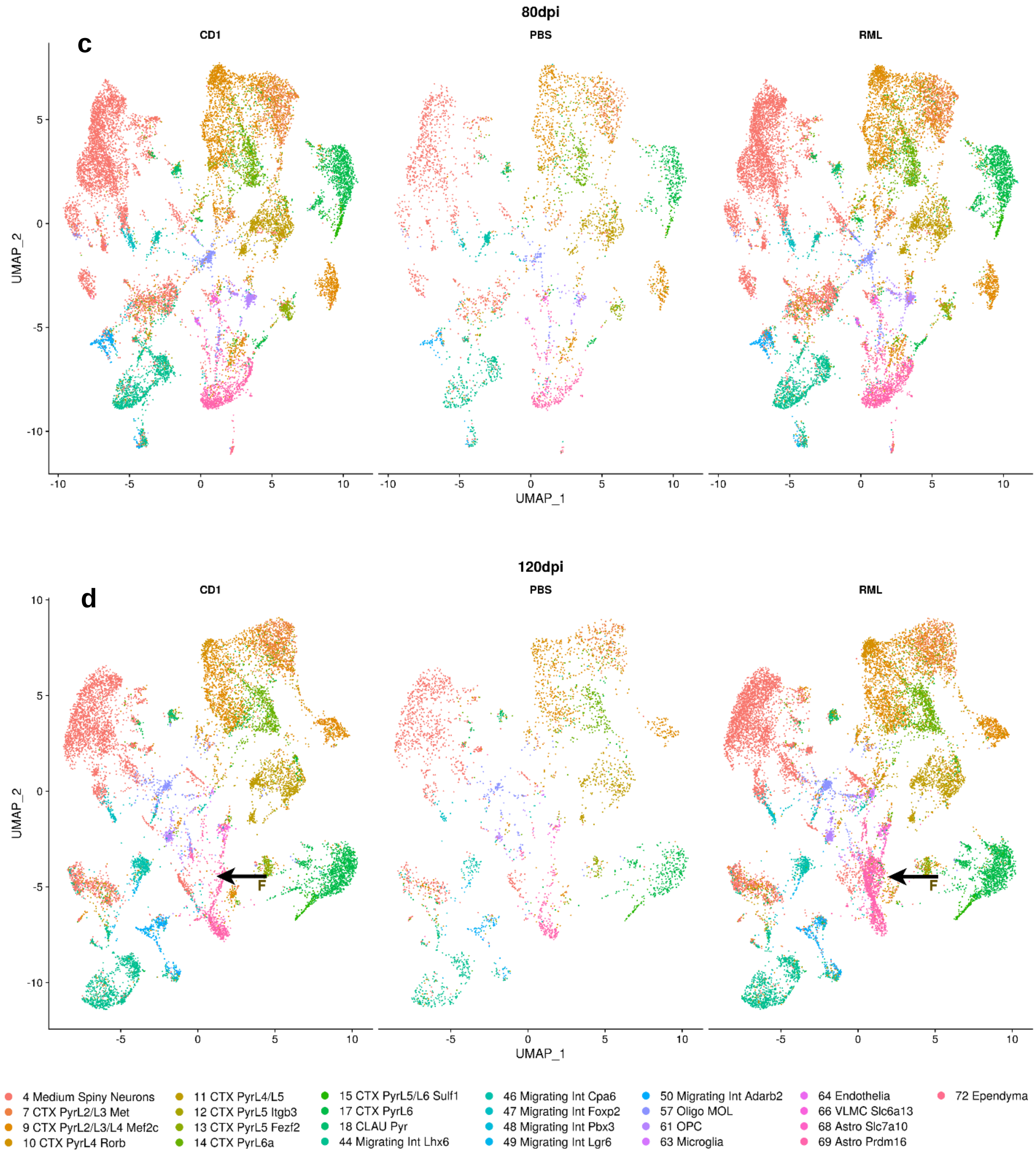


***Figure 4.6: Annotated UMAP plots of the 5 datasets show the relationship of the 26 identified clusters in low-dimensional space.*** *Label transfer and filtering were performed for each dataset separately. After dimensionality reduction using the UMAP algorithm, cell clusters can be visualised in the two-dimensional space. Each plot contains cells from one time point (**a:** 20 dpi, **b:** 40 dpi, **c:** 80 dpi, **d:** 120 dpi, **e:** end-stage) and all three experimental groups (RML, PBS, CD1). Each dot on the plot represents a*

*single cell, with its location defined by the two UMAP coordinates. Cells that cluster together share more similar gene expression patterns than cells further apart. Cell identities are coloured on top of the visualisation and are not considered during the calculation of UMAP coordinates. Most cells that cluster together are also part of the same annotated cluster, confirming the success of the label transfer approach. Cluster names consist of a number, which was maintained from the original dataset as a reference system, then the cluster name (e.g., Migrating Int, OPC, CTX), and an optional anatomical location (e.g., PyrL5) and/or an optional cluster-specific transcription factor (e.g., Slc6a13, Rorb). Pyr: pyramidal; L2/L3/L4 etc.: layer 2,3,4 etc.; CTX: cortex/cortical; CLAU: claustrum; Int: interneurons; MOL: mature oligodendrocytes; OPC: oligodendrocyte precursor cells; VLMC: vascular and leptomeningeal cells; Astro: astrocytes.*

While these visualisations are helpful for giving an overview of the datasets, they include cells from all experimental groups drawn together on a single plot, making group-wise comparisons impossible. To get an estimated visual overview of the impact of the disease on different cell types we used the same dimensionality reduction technique, but split each UMAP plot into three, one plot for each experimental group. Upon careful examination of the resulting plots, we can identify some interesting shifts in gene expression patterns of specific cell types: the first three time points do not show any striking differences, however, there are notable differences in the astrocytic population at 120 dpi, and at the astrocytes (clusters 66 and 68), the medium spiny neurons (cluster 4), and some subpopulations of cortical neurons (clusters 7 and 9) at the end-stage (Figure 4.7). There might be, of course, more subtle changes, but these will be identified by the gene expression analyses that will follow.


20dpi
a
CD1
PBS
RML
UMAP_2
UMAP_1
40dpi
b
CD1
PBS
RML
UMAP_2
UMAP_1
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
12 CTX PyrL5 Itgb3
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
48 Migrating Int Pbx3
49 Migrating Int Lgr6
50 Migrating Int Adarb2
57 Oligo MOL
61 OPC
63 Microglia
64 Endothelia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma


80dpi
c
CD1
PBS
RML
UMAP_2
5
0
-5
-10
-10
-5
0
5
10
UMAP_1
120dpi
d
CD1
PBS
RML
10
5
0
-5
-10
F
F
UMAP_2
UMAP_1
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
12 CTX PyrL5 Itgb3
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
48 Migrating Int Pbx3
49 Migrating Int Lgr6
50 Migrating Int Adarb2
57 Oligo MOL
61 OPC
63 Microglia
64 Endothelia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma

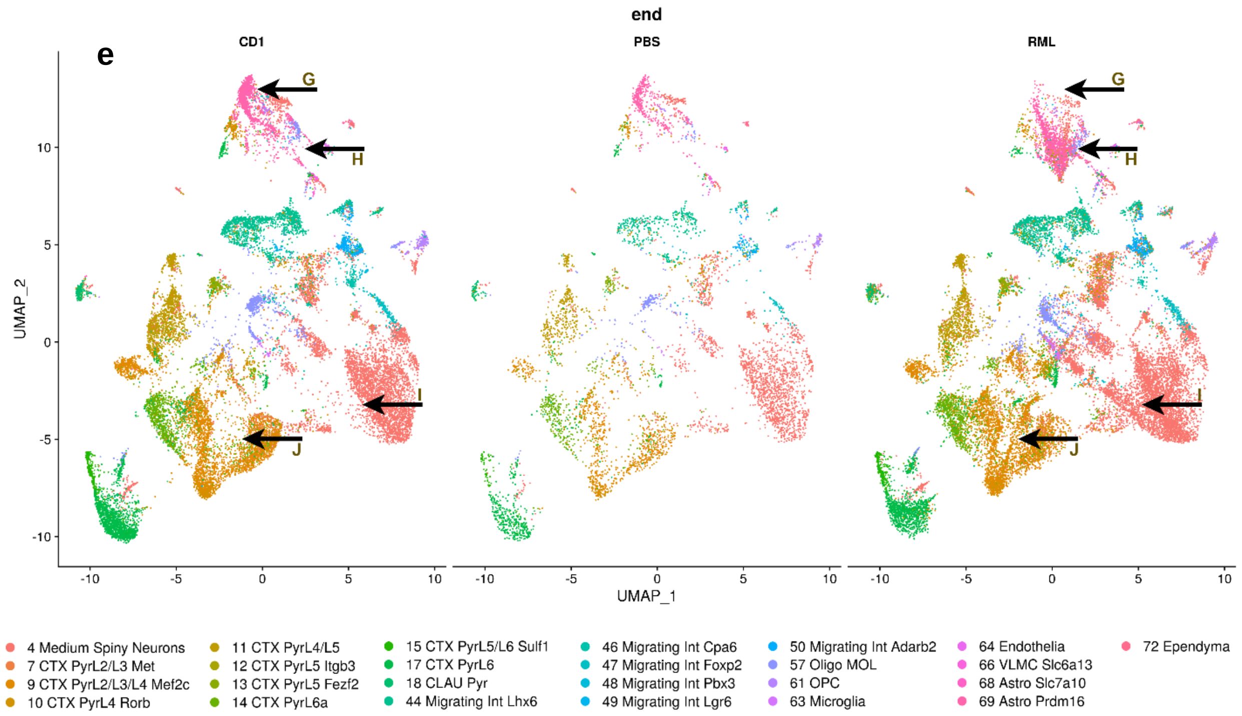


***Figure 4.7: UMAP plots split by experimental group suggest transcriptomic differences in neuronal and astrocytic populations at the last two time points.*** *The first three time points (**a** to **c**) do not show pronounced differences between the CD1 and RML plots. In **(d)** there is an observable shift of astrocytic populations (clusters 68 and 69; indicated by arrow F). In **(e)** there is a more pronounced shift of astrocytic populations (clusters 68 and 69; indicated by arrows G and H), while there are differences in the medium spiny neurons (cluster 4; indicated by arrow I) and subpopulations of cortical neurons (clusters 7 and 9; indicated by arrow J). Abbreviations are the same as in Figure 4.6.*

In total, 26 different cell populations that passed filtering criteria were annotated. Reassuringly, the number of cells in each population was similar across the 5 datasets (Table 5). We kept the numbering system of those populations as the original dataset published by Rosenberg et al. to facilitate comparisons between time points and experiments by having a common reference. The numbers are not continuous, since some of the original clusters were not relevant to the mouse frontal lobe and have been excluded before label transfer. Full cluster names consist of a number, which was maintained from the original dataset as a reference system, then the cluster name (e.g., Migrating Int, OPC, CTX), and an optional anatomical location (e.g., PyrL5) and/or an optional cluster-specific transcription factor (e.g., Slc6a13, Rorb). Other abbreviations include: Pyr: pyramidal; L2/L3/L4 etc.: layer 2,3,4 etc.; CTX: cortex/cortical; CLAU:

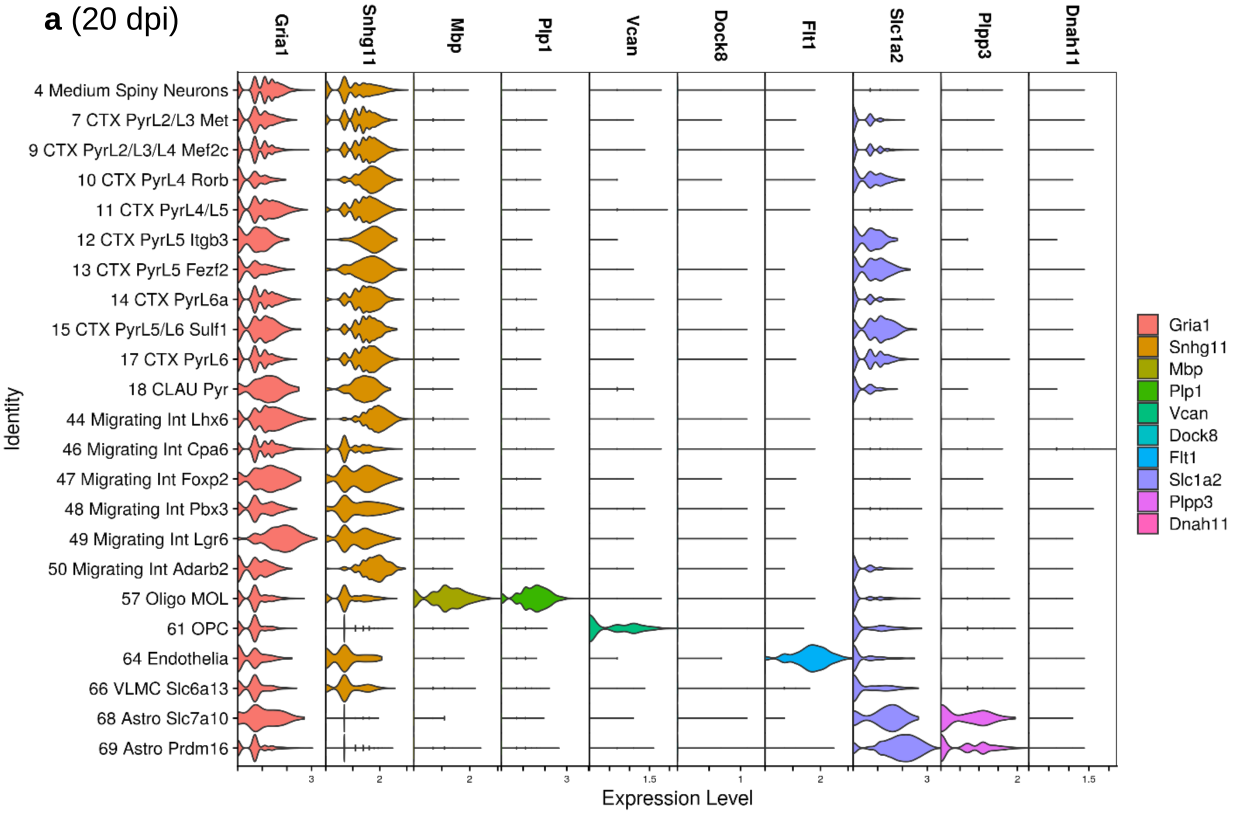
claustrum; Int: interneurons; MOL: mature oligodendrocytes; OPC: oligodendrocyte precursor cells; VLMC: vascular and leptomeningeal cells; Astro: astrocytes. Cluster names and numbers will be used interchangeably, and their relationship will be constant as described in Table 5 throughout the whole chapter and the discussion that follows. More cluster metrics can be found in Supplementary Table 4.

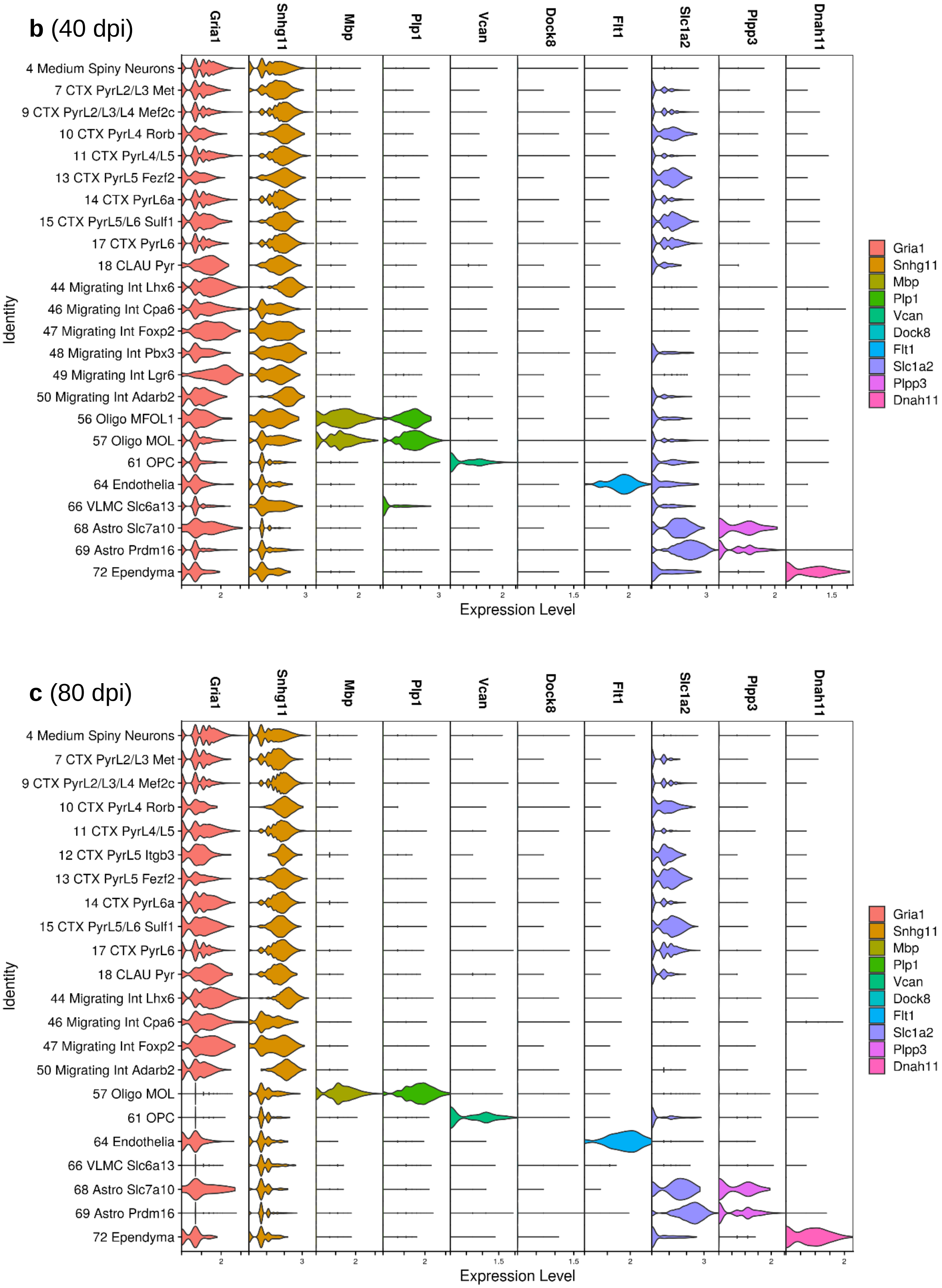
| **Cluster number** | **Cluster name** | **Cluster group** | **Number of cells** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **20 dpi** | **40 dpi** | **80 dpi** | **120 dpi** | **end-stage** |
| **4** | Medium Spiny Neurons | Medium Spiny Neurons | 9546 | 9977 | 11238 | 9111 | 14999 |
| **7** | CTX PyrL2/L3 Met | Cortical neurons | 1590 | 1165 | 1776 | 1561 | 816 |
| **9** | CTX PyrL2/L3/L4 Mef2c | Cortical neurons | 7758 | 7016 | 7424 | 6406 | 8786 |
| **10** | CTX PyrL4 Rorb | Cortical neurons | 581 | 425 | 212 | 471 | 667 |
| **11** | CTX PyrL4/L5 | Cortical neurons | 3033 | 2558 | 2349 | 2478 | 3131 |
| **12** | CTX PyrL5 Itgb3 | Cortical neurons | 124 | N/A | 116 | 119 | 113 |
| **13** | CTX PyrL5 Fezf2 | Cortical neurons | 413 | 293 | 395 | 371 | 412 |
| **14** | CTX PyrL6a | Cortical neurons | 2288 | 1827 | 1960 | 2228 | 3034 |
| **15** | CTX PyrL5/L6 Sulf1 | Cortical neurons | 405 | 392 | 451 | 450 | 644 |
| **17** | CTX PyrL6 | Cortical neurons | 3401 | 2920 | 3094 | 3126 | 3988 |
| **18** | CLAU Pyr | Cortical neurons | 225 | 184 | 342 | 235 | 470 |
| **44** | Migrating Int Lhx6 | Migrating interneurons | 2780 | 2198 | 2679 | 2447 | 3375 |
| **46** | Migrating Int Cpa6 | Migrating interneurons | 3437 | 1052 | 492 | 880 | 347 |
| **47** | Migrating Int Foxp2 | Migrating interneurons | 409 | 267 | 431 | 408 | 675 |
| **48** | Migrating Int Pbx3 | Migrating interneurons | 356 | 200 | N/A | N/A | 114 |
| **49** | Migrating Int Lgr6 | Migrating interneurons | 433 | 162 | N/A | 155 | 104 |
| **50** | Migrating Int Adarb2 | Migrating interneurons | 598 | 409 | 564 | 622 | 736 |
| **56** | Oligo MFOL1 | Oligodendrocytes | N/A | 188 | N/A | N/A | N/A |
| **57** | Oligo MOL | Oligodendrocytes | 1268 | 1539 | 885 | 918 | 1752 |
| **61** | OPC | Oligodendrocyte precursors | 641 | 460 | 570 | 457 | 708 |
| **63** | Microglia | Immune | N/A | N/A | N/A | 171 | 238 |
| **64** | Endothelia | Vascular | 229 | 152 | 118 | N/A | 152 |

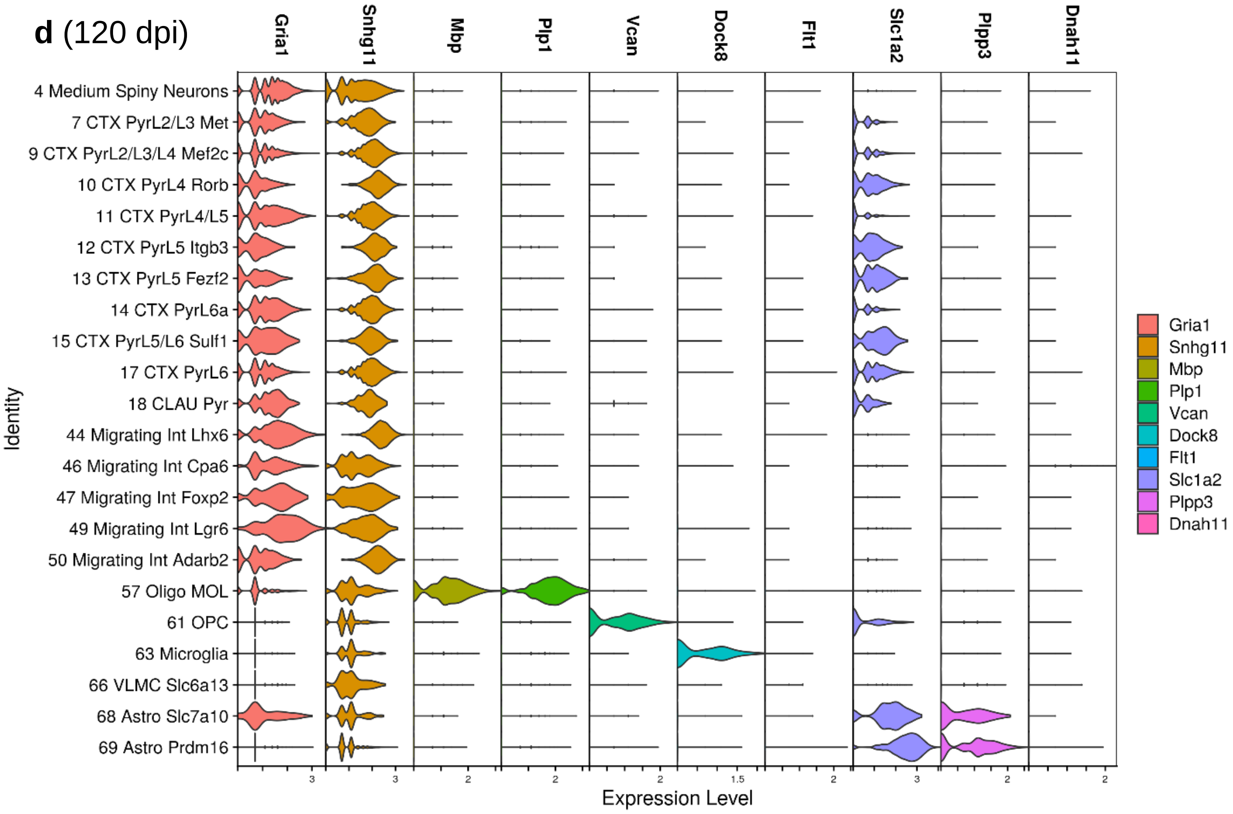
|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| **66** | VLMC Slc6a13 | VLMC | 333 | 262 | 178 | 210 | 242 |
| **68** | Astro Slc7a10 | Astrocytes | 122 | 119 | 104 | 138 | 130 |
| **69** | Astro Prdm16 | Astrocytes | 2781 | 1917 | 2127 | 2134 | 3103 |
| **72** | Ependyma | Ependymal | N/A | 157 | 112 | N/A | 120 |

***Table 5: Number of cells identified per time point dataset across the 26 clusters.*** *A similar number of cells were identified in each cluster across the different time points, highlighting the advantage and reproducibility of the label transfer approach. Cluster numbers are included as a reference system and will correspond to the same cluster name across the chapter and discussion. N/A values represent clusters that did not pass filtering as they had less than 100 cells, but not necessarily zero.*

We ensured that label transfer was successful by visualising the expression of a set of marker genes in each cluster. We selected a set of genes from the available literature, including the adolescent mouse brain atlas from the Linnarsson Lab and the SPLiT-seq manuscript (Rosenberg et al., 2018; Zeisel et al., 2018). The expression of these marker genes corroborated cluster identities.




b (40 dpi)
Gria1
Snhg11
Mbp
Plp1
Vcan
Dock8
Flt1
Slc1a2
Plpp3
Dnah11
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
48 Migrating Int Pbx3
49 Migrating Int Lgr6
50 Migrating Int Adarb2
56 Oligo MFOL1
57 Oligo MOL
61 OPC
64 Endothelia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma
Identity
2
3
2
3
2
1.5
2
3
2
1.5
Expression Level
Gria1
Snhg11
Mbp
Plp1
Vcan
Dock8
Flt1
Slc1a2
Plpp3
Dnah11
c (80 dpi)
Gria1
Snhg11
Mbp
Plp1
Vcan
Dock8
Flt1
Slc1a2
Plpp3
Dnah11
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
12 CTX PyrL5 Itgb3
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
50 Migrating Int Adarb2
57 Oligo MOL
61 OPC
64 Endothelia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma
Identity
2
3
2
2
1.5
1.5
2
3
2
2
Expression Level
Gria1
Snhg11
Mbp
Plp1
Vcan
Dock8
Flt1
Slc1a2
Plpp3
Dnah11


d (120 dpi)
Gria1
Snhg11
Mbp
Plp1
Vcan
Dock8
Flt1
Slc1a2
Plpp3
Dnah11
Identity
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
12 CTX PyrL5 Itgb3
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
49 Migrating Int Lgr6
50 Migrating Int Adarb2
57 Oligo MOL
61 OPC
63 Microglia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
3
3
2
2
2
1.5
2
3
2
2
Expression Level
Gria1
Snhg11
Mbp
Plp1
Vcan
Dock8
Flt1
Slc1a2
Plpp3
Dnah11

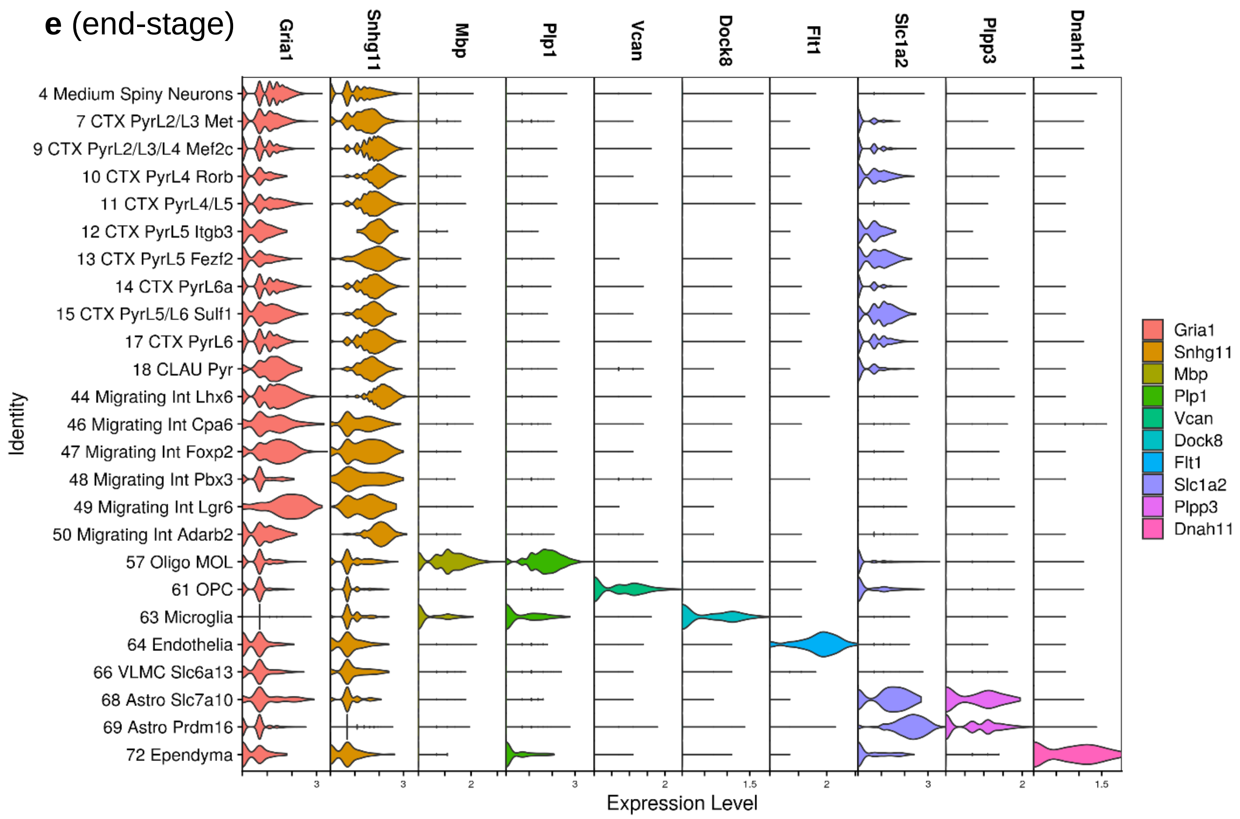


***Figure 4.8: The expression of known marker genes corroborates cluster identities.*** *Violin plots of the expression of known marker genes across the identified clusters in 5 time points. Marker genes used: Gria1 and Snhg11 for neurons; Mbp and Plp1 for oligodendrocytes; Vcan for oligodendrocyte precursor cells; Dock8 for microglia; Flt1 for endothelia; Slc1a2 and Plpp3 for astrocytes; Dnah11 for Ependyma. The expression level corresponds to the sctransform normalised expression values of the dataset.*

## Cell-type proportions – Selective toxicity

Having annotated the datasets, we then quantified differences between cell-type proportions to investigate the effect of prion disease on the abundance of different cell types and investigate selective cell toxicity. We grouped the cell clusters in 10 groups (migrating interneurons, cortical neurons, medium spiny neurons, astrocytes, OPCs, oligodendrocytes, VLMCs, ependymal, immune, and vascular cells) and quantified the relative proportions in all experimental groups across the 5 time points. A visual comparison between the CD1 and RML stacked bar plots suggests an increase of astrocytic populations at the 120 dpi time point and a decrease of cortical neurons and migrating interneurons, and an increase of medium spiny neurons and immune populations at the end-stage (Figure 4.9).

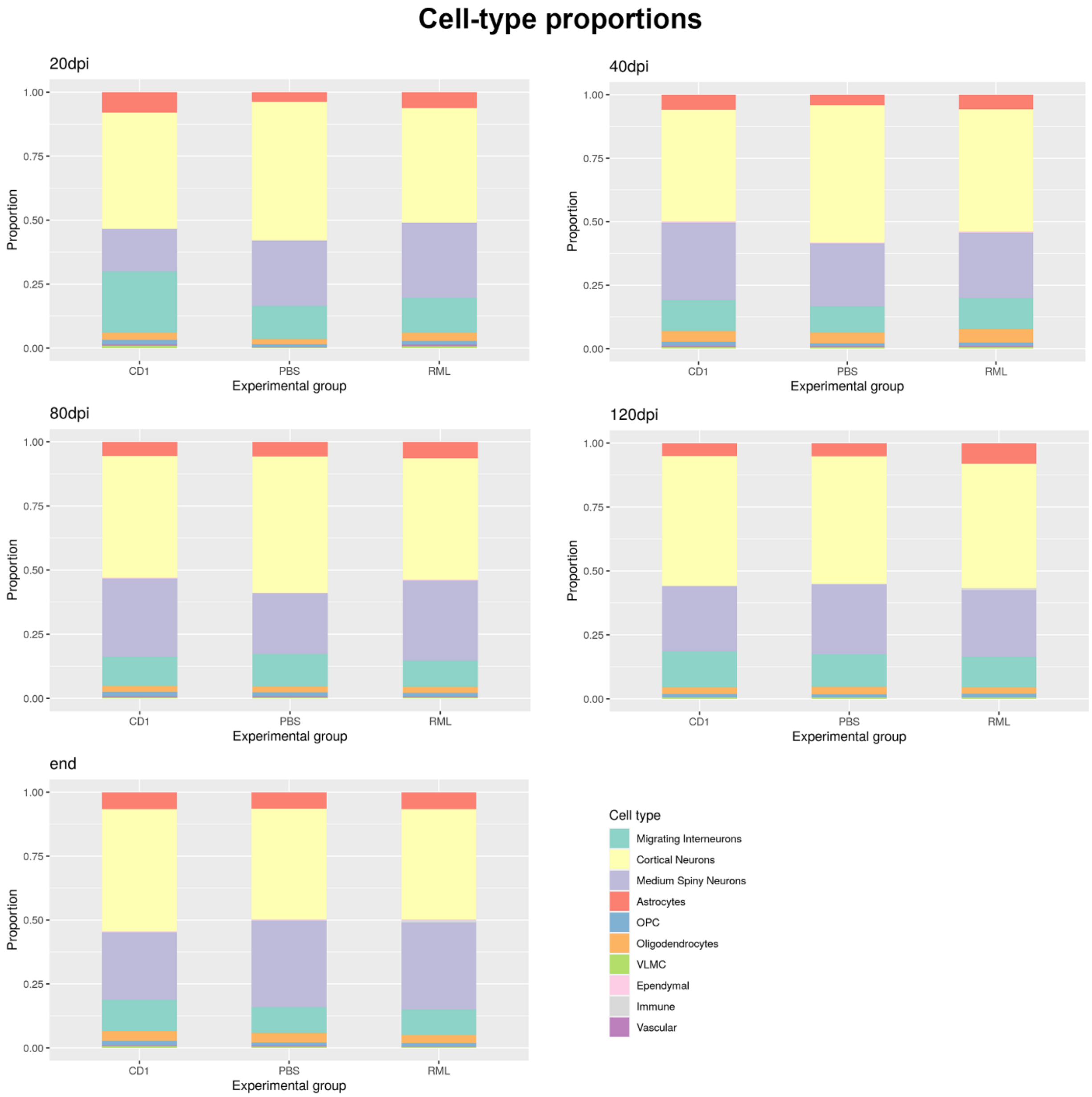


***Figure 4.9: Cell-type proportions of the three experimental groups across the 5 time points.*** *A visual examination of the plots suggests an increase of astrocytic populations at the 120 dpi time point and a decrease of cortical neurons and migrating interneurons, and an increase of medium spiny neurons and immune populations at the end-stage when CD1 and RML groups are compared. Proportions are calculated based on the total number of cells in each time point separately. The proportions of the PBS group are more variable, probably owing to the smaller number of total cells, and appear to be visually different in the*

*first three time points, while it mirrors more closely the CD1 group for the 120 dpi time point, and the RML group at the end-stage. Ependymal, vascular and VLMC cells are identified in minute proportions and are not clearly visible on the plots.*

While a simple calculation of the cell-type proportions can give a rough estimation of the population shifts, the results are inherently biased due to sampling differences and differences in the total number of cells identified, as it is highlighted by the variation that the PBS group exhibits. To begin to investigate the selective toxicity question, a more statistically rigorous approach was considered to be necessary. Although a plethora of bioinformatics packages have increasingly become available, we decided to employ a simple tool that is based on permutation testing and evaluates the null hypothesis that the difference in cell proportions for each cluster between the two conditions is a consequence of random sampling a subset of cells in each condition. This method, called scProportionTest, produced interesting results, especially for the first and last time points. At 20 dpi the analysis suggested a decrease of migrating interneurons, VLMCs, astrocytes, and OPCs and an increase in medium spiny neurons; at 40 dpi there was a small increase in the numbers of oligodendrocytes; at 80 dpi there were no changes reported; at 120 dpi an increase of astrocytes and microglia is observed; finally, at the end-stage, we observed a decrease in VLMCs and vascular cells, OPCs and mature oligodendrocytes, and migrating interneurons, while there was a small increase of medium spiny neurons and a considerable increase of microglia populations (Figure 4.10). Interestingly, some aspects of the 20 dpi plot are also present in the end-stage plot, i.e., the decrease in migrating interneurons, VLMCs, OPCs and the decrease in the medium spiny neurons. In addition, a more abundant immune population is suggested by the last two time points. Some of those findings are in accordance with our current understanding of prion disease pathophysiology (such as an increase of immune populations during the later stages of the disease), while others come in strong contrast (such as the increase of neuronal populations at the end-stage).

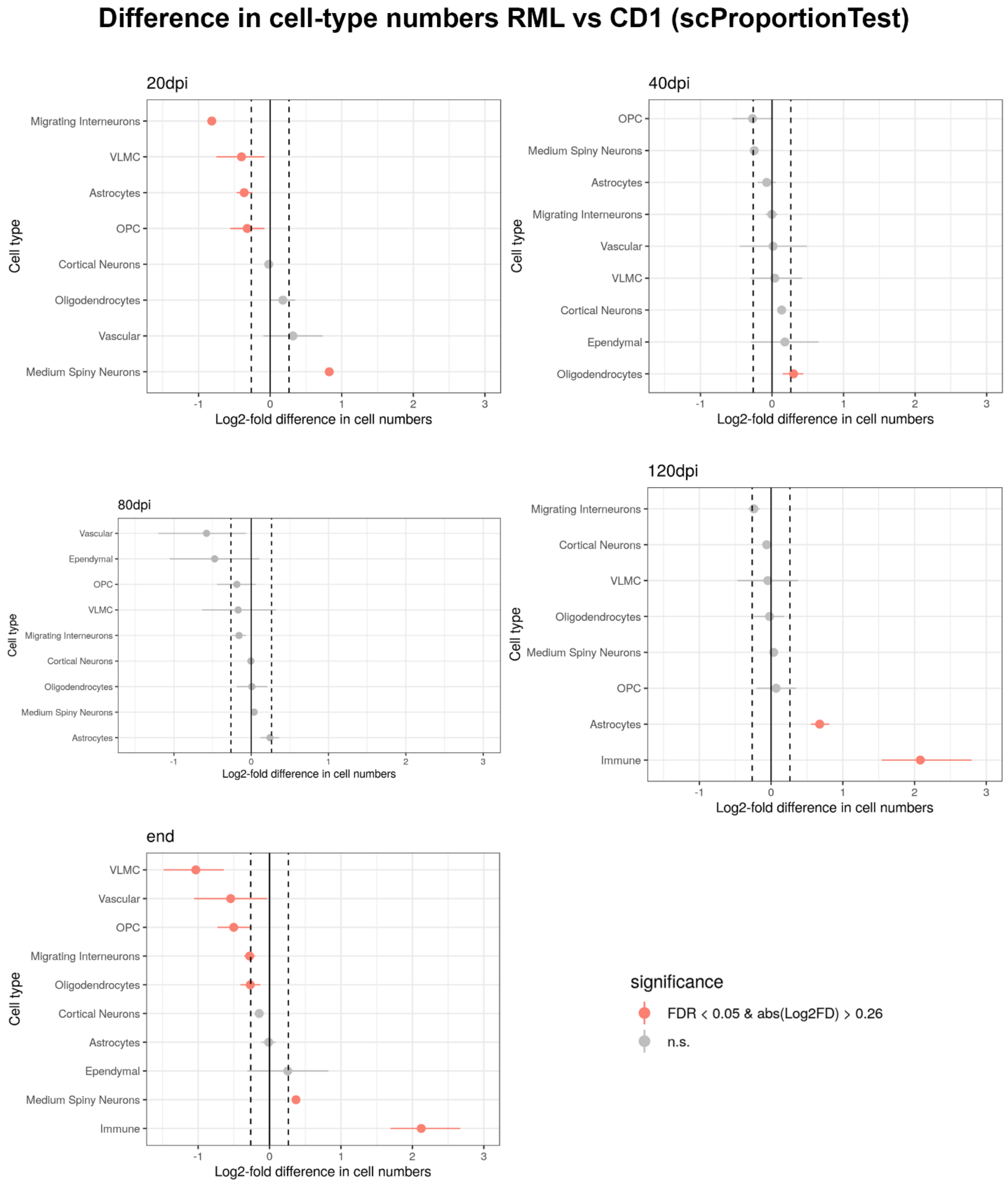


***Figure 4.10: A permutation test identified differences in cell-type abundance, which were more pronounced at the first and last time points.*** *At 20 dpi the analysis suggested a decrease of migrating interneurons, VLMCs, astrocytes, and OPCs and an increase in medium spiny neurons; at 40 dpi there was a small increase in the numbers of oligodendrocytes; at 80 dpi there were no changes reported; at 120 dpi*

*an increase of astrocytes and microglia is observed; finally, at the end-stage, we observed a decrease in VLMCs and vascular cells, OPCs and mature oligodendrocytes, and migrating interneurons, while there was a small increase of medium spiny neurons and a considerable increase of microglial populations. Dashed lines indicate the cut-off values for the difference in the cell numbers to be acceptable; the threshold has been set to an absolute log2-fold difference greater than 0.26, which corresponds to a 20% difference in cell numbers. Dots represent the mean of the calculated cell number differences. Vertical lines represent the 95% confidence interval of the mean. Coloured dots represent acceptable cell number differences, where both the magnitude of change criterion has been met and the calculated false discovery rate is lower than 0.05. Grey dots represent non-significant results. N.s: non-significant; FDR: False discovery rate; abs: absolute. For a detailed explanation of the methods used refer to the methodology section 1.3.4.*

The existence of the additional PBS inoculated group allowed us to assess the efficacy of the permutation test approach by comparing the cell numbers between the two control groups (CD1 vs PBS). The PBS-inoculated mouse samples were used as the reference baseline and the comparison identified an increase in multiple populations at the 20 dpi time point (oligodendrocytes, OPCs, migrating interneurons, vascular and VLMCs, and astrocytes) and a decrease in the number of medium spiny neurons; a small decrease in cortical neurons and a small increase in medium spiny neurons, OPCs, astrocytes and ependymal cells at 40 dpi; an increase in medium spiny neurons at 80 dpi; no differences at 120 dpi; and a decrease in ependymal cells and medium spiny neurons, and an increase in migrating interneurons and OPCs at the end-stage (Figure 4.11). The shift in population abundance during the first time point could be relevant to the introduction of external brain homogenate eliciting an immune response, as the data corroborates the existence of neuroinflammation (higher abundance of OPCs, oligodendrocytes and astrocytes; differences in vascular and leptomeningeal cells), while only small changes are suggested for the other four time points.

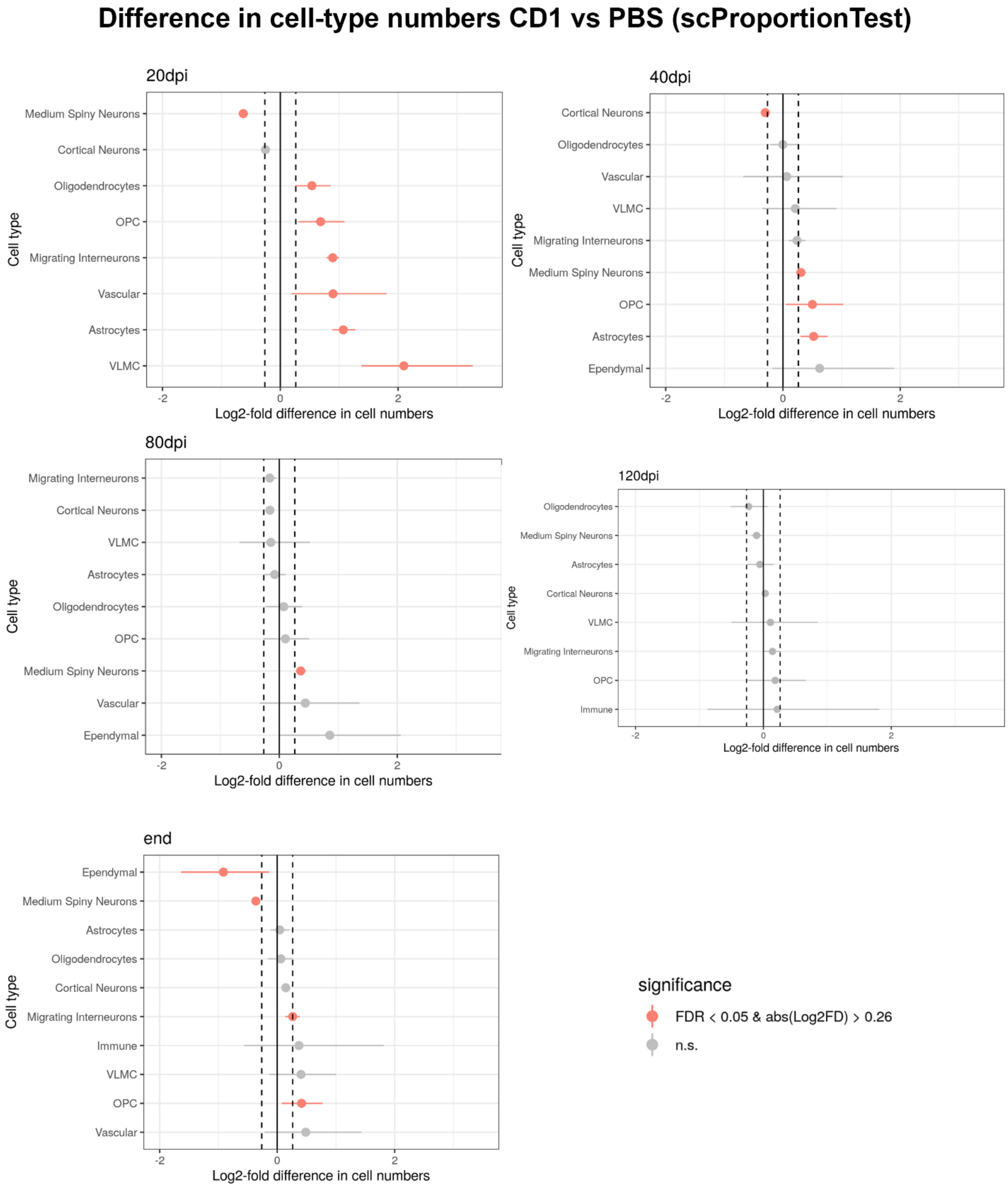


***Figure 4.11: A permutation test between the two control groups identified changes in cell-type abundance, especially at the first time point.*** *The comparison between CD1 and PBS control groups identified an increase in multiple populations at the 20 dpi time point (oligodendrocytes, OPCs, migrating interneurons, vascular and VLMCs, and astrocytes) and a decrease in the number of medium spiny*

*neurons; a small decrease in cortical neurons and a small increase in medium spiny neurons, OPCs, astrocytes and ependymal cells at 40 dpi; an increase in medium spiny neurons at 80 dpi; no differences at 120 dpi; and a decrease in ependymal cells and medium spiny neurons, and an increase in migrating interneurons and OPCs at the end-stage. The PBS group is used as the baseline reference. Plot elements are the same as described in Figure 4.10.*

We then focused on the most diverse cell type and the most relevant to neurodegenerative diseases, neurons. We subset our datasets to include only the migrating interneurons, cortical neurons and medium spiny neurons and performed the same permutation tests for each neuronal cluster separately. We observed that most of the neuronal populations seem to decrease in numbers across all time points, while the medium spiny neurons were more abundant in the first and last time points, as expected by the results of the previous plots. In accordance with the previous findings, most of the differences in neuronal populations are found at the first and last time points, 20 dpi and end-stage (Figure 4.12). Clusters of cortical neurons 10 and 13 exhibit a fluctuating pattern, where they are found to be more abundant in some time points and less in others. Migrating interneurons cluster 46 is shown to be less abundant in all time points, except from 80 dpi. Migrating interneurons cluster 50 and claustrum neurons of cluster 18 are less abundant during the first and last time points.

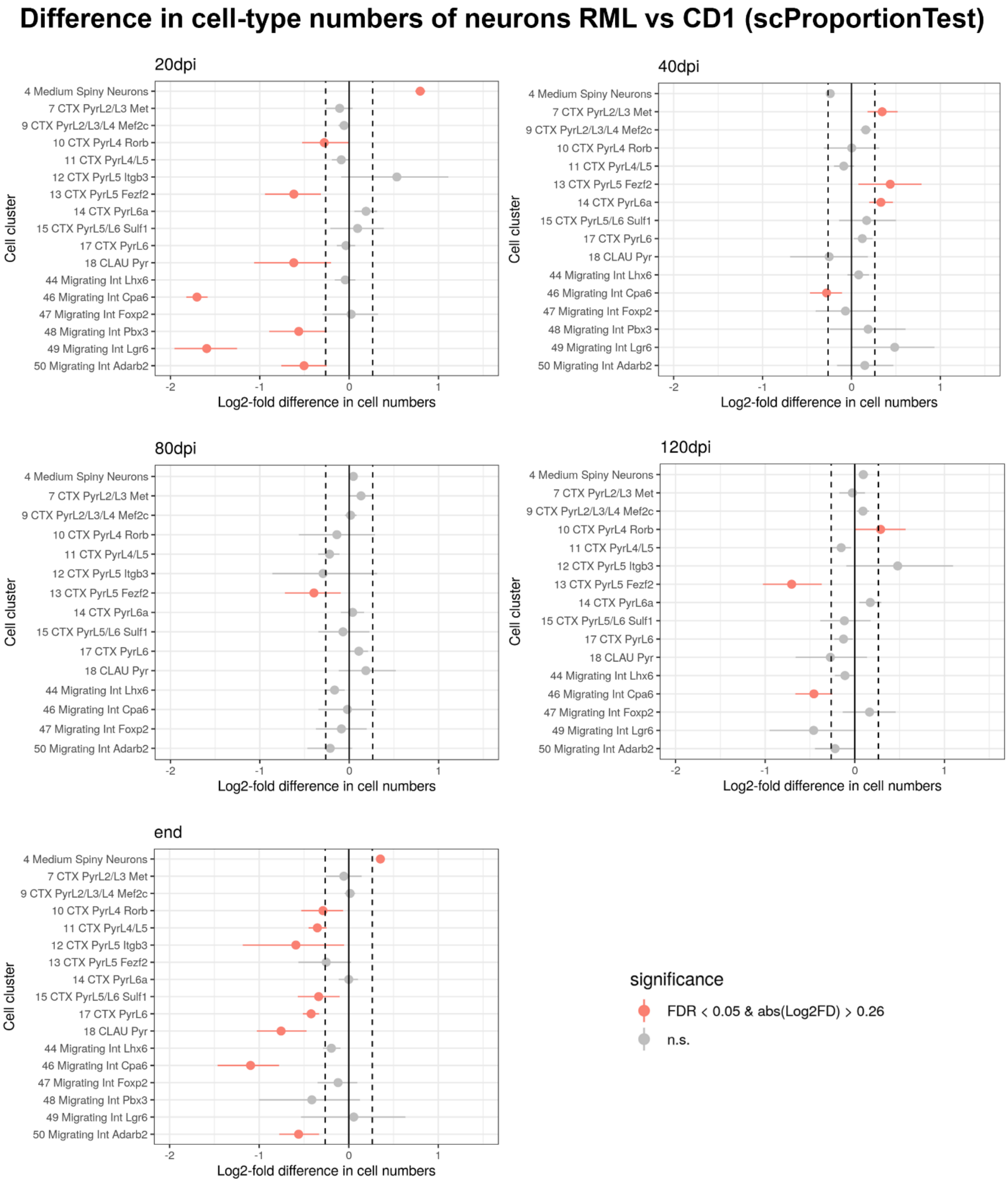


***Figure 4.12: A permutation test of the neuronal populations suggests a reduction of cell numbers for most of the neuronal clusters, which is more pronounced at the first and last time points.*** *Clusters of cortical neurons 10 and 13 exhibit a fluctuating pattern, where they are found to be more abundant in some time points and less in others. Migrating interneurons cluster 46 is shown to be less abundant in all*

*time points, except from 80 dpi. Migrating interneurons cluster 50 and claustrum neurons of cluster 18 are less abundant during the first and last time points. Plot elements are the same as described in Figure 4.10.*

Overall, while the permutation tests suggest that there are differences in the abundance of some cell types, these are of small magnitude and with large confidence intervals. Some of the evidence is in accordance with known pathophysiological changes in prion diseases, mainly the increase of immune and astrocyte populations during the later stages of the disease, while we could not observe consistent differences in the numbers of different clusters of neurons, even though a trend towards a lower abundance of specific populations can be identified especially at the 20 dpi and end-stage. No significant claims regarding selective toxicity can be made based on our findings.

**Differential gene expression analysis**

We performed a differential gene expression analysis to identify transcriptomic differences between cell clusters and investigate the fluctuation of the transcriptomic landscape that follows disease progression.

We first started by comparing the two sets of controls, each cluster of the CD1 group with the same cluster of the PBS group, to identify differences in gene expression that can be attributed to technical noise or are not specific to prion infection. From the 331 genes identified with an adjusted p-value less than 0.05, the vast majority were only found to be differentially expressed (DE) in one cluster of one time point and were therefore not excluded from the analysis (Supplementary Figure 7 and External Supplementary Table 1). However, there was a small set of genes that were found to be DE in more clusters and across many time points, their lack of specificity indicating that these genes were possible artefacts of the methodology. Setting a threshold of more than 5 occurrences allowed us to identify a set of 7 DE genes that were flagged for removal *(Calm1, Cdk8, Cmss1, Malat1, mt-Rnr1, mt-Rnr2,* and *Rn18s*). Their number of occurrences deviated substantially from 1 which was the case for the majority of genes, with Rn18s being identified 37 individual times (*Calm1*: 6 times, *mt-Rnr1* and *mt-Rnr2*: 10 times, *Malat1*: 14 times, *Cmss1*: 23 times, *Cdk8*: 31 times). The observation that most of those genes encode ribosomal RNAs or protein (*Cmss1, Rn18s*), are of mitochondrial origin (*mt-Rnr1, mt-Rnr2*), or are highly expressed in brain tissue (*Malat1, Calm1*) strengthens our hypothesis that they are indeed methodological artefacts introduced due to their high abundance or due to library contamination from mitochondrial RNA.

We then proceeded to compare the RML-prion brain homogenate inoculated group to the uninfected CD1 brain homogenate group to identify transcriptomic differences that are specific to prion disease. We identified approximately 8 thousand differentially expressed genes (DEGs) using the default settings of the FindMarkers function of Seurat (log2-fold change higher than 0.25, only testing genes that are expressed in at least 10% of cells in either group, using the Wilcoxon rank-sum test) (Table 6). These results were further filtered to keep DEGs with a Bonferroni-corrected p-value of less than 0.05. Finally, we removed the set of 7 spurious genes that were previously identified from the comparison between the two controls. This resulted in a total of 928 DEGs, most of those identified at the last time point, followed by the 120 and 20 dpi time points. A very low number of genes was identified for the 40 and 80 dpi time points. No bias was identified regarding the transcript lengths or chromosomes of the DEGs, and no outlier samples were found to drive the differences in gene expression (Supplementary Figure 8). External Supplementary Table 2 includes detailed information regarding all identified genes.

| Time point | DEGs before filtering | DEGs after p-value filtering | DEGs after removal of spurious transcripts |
|---|---|---|---|
| **20 dpi** | 820 | 127 | 60 |
| **40 dpi** | 979 | 22 | 6 |
| **80 dpi** | 1091 | 12 | 5 |
| **120 dpi** | 1876 | 228 | 174 |
| **end-stage** | 3260 | 758 | 683 |
| **Total** | **8026** | **1147** | **928** |

***Table 6: A comparison between the RML and CD1 groups identified 928 differentially expressed genes in total after filtering.*** *Approximately 8 thousand genes were initially identified. These were filtered by adjusted p-value keeping the ones that do not pass the threshold of 0.05. The 7 flagged genes from the comparison between controls were then excluded from the analysis, resulting in the final number of 928 genes in total. Most of the DEGs are identified at the disease end-stage. The 120 and 20 dpi time points follow in numbers, while only a handful of genes were identified for the 40 and 80 dpi time points.*

Before moving forward with the rest of the analysis we decided to employ a different methodology to assess the robustness of the differential gene expression analysis. We transformed the data to generate pseudo-bulk transcript counts by summing the identified transcripts across cells of the same cluster and time point. This resulted in 5 datasets that resembled bulk sequencing experiments with 8 samples per experimental group (8 RML and 8 CD1 samples). We were then able to employ more traditional DE analysis tools, namely two pipelines, one based on the well-established DESeq2, and another based on the newer glmGamPoi. This approach allows us to include sample-to-sample heterogeneity information in the calculation of the relevant statistics, which is lost when using Seurat (Seurat and other single-cell specific tools treat each cell individually, regardless of the biological sample of origin). DESeq2 employs a different statistical test, the Wald test, for hypothesis testing when comparing the two groups, and the calculation of the false discovery rate (FDR) is done using the Benjamini–Hochberg procedure. glmGamPoi fits a Gamma-Poisson Generalised Linear Model on the data and employs quasi-likelihood ratio testing to identify differentially expressed genes. Overall, the use of a different methodology can contribute additional support regarding the validity of our results if the new data corroborates the previous findings.

Both pseudo-bulk methods identified approximately 5 thousand DE genes in total, highly exceeding the number of genes identified by Seurat, even though pseudo-bulk approaches are considered more conservative. Reassuringly, the pattern of differential gene expression did follow previous findings, i.e., most of the DE genes were identified at the last two time points, with the end-stage having the highest number, then a smaller set of genes was reported to be DE at 20 dpi, while only a handful of genes were identified at 40 and 80 dpi (Table 7).

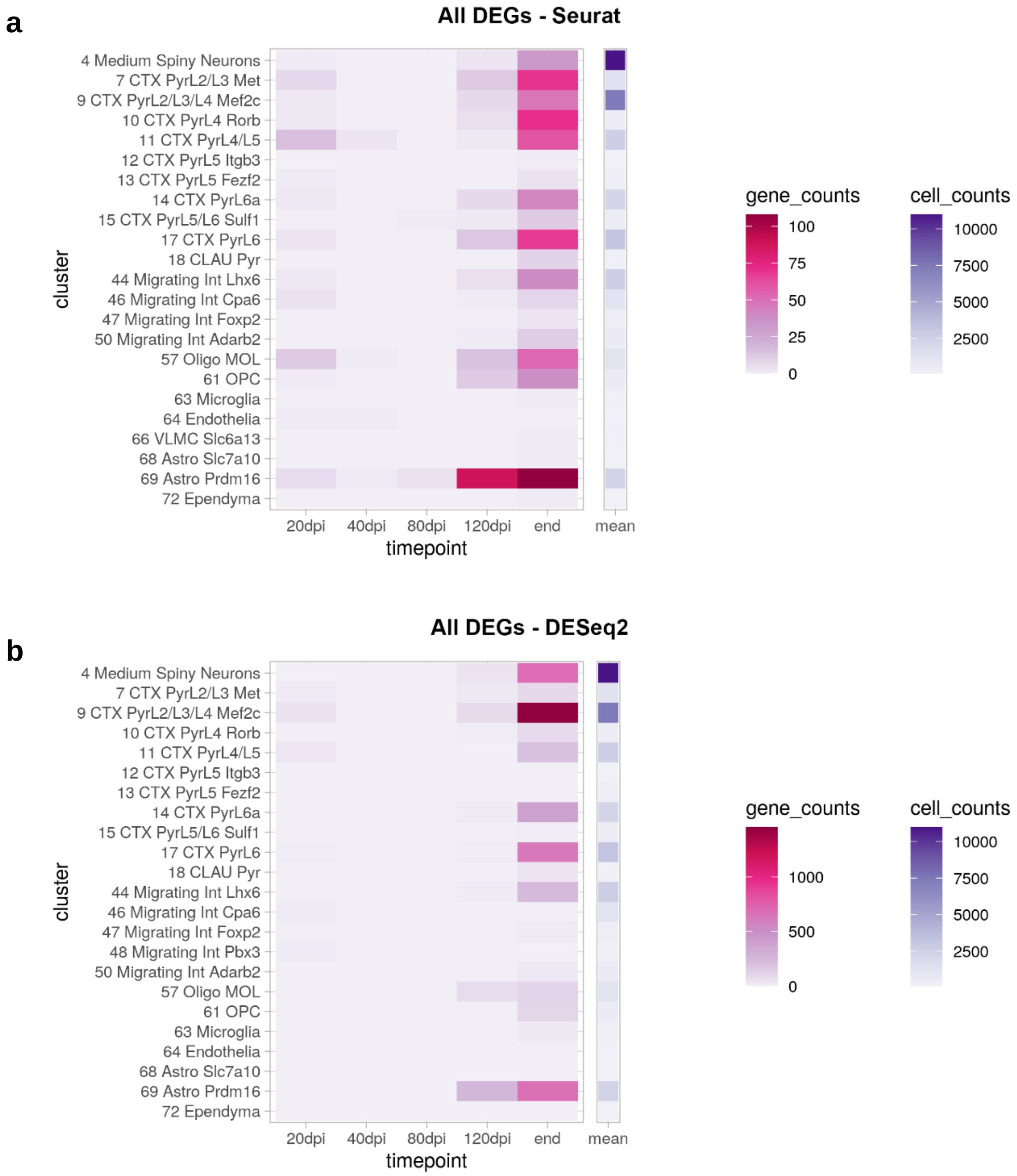
| Time point | DEGs after p-val. filtering (DESeq2) | DEGs after removal of spurious transcripts (DESeq2) | DEGs after p-val. filtering (glmGamPoi) | DEGs after removal of spurious transcripts (glmGamPoi) |
|---|---|---|---|---|
| **20 dpi** | 148 | 137 | 98 | 73 |
| **40 dpi** | 2 | 2 | 3 | 3 |
| **80 dpi** | 1 | 1 | 3 | 2 |
| **120 dpi** | 598 | 560 | 359 | 316 |
| **end-stage** | 4870 | 4811 | 4582 | 4517 |
| **Total** | **5619** | **5511** | **5045** | **4911** |

***Table 7: A comparison between the RML and CD1 groups identified more than 5 thousand genes in total when using the pseudo-bulk approach.*** *The number of DE genes reported by DESeq2 and glmGamPoi using a pseudo-bulk approach follows the same trend as the ones reported by Seurat: most*

*DE genes were identified at the last two time points, while a smaller number was reported for the 20 dpi time point. Just one or two genes were reported for the 40 and 80 dpi time points. The p-value filtering criterion was a BH-adjusted p-value of less than 0.05. Spurious transcripts refer to the set of 7 genes identified from the comparison of the two control groups.*

To visualise the number of DEGs across all time points we plotted heatmaps for each of the 3 DE analysis methods (Figure 4.13). The plots of the pseudo-bulk approaches (Figure 4.13b and Figure 4.13c) are more similar to each other, with most identified DE genes belonging to the same clusters. Clusters of cortical neurons 9, 14 and 17 show the highest number of DEGs, cluster 4 of medium spiny neurons is also shown to be dysregulated, astrocytes and to a lesser degree oligodendrocytes show a high number of DEGs, especially in the last two time points. When we compare the pseudo-bulk methods with Seurat, some overall trends seem to be characteristic for all three plots. Clusters of cortical neurons 9, 14, and 17 are shown to have a high number of DEGs, while Seurat also identified the additional clusters 7, 10, 11 and 14 with a high number of DEGs at the end-stage. Perturbations in astrocytes, mature oligodendrocytes, and to a lesser degree oligodendrocyte precursor cells are also commonly identified, especially at the end-stage.

Focusing on the 20 dpi time point only (so that the scale is shorter and small differences are more obvious), the 20 dpi signature of astrocytes and oligodendrocytes identified by Seurat is not identified when using the pseudo-bulk methods (Figure 4.13d). In contrast, the 20 dpi signature of cortical neurons does seem to be concordant, which is especially evident for clusters 7 and 11. Interestingly, cluster 9 is identified to have the highest number of DEGs by DESeq2 and a high number by glmGamPoi, however, the signature is not as pronounced when using Seurat for the analysis. Results are more ambiguous regarding the three clusters of migrating interneurons, where Seurat reports a low number of genes, while DESeq2 identified more DEGs for clusters 46 and 48, and glmGamPoi only for cluster 44.


a
All DEGs - Seurat
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
12 CTX PyrL5 Itgb3
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
50 Migrating Int Adarb2
57 Oligo MOL
61 OPC
63 Microglia
64 Endothelia
66 VLMC Slc6a13
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma
cluster
20dpi
40dpi
80dpi
120dpi
end
mean
timepoint
gene_counts
100
75
50
25
0
cell_counts
10000
7500
5000
2500
b
All DEGs - DESeq2
4 Medium Spiny Neurons
7 CTX PyrL2/L3 Met
9 CTX PyrL2/L3/L4 Mef2c
10 CTX PyrL4 Rorb
11 CTX PyrL4/L5
12 CTX PyrL5 Itgb3
13 CTX PyrL5 Fezf2
14 CTX PyrL6a
15 CTX PyrL5/L6 Sulf1
17 CTX PyrL6
18 CLAU Pyr
44 Migrating Int Lhx6
46 Migrating Int Cpa6
47 Migrating Int Foxp2
48 Migrating Int Pbx3
50 Migrating Int Adarb2
57 Oligo MOL
61 OPC
63 Microglia
64 Endothelia
68 Astro Slc7a10
69 Astro Prdm16
72 Ependyma
cluster
20dpi
40dpi
80dpi
120dpi
end
mean
timepoint
gene_counts
1000
500
0
cell_counts
10000
7500
5000
2500

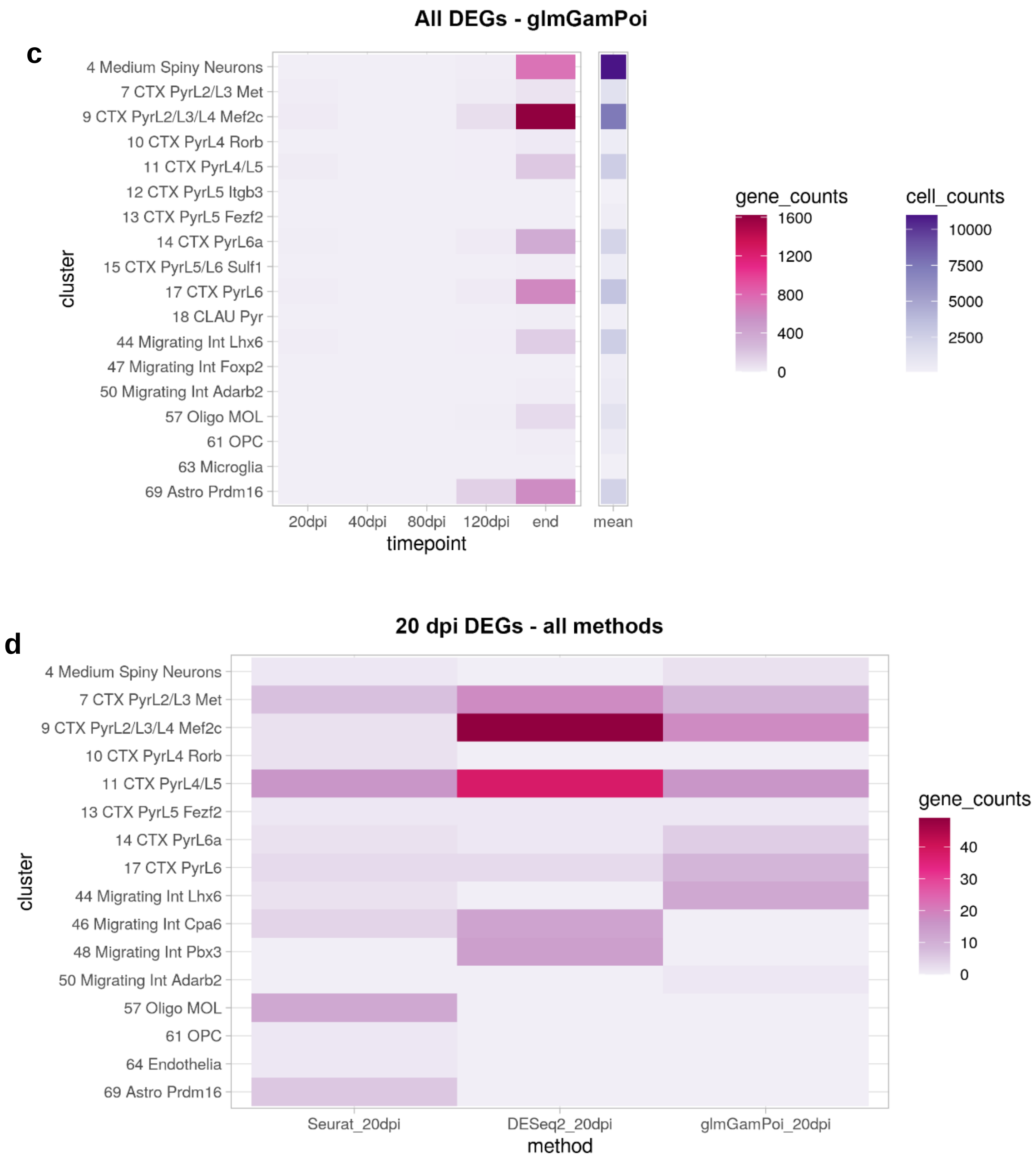


***Figure 4.13: Seurat and pseudo-bulk approaches based on DESeq2 and glmGamPoi identify the same patterns of gene expression across the 5 time points. (a)** Heatmap of the number of DEGs identified in each time point (x-axis) and cluster (y-axis) when using Seurat for the DE analysis. **(b)** Heatmap of the number of DEGs identified in each time point and cluster when using pseudo-bulk data with DESeq2. **(c)** Heatmap of the number of DEGs identified in each time point and cluster when using pseudo-bulk data with glmGamPoi. The separate column on the right of each heatmap shows the number of cells that each cluster comprises. A visual comparison of the number of DEGs and cells of a cluster suggests that clusters with higher numbers of cells also tend to have higher numbers of DEGs. The counts of differentially expressed genes and the counts of the cells in each cluster use different scales, which are denoted using*

*two different colours.* ***(d)*** *Heatmap of the number of DEGs identified using the 3 different methods (x-axis), only for the 20 dpi time point. The information of this plot is included in plots* ***a, b,*** *and* ***c****, however, it is shown here using a different scale to allow easier visual comparison.*

Figure 4.13 also includes a small heatmap of the number of cells in each cluster. A visual comparison of the number of DEGs and cells in a cluster suggests that clusters with higher numbers of cells are also associated with a higher number of DEGs. This trend is especially evident when considering the pseudo-bulk methods, while it is not as pronounced for clusters 7, 10, 17, 57, and 69 when using Seurat. A further investigation of this relationship between DEGs and numbers of cells also confirmed a positive correlation between the two variables with correlation coefficients being 0.28 for Seurat, 0.78 for DESeq2 and 0.78 for glmGamPoi (Supplementary Figure 9). This positive correlation is to be expected since having a larger sample allows the DE genes to pass the statistical thresholds and be included in the final results. It also highlights the fact that low numbers of DE genes in clusters with low numbers of cells can be attributed to shallower sampling and not a lack of differential expression. In our case, we are mostly concerned with tracking the dynamic changes of DE in the same clusters across different time points (which have similar numbers of cells as we have previously demonstrated). However, we do need to point out that we cannot make strong claims regarding the lack of DE in clusters with low numbers of cells.

We then investigated the concordance between the genes identified by Seurat and the pseudobulk methods and found that the pseudo-bulk methods agreed with the results of Seurat more at the last two time points, while the agreement was lower for the 20 dpi time point (Figure 4.14). We did not consider the 40 and 80 dpi time points, as all three methods identified only a handful of genes for those. This higher replication by pseudo-bulk methods of DEGs identified by Seurat could be attributed to the higher numbers of DEGs identified by both DESeq2 and glmGamPoi at 120 dpi and the end-stage, which were an order of magnitude more than the DEGs identified by Seurat.

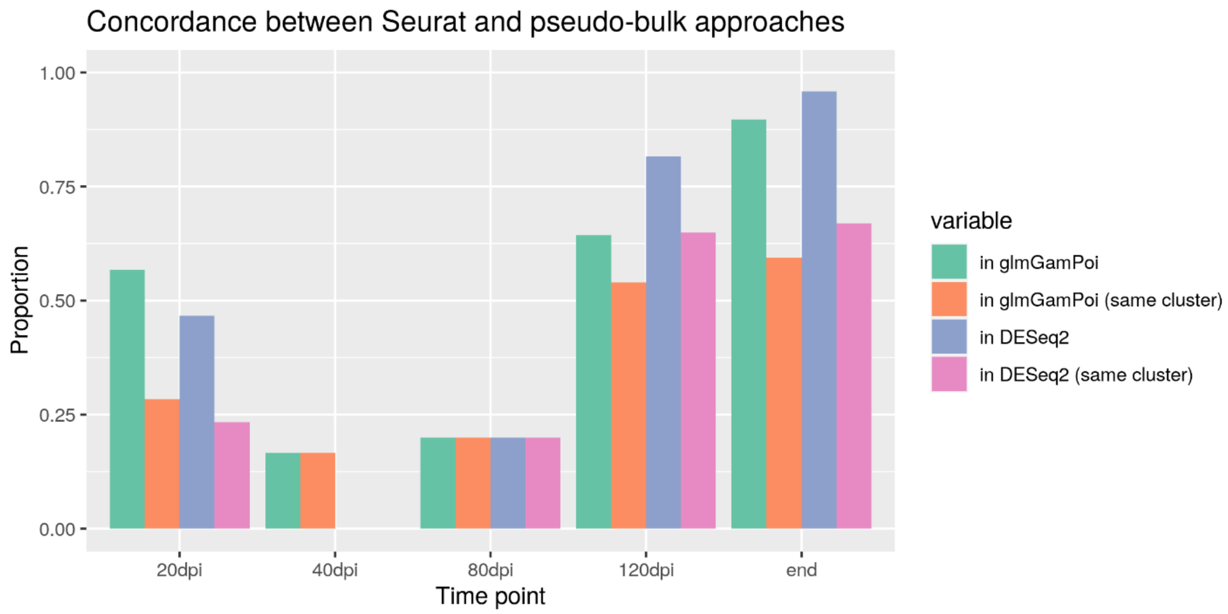


***Figure 4.14: Pseudo-bulk analysis methods show higher concordance with Seurat at the last two time points.*** *The plot shows the proportion of the DEGs identified by Seurat that were also identified by glmGamPoi or DESeq2 in the same time point, or in the same time point and same cluster. glmGamPoi shows higher concordance with Seurat results than DESeq2 for the 20 dpi time point, while this is reversed for the 120 dpi and end-stage time point, where DESeq2 identified more of the DEGs identified by Seurat. For both 40 and 80 dpi time points, only a handful of genes were identified by any of the three approaches, so these proportions are not as relevant.*

Overall, a purely single-cell-based DE analysis based on Seurat Wilcoxon rank-sum test produced results that were partially concordant with pseudo-bulk approaches based on the DESeq2 Wald test and the glmGamPoi quasi-likelihood ratio test. All approaches produced a similar pattern of differential gene expression where less than a hundred genes were identified at the first time point (20 dpi), then practically no differential expression was evident at 40 and 80 dpi, followed by an increase of DE genes identified at 120 dpi, which was then amplified at the end-stage. While all tools identified a similar number of genes for the 20 dpi, Seurat was shown to be more stringent at the last two time points, identifying an order of magnitude fewer genes than the pseudo-bulk tools. Regarding the genes identified, the pseudo-bulk approaches identified a concordant perturbation signature in various clusters of neurons, while the astrocytic and oligodendrocyte signatures were not identified at 20 dpi. Most of the genes identified by Seurat were concordant with the pseudo-bulk dataset at the last two time points. Taking all this information into account, we decided to proceed with the analysis focusing on the DE gene lists generated by Seurat, which was shown to be more stringent and more integrated into the analysis pipeline. However, the results from all these methods provided useful insight which will be invaluable in guiding us through further exploration of the data.

Having the DE gene lists from all 5 time points, we proceeded with identifying common gene dysregulation patterns during the disease (Figure 4.15). One-third of the DEGs at 20 dpi were unique to this time point. Interestingly, the other two-thirds of the DEGs, 25 in total, were found to be dysregulated at 20 dpi, and then showed up again in our analysis at the last two time points (indicated by black arrows in Figure 4.15, Table 8). Approximately half of those were found to be DE again at 120 dpi and the end-stage, while the other half were only identified again at the end-stage. Moving on to the 120 dpi time point, 31 of those genes were uniquely DE during this time point only, while 93 started being DE at 120 dpi and continued being DE at the end-stage. Finally, the end-stage had the highest number of uniquely DE genes, 267 DEGs that were not reported at any other time point. Lists of all possible intersections are provided in External Supplementary Table 5.

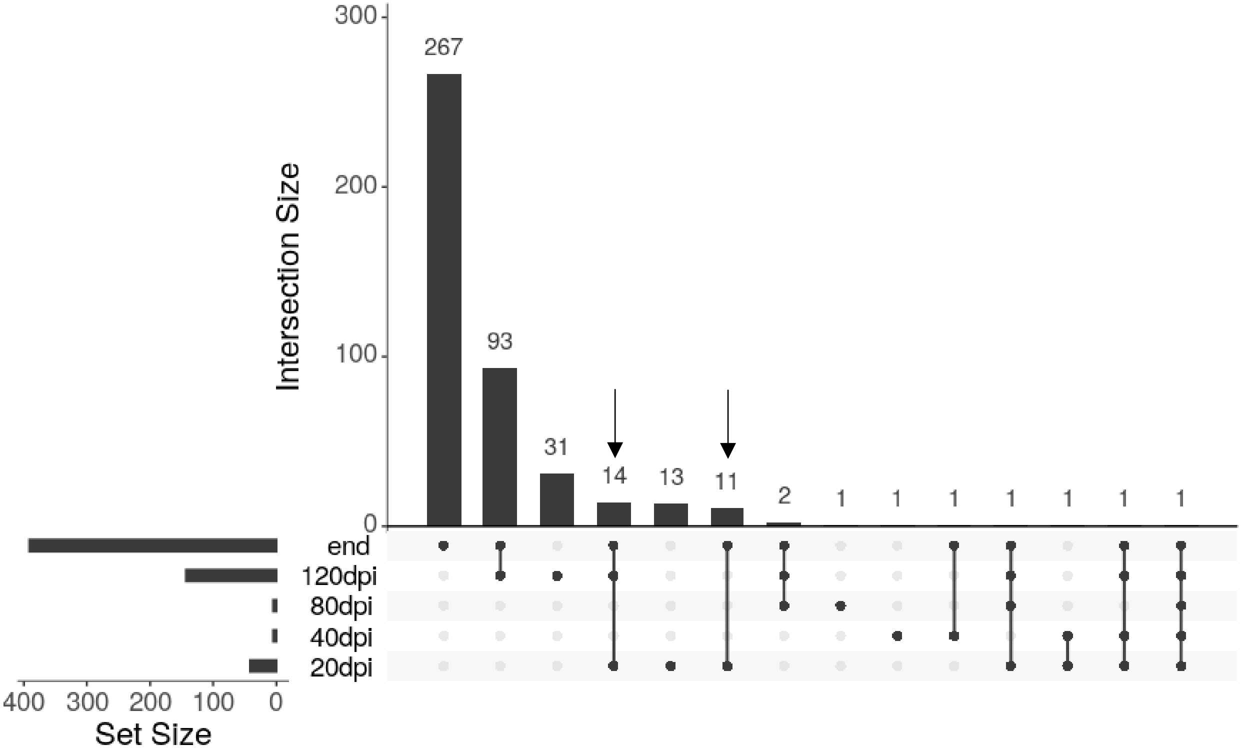


***Figure 4.15: The intersection of DEGs during the course of the disease reveals interesting patterns of gene expression perturbations.*** *The UpSet plot offers an overview of the intersections between DE gene sets across all time points. The horizontal bar plot on the left shows the number of DEGs at each time point. The matrix in the centre-bottom of the plot shows all possible combinations of unique gene sets with at least one gene. Each set intersection is defined by vertical black lines that connect two or more dots, each dot representing the time point defined on the y-axis of the matrix. The bar plot on top of the matrix shows the number of members of each intersection defined in the matrix below (for example the second vertical bar indicates that the intersection between the 120 dpi gene set and the end-stage gene set consists of 93 genes). A set of 25 genes were found to be DE at the early and then the late stages of the disease (indicated by the black arrows). 13, 31 and 267 genes were only found to be DE at 20 dpi, 120 dpi and the end-stage, respectively. 93 genes were found to be DE at 120 dpi and then continued being DE at the end-stage of the disease.*

Based on the gene set intersections and DE patterns, we separated the genes into two groups: the early/late set consisting of those genes that exhibit the DE pattern 20 dpi – 120 dpi – end-stage or 20 dpi – end-stage, and the late set consisting of genes DE at 120 dpi or end-stage or exhibiting the pattern 120 dpi – end-stage (Table 8 and Table 9). Importantly, these two groups do not intersect, i.e., genes that exhibit the early/late pattern will not be included in the set of late genes.

A consequent validation by RT-qPCR of the expression signature of 6 selected genes from the first set that were shown to be DE by Seurat and pseudo-bulk approaches (*Ndst4, Gphn, Pde10a, Abi3bp, Il31ra, Auts2*) confirmed a statistically significant differential expression in the early or late time points for 2 genes (*Abi3bp, Auts2*), while 4 genes could not be validated (Supplementary Figure 10). Further validation of 5 genes from the late set (*Apoe, Grin2a, Nrp1, Ptk2, Rph3a*) confirmed a statistically significant differential expression in the late time points of 2 genes (*Apoe, Grin2a*) (Supplementary Figure 11). In both cases the starting material was bulk brain nuclei suspension, so the effect of specific cell populations could have been diluted in the bulk material.

| Early/late DEGs | | | | |
|---|---|---|---|---|
| Abi3bp | Gphn | Lrrtm4 | Pde10a | Tafa1 |
| Adarb2 | Grm8 | Lsamp | Pdzrn4 | Tenm3 |
| Auts2 | Il31ra | Meg3 | Phactr1 | Tnik |
| Dlgap1 | Kcnb2 | Mgat4c | Prkg1 | Xylt1 |
| Ext1 | Kcnc2 | Ndst4 | Rora | Zfp804b |

***Table 8: A set of 25 genes were found to exhibit an early and late signature of differential expression.*** *The DE pattern of these genes was either 20 dpi – 120 dpi – end-stage or 20 dpi – end-stage. These genes are not included in the late set. Double underlined are the genes with expression patterns validated by real-time PCR analysis, while single underlined are the ones with expression patterns that could not be validated.*

| Late DEGs | | | | | |
|---|---|---|---|---|---|
| 4930488L21Rik | Clstn2 | Gabbr2 | Lrfn5 | Pde8b | Slc24a2 |
| 5031425E22Rik | Clu | Gabrb1 | Lrrc4c | Pdzrn3 | Slc24a3 |
| 9330162G02Rik | Cntnap2 | Gabrg3 | Lrrk2 | Penk | Slc2a13 |
| A230001M10Rik | Cntnap4 | Galntl6 | Lrrtm3 | Pex5l | Slc35f1 |
| A230057D06Rik | Cntnap5c | Garnl3 | Luzp2 | Phlpp1 | Slc8a1 |
| A330015K06Rik | Cobl | Gjc3 | Magi2 | Pid1 | Slco1c1 |
| Abca1 | Col19a1 | Gli2 | Maml2 | Pitpnc1 | Slit2 |
| Ablim1 | Crtac1 | Gm16168 | Maml3 | Pitpnm2 | Slit3 |
| Abr | Csgalnact1 | Gm26871 | Mapk4 | Pknox2 | Smarca2 |
| Actb | Csmd1 | Gm28905 | Mast3 | Plcb1 | Snap25 |
| Adcy2 | Csmd2 | Gm30382 | Mast4 | Plce1 | Sorbs1 |
| Adgrb3 | Cst3 | Gm3764 | Mbp | Plp1 | Sorbs2 |
| Adgrl3 | Ctnna2 | Gnao1 | Mdga2 | Plxdc2 | Sorcs2 |
| Afap1 | Ctnnd2 | Gng12 | Mef2c | Plxna4 | Sorcs3 |
| Aig1 | D430041D05Rik | Gpc5 | Meis2 | Ppm1l | Sox2ot |

| | | | | | |
|---|---|---|---|---|---|
| Ak5 | Dab1 | Gpc6 | Mertk | Ppme1 | Sox5 |
| Alk | Dcc | Gpm6a | Mgat5 | Ppp2r2b | Sox6 |
| Ankrd33b | Dclk1 | Gpm6b | Mical2 | Prex2 | Spock1 |
| Anks1b | Ddx5 | Gria4 | Mir9-3hg | Prickle1 | Spock3 |
| Ano4 | Dennd1a | Grid2 | Mir99ahg | Prkag2 | Srgap3 |
| Aopep | Dgkb | Grik3 | Mobp | Prkca | Srrm2 |
| Apc | Dgki | Grin2a | Msi2 | Prkcb | St18 |
| Apod | Dip2a | Grip1 | Mtcl1 | Psd3 | St6galnac3 |
| Apoe | Dlgap2 | Grm1 | Ncam1 | Ptk2 | St6galnac5 |
| Appl2 | Dlx1as | Grm3 | Nckap5 | Ptn | Stox2 |
| Arhgef4 | Dlx6os1 | Grm5 | Nebl | Ptprj | Stxbp6 |
| Arpp21 | Dmd | Hcn1 | Nedd4l | Ptprm | Syt1 |
| Asap1 | Dnajc6 | Hdac4 | Negr1 | Ptprt | Tenm4 |
| Asic2 | Dnm3 | Hdac9 | Neto1 | Qk | Thrb |
| Astn1 | Dock10 | Hecw1 | Nfia | R3hdm1 | Thsd7b |
| Astn2 | Dock4 | Hivep2 | Nhsl1 | R3hdm2 | Tmeff2 |
| Atg4a | Dpp10 | Homer1 | Nkain2 | Rapgef2 | Tmem108 |
| Atp1a2 | Dscam | Hs3st2 | Nlgn1 | Rarb | Tmem132d |
| Atp1b2 | Dscaml1 | Hs3st4 | Nlk | Rasal2 | Tmem178 |
| Atp2b2 | Dst | Hs6st3 | Nol4 | Rasgrf1 | Tmtc1 |
| Atp8a2 | Dtna | Hsp90aa1 | Nos1ap | Rbfox1 | Tmtc2 |
| Atrnl1 | Edil3 | Hspa12a | Npas3 | Rbms3 | Tnr |
| Atxn1 | Elavl2 | Htr2c | Npsr1 | Rfx3 | Tox |
| B3galt1 | Eml5 | Igfbp5 | Nrg1 | Rgs20 | Trf |
| Brinp3 | Enox1 | Igsf21 | Nrg3 | Rgs6 | Trhde |
| C4b | Enpp2 | Igsf9b | Nrp1 | Rgs7 | Trim9 |
| Cacna1a | Epha6 | Iqgap2 | Nrxn1 | Rgs9 | Trpm3 |
| Cacna1e | Ephb1 | Iqsec1 | Ntm | Rims1 | Trps1 |
| Cacna2d1 | Epn2 | Jazf1 | Ntrk2 | Rnf220 | Tshz2 |
| Cacna2d3 | Eps8 | Kcnab1 | Ntrk3 | Robo1 | Tspan5 |
| Cacnb2 | Erbb4 | Kcnd2 | Numb | Rock2 | Tspan7 |
| Cacng3 | Etl4 | Kcnd3 | Nwd2 | Rorb | Ugt8a |
| Cadm1 | Exph5 | Kcnh1 | Nxph1 | Rph3a | Unc13a |
| Cadm2 | Fam13c | Kcnh7 | Osbp2 | Rps6ka2 | Unc13c |
| Caln1 | Fam155a | Kcnip1 | Osbpl8 | Ryr2 | Unc5c |
| Camk1d | Fam189a1 | Kcnip4 | Otud7a | Scd2 | Unc5d |
| Camk2a | Fam20a | Kcnj3 | Oxr1 | Scube1 | Unc80 |
| Camta1 | Fars2 | Kcnj6 | Pacrg | Sema5a | Utrn |
| Cap2 | Fat3 | Kcnma1 | Pak5 | Sema6d | Vav3 |
| Car10 | Fgf12 | Kcnmb2 | Pard3 | Septin7 | Vmp1 |
| Cdc42bpa | Fgf13 | Kcnn2 | Pbx1 | Setbp1 | Vsnl1 |

| | | | | | |
|---|---|---|---|---|---|
| Cdh12 | Fgf14 | Kcnq3 | Pcdh11x | Sgcd | Wdr17 |
| Cdh13 | Fgfr2 | Kctd16 | Pcdh7 | Sgcz | Xist |
| Cdh18 | Fhod3 | Kirrel3 | Pcsk2 | Sgip1 | Ypel2 |
| Cdh20 | Fmn1 | Ldb2 | Pde1a | Shisa6 | Zbtb20 |
| Cdh4 | Fnbp1 | Ldlrad4 | Pde4a | Shisa9 | Zdhhc14 |
| Cdyl2 | Frmd4a | Lhfpl3 | Pde4b | Shtn1 | Zeb1 |
| Cemip | Frmd5 | Lingo1 | Pde4d | Sik2 | Zfp385b |
| Chrm2 | Frmpd4 | Lingo2 | Pde7b | Sipa1l1 | Zfp536 |
| Chsy3 | Fstl4 | Lrfn2 | Pde8a | Slc1a2 | Zfp804a |
| | | | | | Zswim6 |

***Table 9: A set of 391 genes were identified to be DE at the last two time points of our experiment.*** *These genes were DE at 120 dpi only, at the end-stage only, or both at 120 dpi and end-stage. Genes that belong to the early/late set are not included in this set. Double underlined are the genes with expression patterns validated by real-time PCR analysis, while single underlined are the ones with expression patterns that could not be validated.*

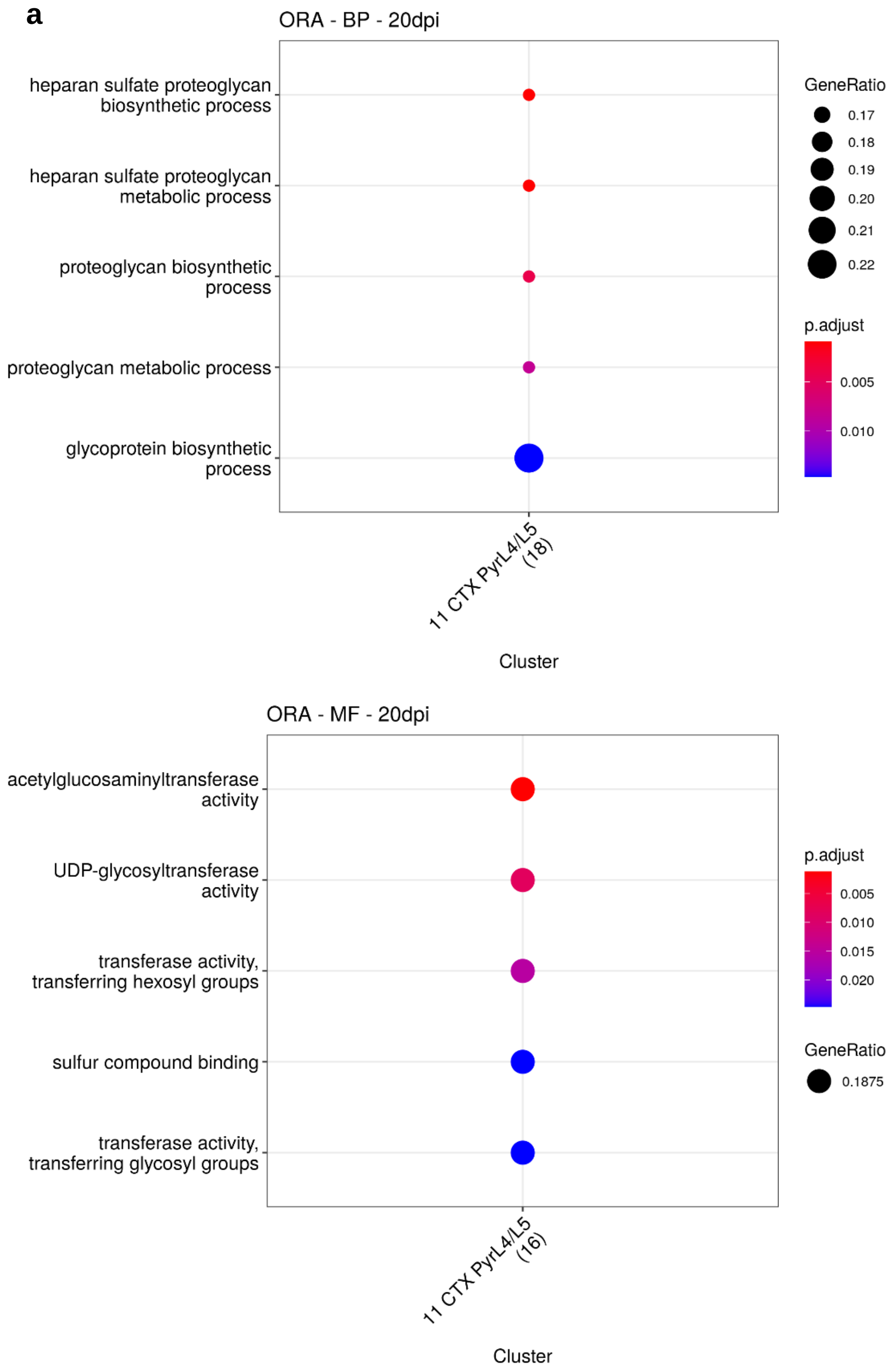
### Enrichment analyses

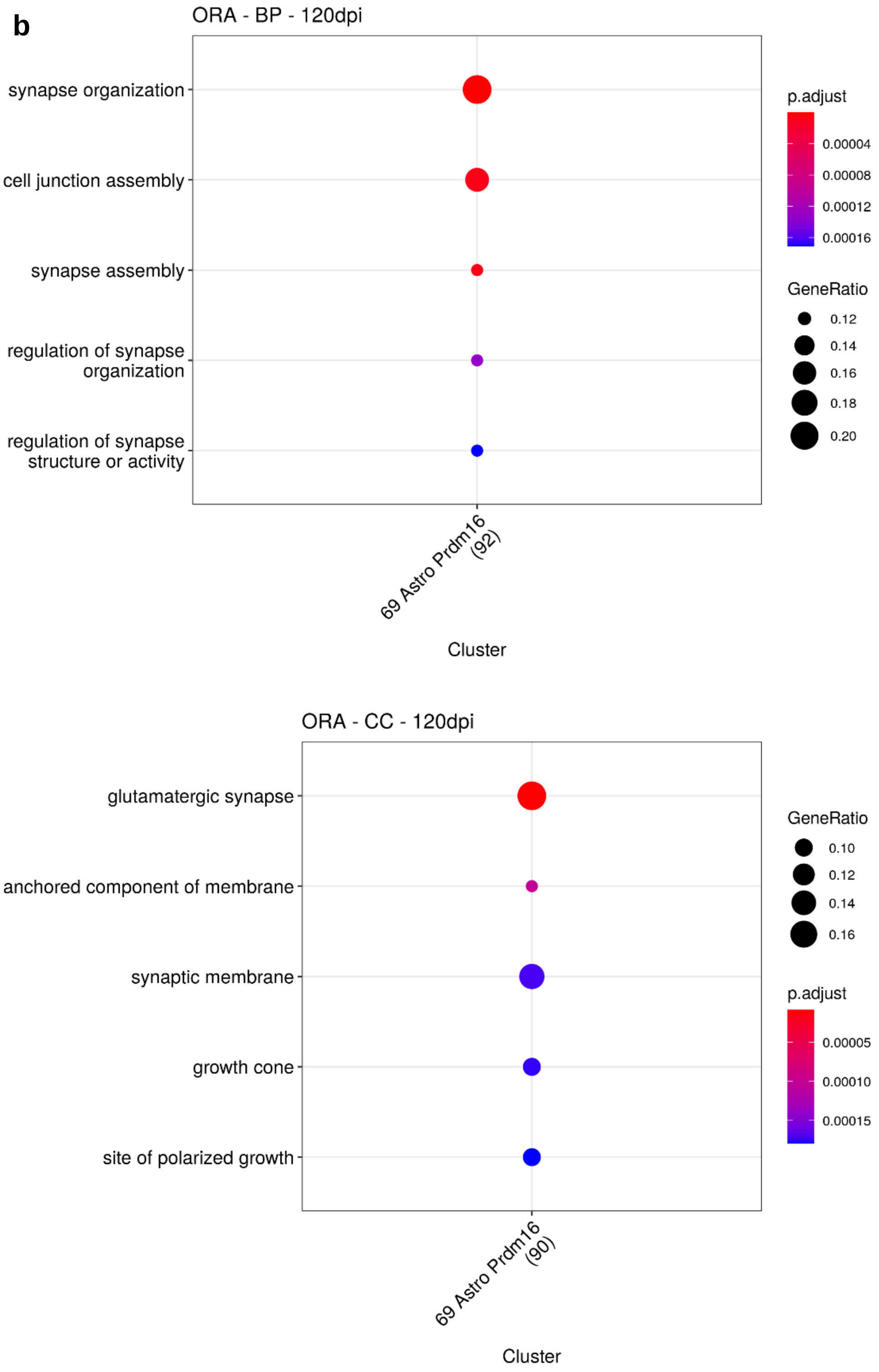
A common approach to analysing gene expression profiles in disease is to identify interesting biological functions of gene sets and select gene-members of interest. One of the ways to uncover perturbed biological processes based on the DE gene lists is to perform enrichment analyses. Here we selected two well-documented and long-standing methodologies, namely the over-representation analysis (ORA) and the gene-set enrichment analysis (GSEA).

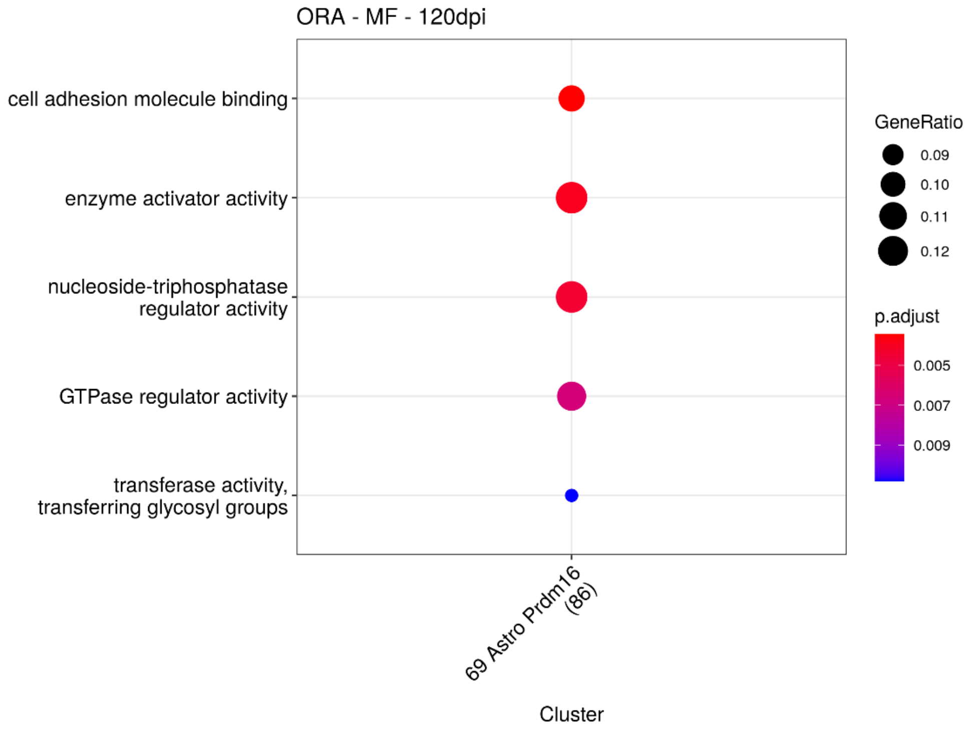
The ORA is a widely used approach to determine if known biological processes or functions are over-represented in an experimentally derived gene list (Boyle et al., 2004). It can identify groups of interesting genes when the differential expression is substantial, however it can miss subtle signatures with small differences in expression that are evidenced in a coordinated way in a set of related genes. While the ORA uses only the set of DEGs, an alternative approach, the GSEA, can be used with all genes, even the ones with slight changes in expression. This allows the method to identify situations where all genes in a predefined set change in small but coordinated ways (Subramanian et al., 2005).

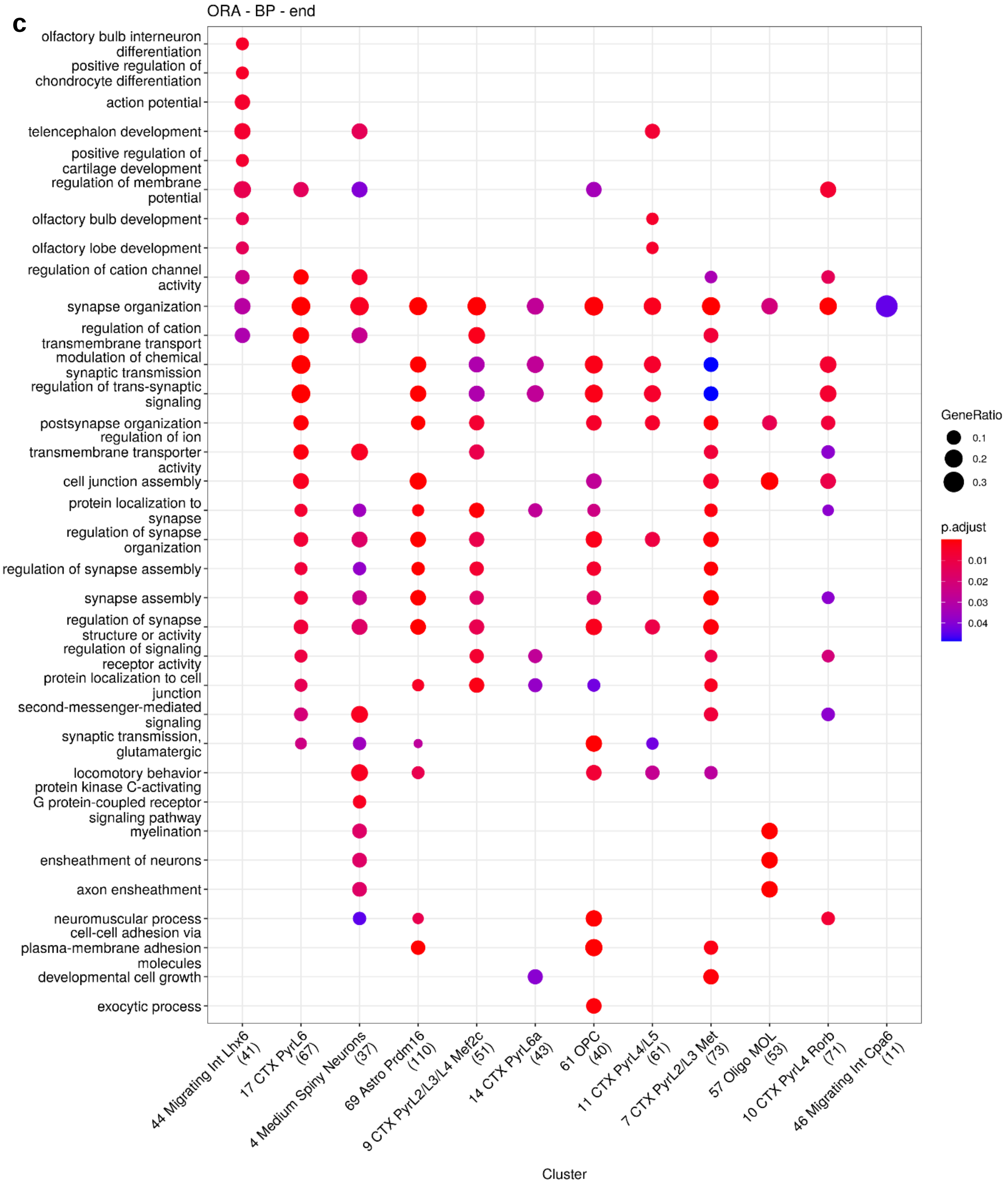
In both cases, the genes need to be mapped to pre-defined gene sets. For our analysis, we used the Gene Ontology (GO) classification which defines concepts and classes to describe gene functions and the relationship between them (Ashburner et al., 2000). GO classifies genes based on the following three aspects: Molecular Function (MF), based on the molecular activity of gene products, Biological Process (BP), based on larger processes and pathways consisting of multiple gene products, and Cellular Component (CC), based on the cellular location where a gene product is active.

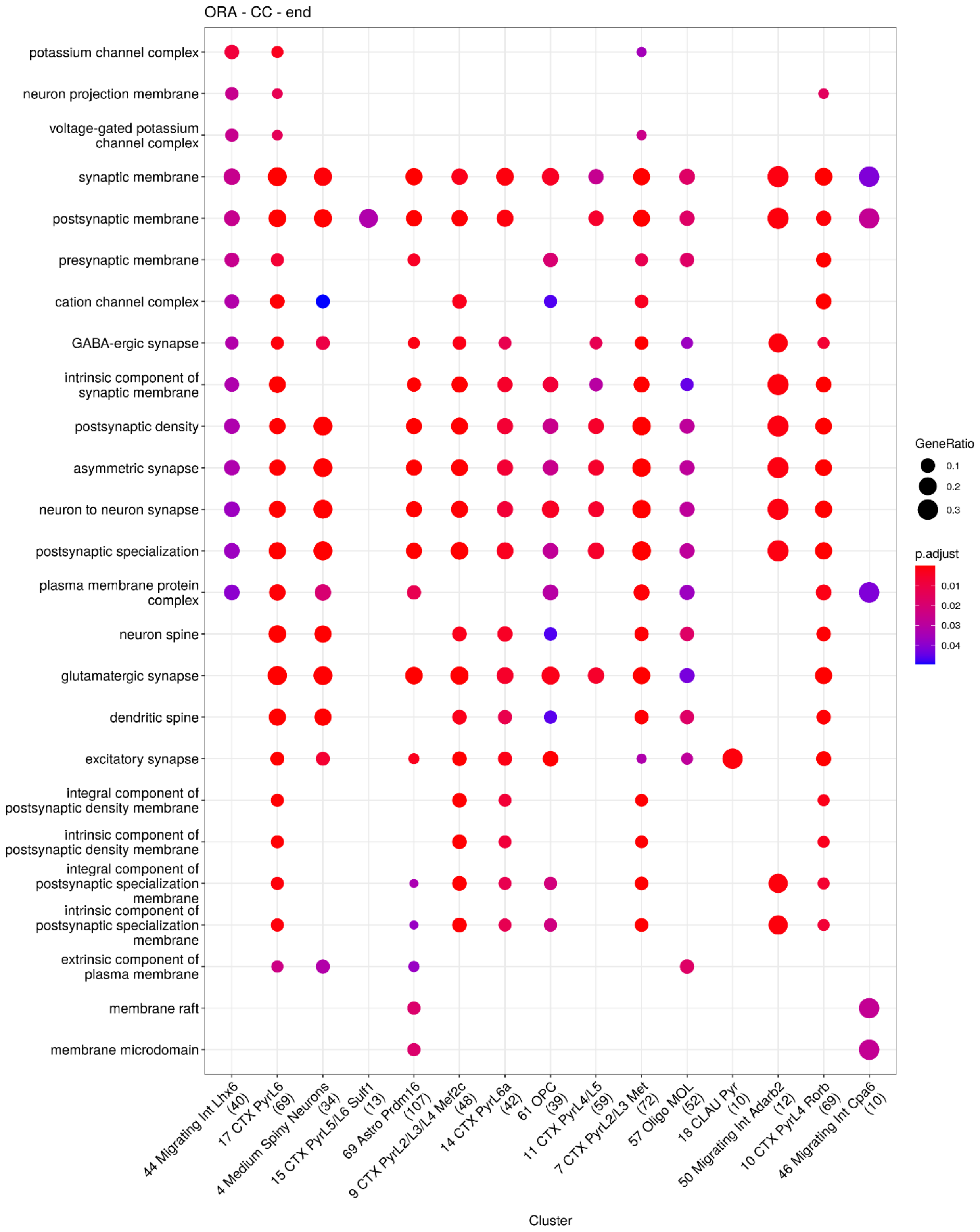
We used the DEGs from Seurat with clusterProfiler, a utility that facilitates enrichment analyses, to uncover perturbed biological functions and classify them based on the BP, CC, and MF systems. The ORA was performed for each cluster individually and identified enriched pathways at the 20 dpi, 120 dpi and end-stage time points (Figure 4.16). All GO terms of the 20 dpi time point were associated with only one cluster —cluster 11 of cortical neurons— which was also the cluster with the highest number of DEGs (Figure 4.16a). The BP classification identified terms related to the metabolism of glycosylated proteins, while the MF classification included relevant functions of transferases in general, including enzymes that direct oligosaccharide processing. No enriched terms were identified for the CC classification. No enrichment was observed for the 40 dpi and 80 dpi time points since the number of DEGs was very low. Cluster 69 of astrocytes was the only cluster with enriched GO terms at 120 dpi (Figure 4.16b). The BP and CC classifications pointed to synaptic dysregulation, especially of the glutamatergic system, while the MF classification identified perturbations in cell adhesion and regulation of nucleoside-triphosphatases. A plethora of GO terms was found to be enriched at the end-stage, involving most clusters of neurons, oligodendrocytes, OPCs, and astrocytes (Figure 4.16c). The biological processes identified suggest a global dysregulation of synaptic function across all cell types; the “synapse organization” pathway was the most prevalent in all clusters, while more than 10 synapse-related pathways were dysregulated in different degrees. We attempted to validate these findings by performing real-time quantitative PCR on bulk brain nuclei suspensions and could identify 2 genes (*Apoe* and *Grin2a*) that showed statistically significant differential expression in the last time point out of a selection of 5 assayed genes involved in the synapse organisation pathway (*Apoe, Grin2a, Nrp1, Ptk2, Rph3a*) (Supplementary Figure 11). Interestingly, the migrating interneurons showed a diverging dysregulation profile, with most of the dysregulated terms being relevant to development and differentiation. Additional terms were identified for Oligodendrocytes and OPCs relevant to myelination, axon ensheathment and cell adhesion. A trend was also evident regarding the CC classification, which recapitulated the BP pathways and identified the synapse as the location of the perturbations for most cell clusters, except the migrating interneurons which had lower associated p-values. The terms identified by the MF classification were not as universal and included fewer genes with lower p-values. Of note are the terms relevant to cell adhesion and syntaxin binding identified in the OPC cluster, and ion regulation, voltage-gated channels, and phosphodiesterase activity identified in clusters of cortical neurons. External Supplementary Table 6 includes all identified GO terms of the ORA that passed the filtering criteria.


a
ORA - BP - 20dpi
heparan sulfate proteoglycan biosynthetic process
heparan sulfate proteoglycan metabolic process
proteoglycan biosynthetic process
proteoglycan metabolic process
glycoprotein biosynthetic process
11 CTX PyrL4/L5 (18)
Cluster
GeneRatio
0.17
0.18
0.19
0.20
0.21
0.22
p.adjust
0.005
0.010
ORA - MF - 20dpi
acetylglucosaminyltransferase activity
UDP-glycosyltransferase activity
transferase activity, transferring hexosyl groups
sulfur compound binding
transferase activity, transferring glycosyl groups
11 CTX PyrL4/L5 (16)
Cluster
p.adjust
0.005
0.010
0.015
0.020
GeneRatio
0.1875


b
ORA - BP - 120dpi
synapse organization
cell junction assembly
synapse assembly
regulation of synapse organization
regulation of synapse structure or activity
69 Astro Prdm16 (92)
Cluster
p.adjust
0.00004
0.00008
0.00012
0.00016
GeneRatio
0.12
0.14
0.16
0.18
0.20
ORA - CC - 120dpi
glutamatergic synapse
anchored component of membrane
synaptic membrane
growth cone
site of polarized growth
69 Astro Prdm16 (90)
Cluster
GeneRatio
0.10
0.12
0.14
0.16
p.adjust
0.00005
0.00010
0.00015


ORA - MF - 120dpi
cell adhesion molecule binding
enzyme activator activity
nucleoside-triphosphatase regulator activity
GTPase regulator activity
transferase activity, transferring glycosyl groups
69 Astro Prdm16 (86)
Cluster
GeneRatio
0.09
0.10
0.11
0.12
p.adjust
0.005
0.007
0.009


c
ORA - BP - end
olfactory bulb interneuron differentiation
positive regulation of chondrocyte differentiation
action potential
telencephalon development
positive regulation of cartilage development
regulation of membrane potential
olfactory bulb development
olfactory lobe development
regulation of cation channel activity
synapse organization
regulation of cation transmembrane transport
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
postsynapse organization
regulation of ion transmembrane transporter activity
cell junction assembly
protein localization to synapse
regulation of synapse organization
regulation of synapse assembly
synapse assembly
regulation of synapse structure or activity
regulation of signaling receptor activity
protein localization to cell junction
second-messenger-mediated signaling
synaptic transmission, glutamatergic
locomotory behavior
protein kinase C-activating G protein-coupled receptor signaling pathway
myelination
ensheathment of neurons
axon ensheathment
neuromuscular process
cell-cell adhesion via plasma-membrane adhesion molecules
developmental cell growth
exocytic process
44 Migrating Int Lhx6 (41)
17 CTX PyrL6 (67)
4 Medium Spiny Neurons (37)
69 Astro Prdm16 (110)
9 CTX PyrL2/L3/L4 Mef2c (51)
14 CTX PyrL6a (43)
61 OPC (40)
11 CTX PyrL4/L5 (61)
7 CTX PyrL2/L3 Met (73)
57 Oligo MOL (53)
10 CTX PyrL4 Rorb (71)
46 Migrating Int Cpa6 (11)
Cluster
GeneRatio
0.1
0.2
0.3
p.adjust
0.01
0.02
0.03
0.04


ORA - CC - end
potassium channel complex
neuron projection membrane
voltage-gated potassium channel complex
synaptic membrane
postsynaptic membrane
presynaptic membrane
cation channel complex
GABA-ergic synapse
intrinsic component of synaptic membrane
postsynaptic density
asymmetric synapse
neuron to neuron synapse
postsynaptic specialization
plasma membrane protein complex
neuron spine
glutamatergic synapse
dendritic spine
excitatory synapse
integral component of postsynaptic density membrane
intrinsic component of postsynaptic density membrane
integral component of postsynaptic specialization membrane
intrinsic component of postsynaptic specialization membrane
extrinsic component of plasma membrane
membrane raft
membrane microdomain
44 Migrating Int Lhx6 (40)
17 CTX PyrL6 (69)
4 Medium Spiny Neurons (34)
15 CTX PyrL5/L6 Sulf1 (13)
69 Astro Prdm16 (107)
9 CTX PyrL2/L3/L4 Mef2c (48)
14 CTX PyrL6a (42)
61 OPC (39)
11 CTX PyrL4/L5 (59)
7 CTX PyrL2/L3 Met (72)
57 Oligo MOL (52)
18 CLAU Pyr (10)
50 Migrating Int Adarb2 (12)
10 CTX PyrL4 Rorb (69)
46 Migrating Int Cpa6 (10)
Cluster
GeneRatio
0.1
0.2
0.3
p.adjust
0.01
0.02
0.03
0.04

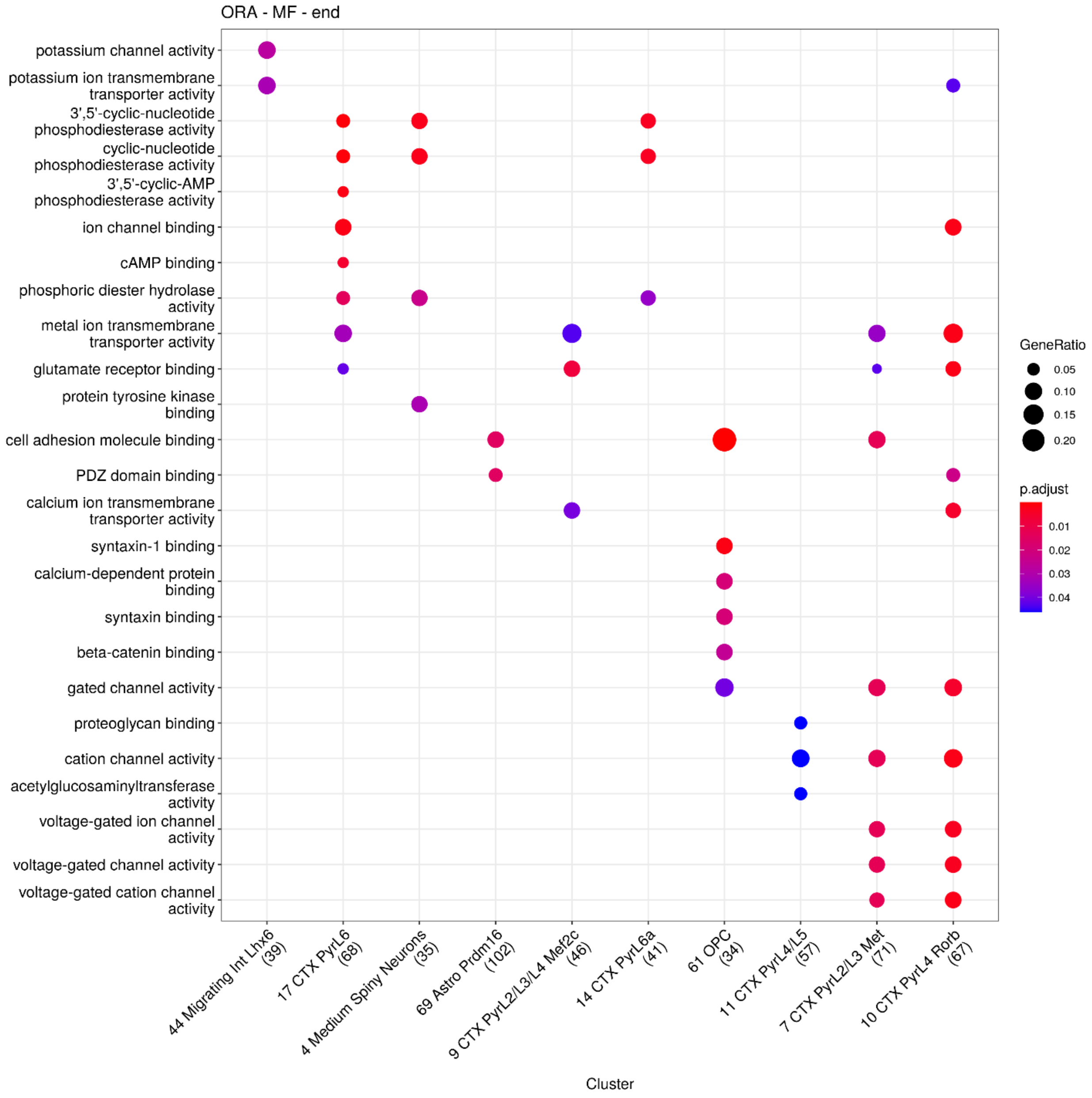
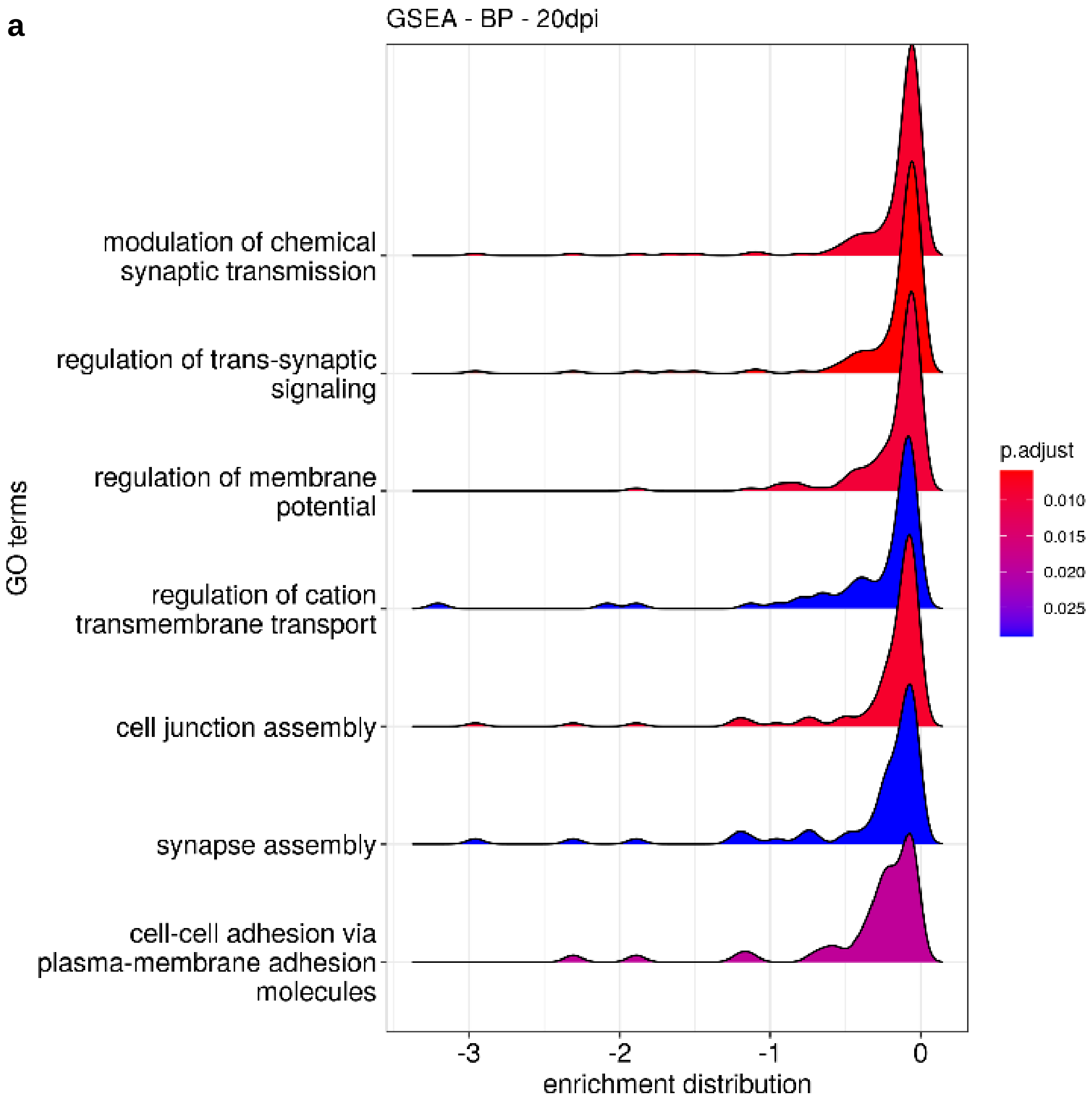


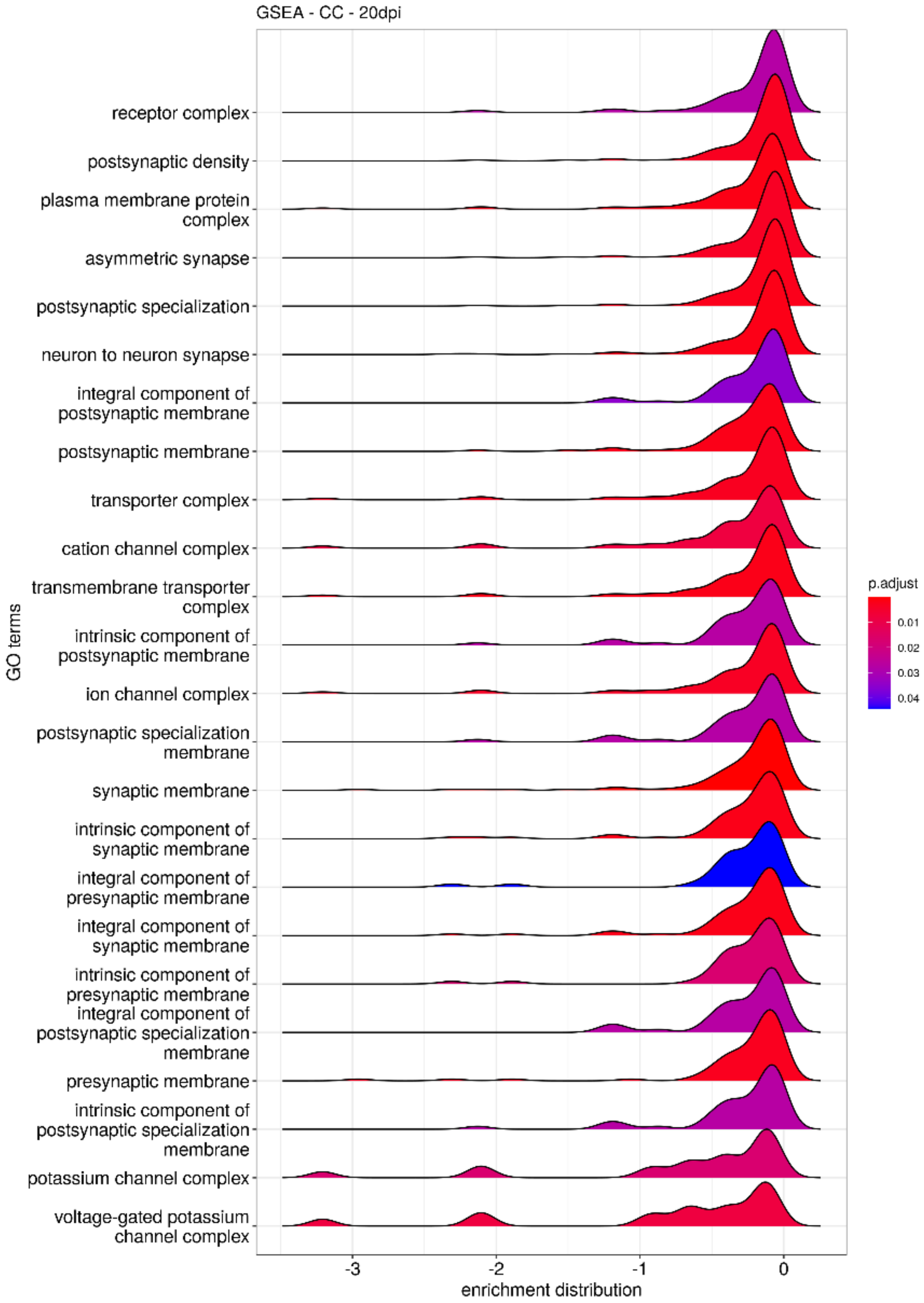
***Figure 4.16: An over-representation analysis identified enriched GO terms at 20 dpi, 120 dpi and end-stage and global synaptic perturbations during the late stages of the disease.*** *We used clusterProfiler to uncover perturbed biological functions and classify them based on the BP, CC, and MF systems. The ORA was performed for each cluster individually and identified enriched pathways at the 20 dpi **(a)**, 120 dpi **(b)** and end-stage time points **(c)**. **(a)** All GO terms of the 20 dpi time point were associated with cluster 11 of cortical neurons. The BP classification identified terms related to the metabolism of glycosylated proteins, while the MF classification included functions of transferases in general, including enzymes that direct oligosaccharide processing. No enriched terms were identified for the CC classification. **(b)** Cluster 69 of astrocytes was the only cluster with enriched GO terms at 120 dpi. The BP and CC*

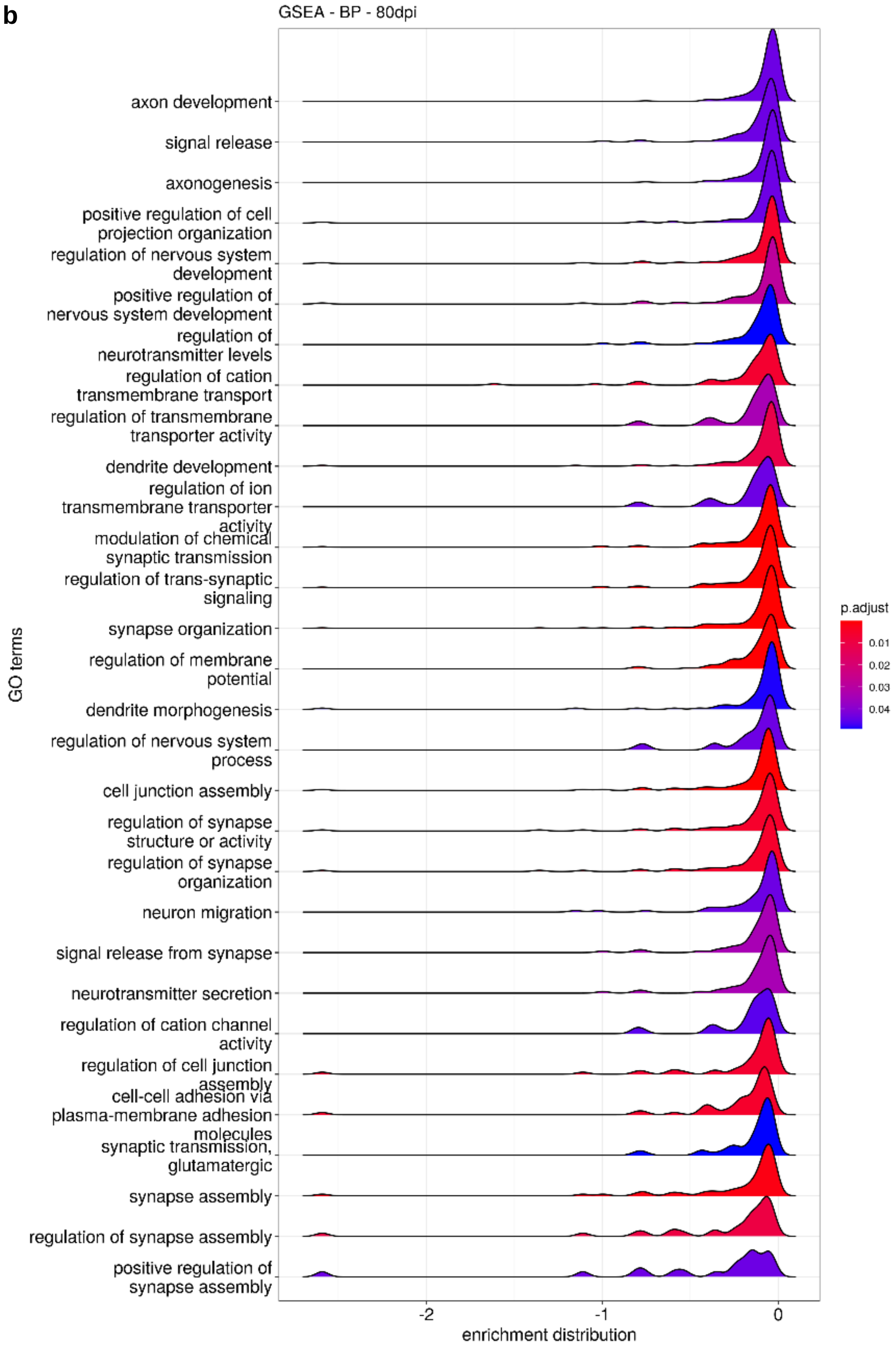
*classifications pointed to synaptic dysregulation, especially of the glutamatergic system, while the MF classification identified perturbations in cell adhesion and regulation of nucleoside-triphosphatases.* ***(c)*** *A plethora of GO terms was found to be enriched at the end-stage, involving most clusters of neurons, oligodendrocytes, OPCs, and astrocytes. The biological processes identified suggest a global dysregulation of synaptic function across all cell types. The migrating interneurons showed a diverging dysregulation profile, with most of the dysregulated terms being relevant to development and differentiation. Additional terms were identified for Oligodendrocytes and OPCs relevant to myelination, axon ensheathment and cell adhesion. A universal trend was also evident regarding the CC classification, which recapitulated the BP pathways and identified the synapse as the location of the perturbations for most cell clusters, except the migrating interneurons which had lower associated p-values. The terms identified by the MF classification were not as universal and included fewer genes with lower p-values. Of note are the terms relevant to cell adhesion and syntaxin binding identified in the OPC cluster, and ion regulation, voltage-gated channels, and phosphodiesterase activity identified in clusters of cortical neurons. No enrichment was observed for the 40 dpi and 80 dpi time points. Circle size corresponds to the gene ratio (the ratio of the intersection of DE genes in our data with the GO gene set over the intersection of DE genes in our data with all the genes of the GO collection). The colour of the circles corresponds to the Benjamini-Hochberg adjusted p-value. Missing combinations of time points and classifications mean that no enriched pathways were identified (20 dpi – CC, all classifications for 40 and 80 dpi).*

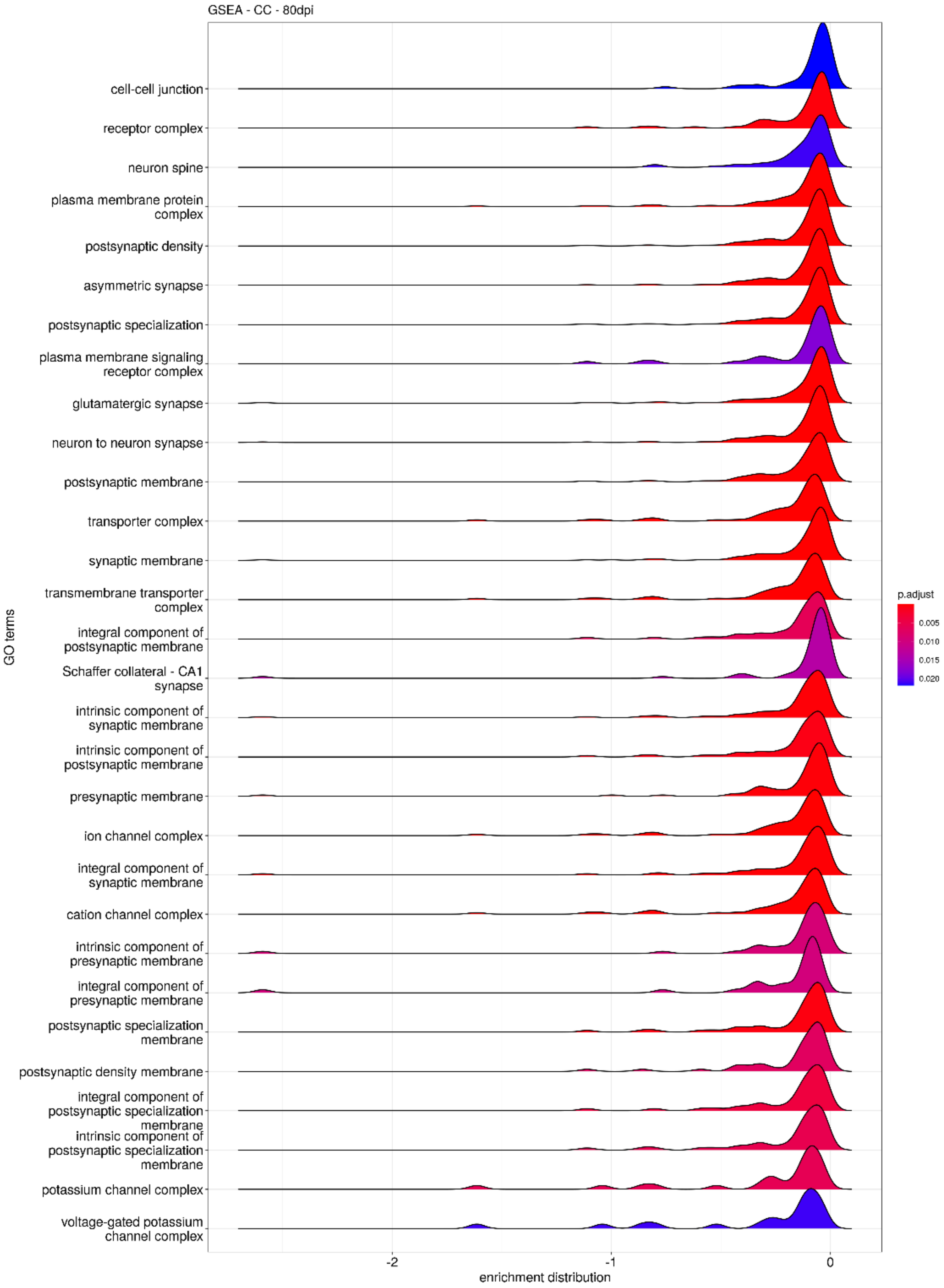
The GSEA is a method that determines whether a pre-defined gene set is differentially enriched between two biological states, in our case disease and controls. The GSEA calculates the enrichment statistic by walking down a ranked list of genes and increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not. It can identify smaller biological differences as it compares the expression of all genes sequenced and does not rely on arbitrary criteria of differential expression. We used clusterProfiler to perform a GSEA across all clusters at each time point and identify globally affected biological processes. We observed a global downregulation of all identified GO gene sets across all time points at 20, 80, 120 dpi and end-stage, while there were no enriched terms at 40 dpi (Figure 4.17). The analysis identified synaptic and cell adhesion dysregulation as early as the 20 dpi (Figure 4.17a - BP) and localisation of gene products in the synapses and ion channel complexes (Figure 4.17a - CC). No enriched gene sets were identified for the 40 dpi time point, which was in agreement with the previous ORA. In contrast with previous results, the GSEA did identify synaptic perturbations at the 80 dpi time point, possibly due to the higher sensitivity of the analysis (Figure 4.17b). The same pathways relevant to synaptic function and cell adhesion were also found to be dysregulated at the last two time points (Figure 4.17 c and d). The MF classification corroborated the results of the ORA, identifying perturbations in systems of ion homeostasis.

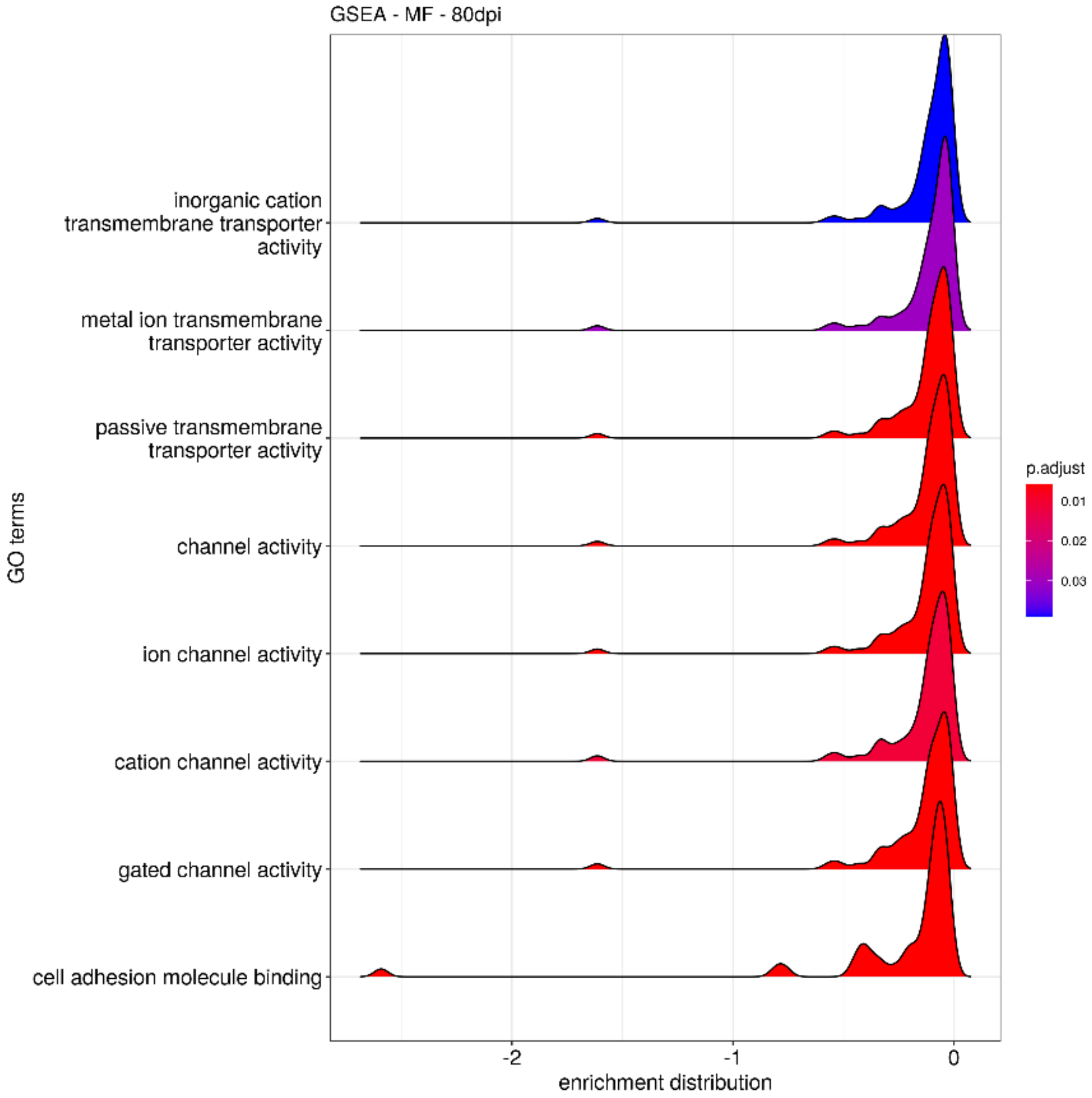
**a**


GSEA - BP - 20dpi
GO terms
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
regulation of membrane potential
regulation of cation transmembrane transport
cell junction assembly
synapse assembly
cell-cell adhesion via plasma-membrane adhesion molecules
-3
-2
-1
0
enrichment distribution
p.adjust
0.010
0.015
0.020
0.025

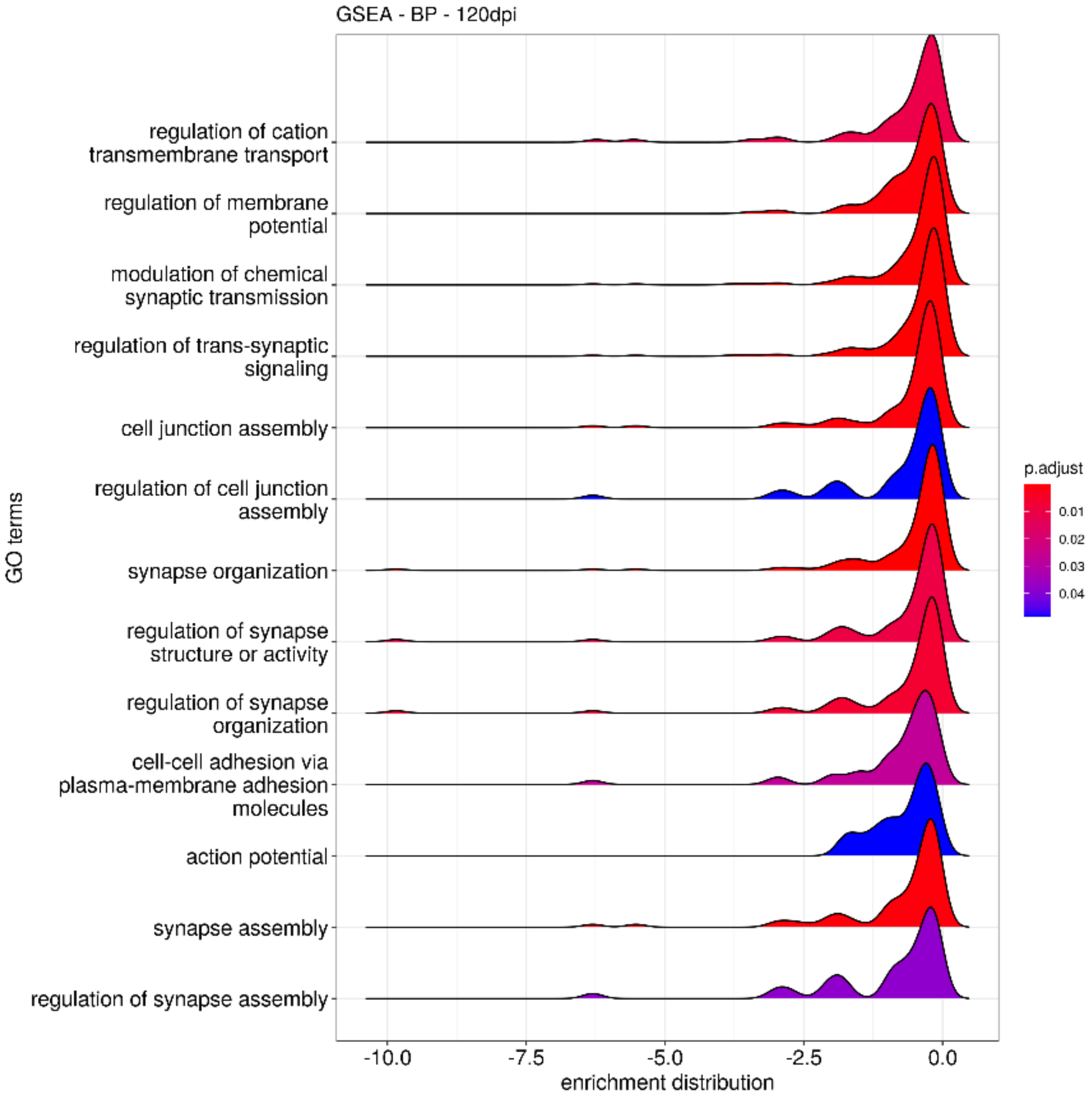

GSEA - CC - 20dpi
GO terms
receptor complex
postsynaptic density
plasma membrane protein complex
asymmetric synapse
postsynaptic specialization
neuron to neuron synapse
integral component of postsynaptic membrane
postsynaptic membrane
transporter complex
cation channel complex
transmembrane transporter complex
intrinsic component of postsynaptic membrane
ion channel complex
postsynaptic specialization membrane
synaptic membrane
intrinsic component of synaptic membrane
integral component of presynaptic membrane
integral component of synaptic membrane
intrinsic component of presynaptic membrane
integral component of postsynaptic specialization membrane
presynaptic membrane
intrinsic component of postsynaptic specialization membrane
potassium channel complex
voltage-gated potassium channel complex
-3
-2
-1
0
enrichment distribution
p.adjust
0.01
0.02
0.03
0.04

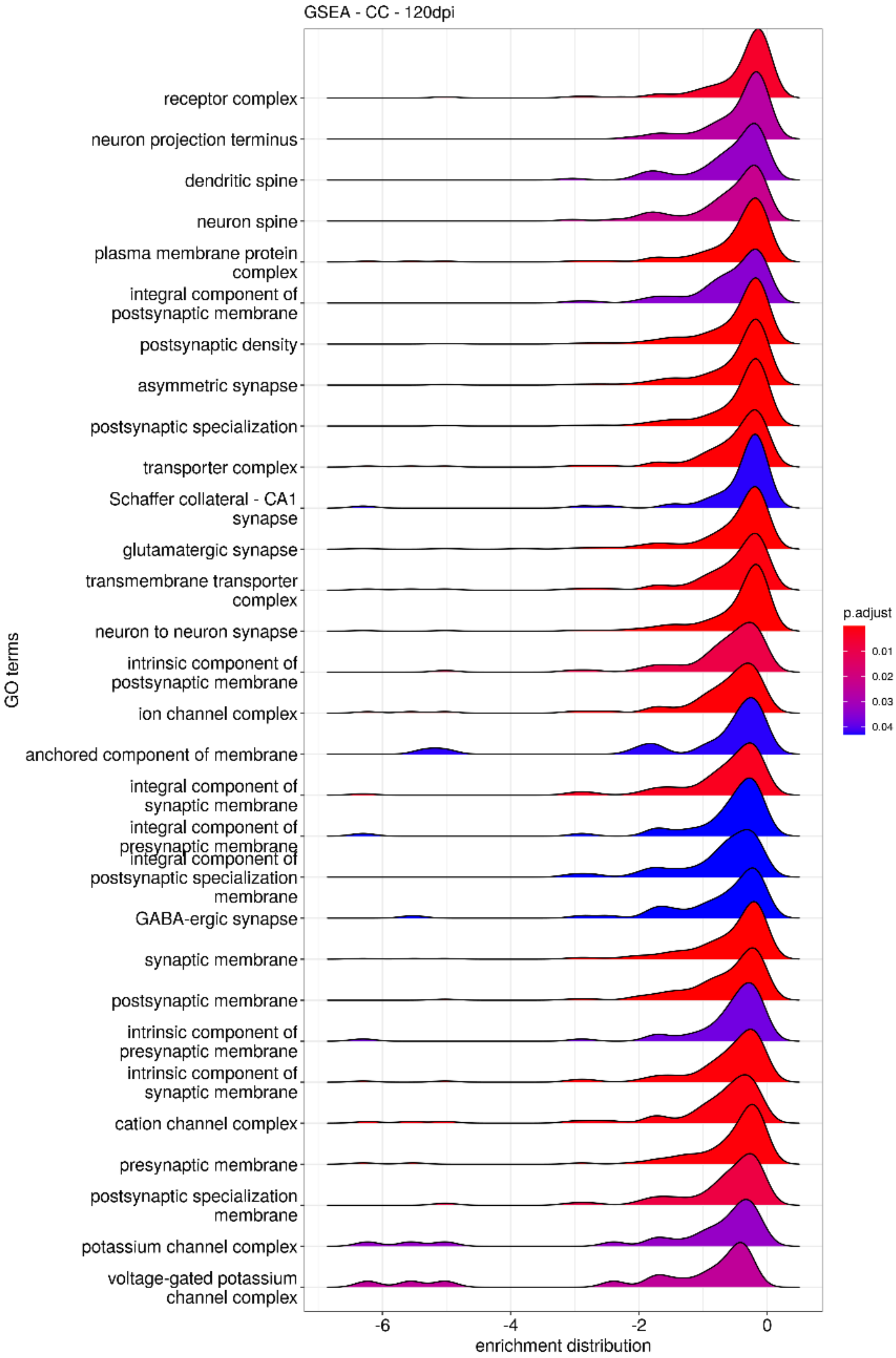

b
GSEA - BP - 80dpi
GO terms
axon development
signal release
axonogenesis
positive regulation of cell projection organization
regulation of nervous system development
positive regulation of nervous system development
regulation of neurotransmitter levels
regulation of cation transmembrane transport
regulation of transmembrane transporter activity
dendrite development
regulation of ion transmembrane transporter activity
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
synapse organization
regulation of membrane potential
dendrite morphogenesis
regulation of nervous system process
cell junction assembly
regulation of synapse structure or activity
regulation of synapse organization
neuron migration
signal release from synapse
neurotransmitter secretion
regulation of cation channel activity
regulation of cell junction assembly
cell-cell adhesion via plasma-membrane adhesion molecules
synaptic transmission, glutamatergic
synapse assembly
regulation of synapse assembly
positive regulation of synapse assembly
-2
-1
0
enrichment distribution
p.adjust
0.01
0.02
0.03
0.04

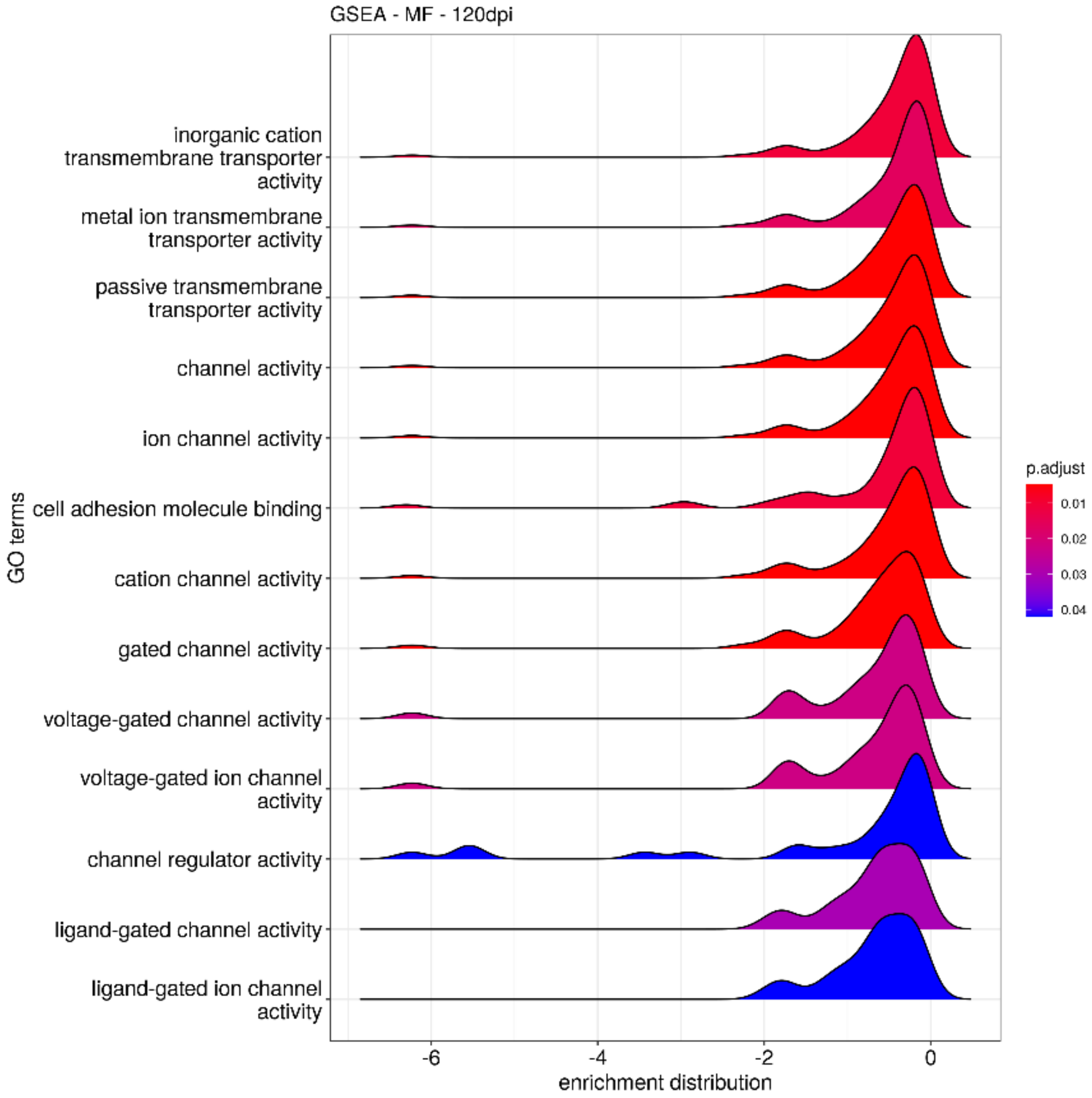

GSEA - CC - 80dpi
cell-cell junction
receptor complex
neuron spine
plasma membrane protein complex
postsynaptic density
asymmetric synapse
postsynaptic specialization
plasma membrane signaling receptor complex
glutamatergic synapse
neuron to neuron synapse
postsynaptic membrane
transporter complex
synaptic membrane
transmembrane transporter complex
integral component of postsynaptic membrane
Schaffer collateral - CA1 synapse
intrinsic component of synaptic membrane
intrinsic component of postsynaptic membrane
presynaptic membrane
ion channel complex
integral component of synaptic membrane
cation channel complex
intrinsic component of presynaptic membrane
integral component of presynaptic membrane
postsynaptic specialization membrane
postsynaptic density membrane
integral component of postsynaptic specialization membrane
intrinsic component of postsynaptic specialization membrane
potassium channel complex
voltage-gated potassium channel complex
GO terms
-2
-1
0
enrichment distribution
p.adjust
0.005
0.010
0.015
0.020

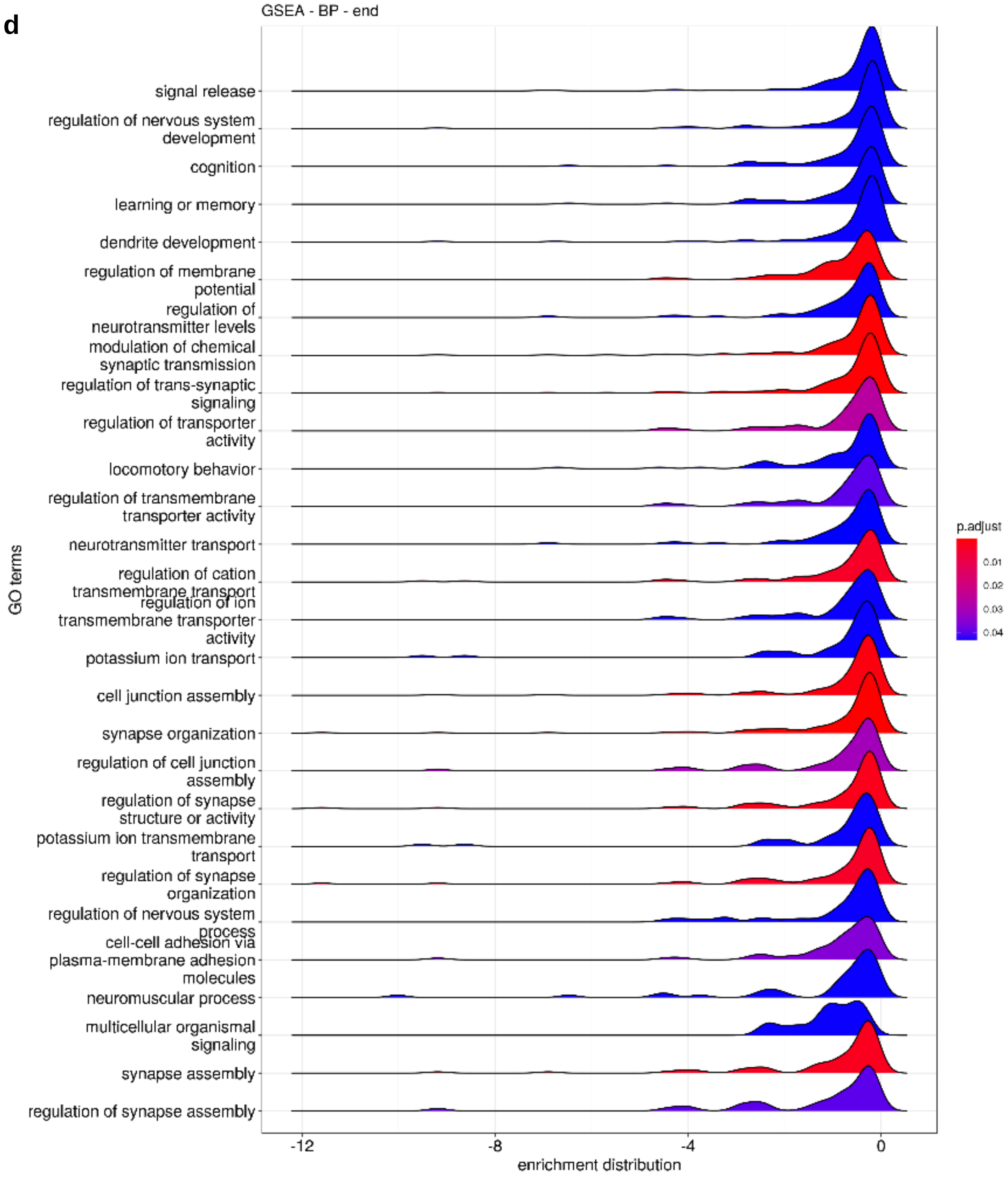

GSEA - MF - 80dpi
GO terms
inorganic cation transmembrane transporter activity
metal ion transmembrane transporter activity
passive transmembrane transporter activity
channel activity
ion channel activity
cation channel activity
gated channel activity
cell adhesion molecule binding
-2
-1
0
enrichment distribution
p.adjust
0.01
0.02
0.03

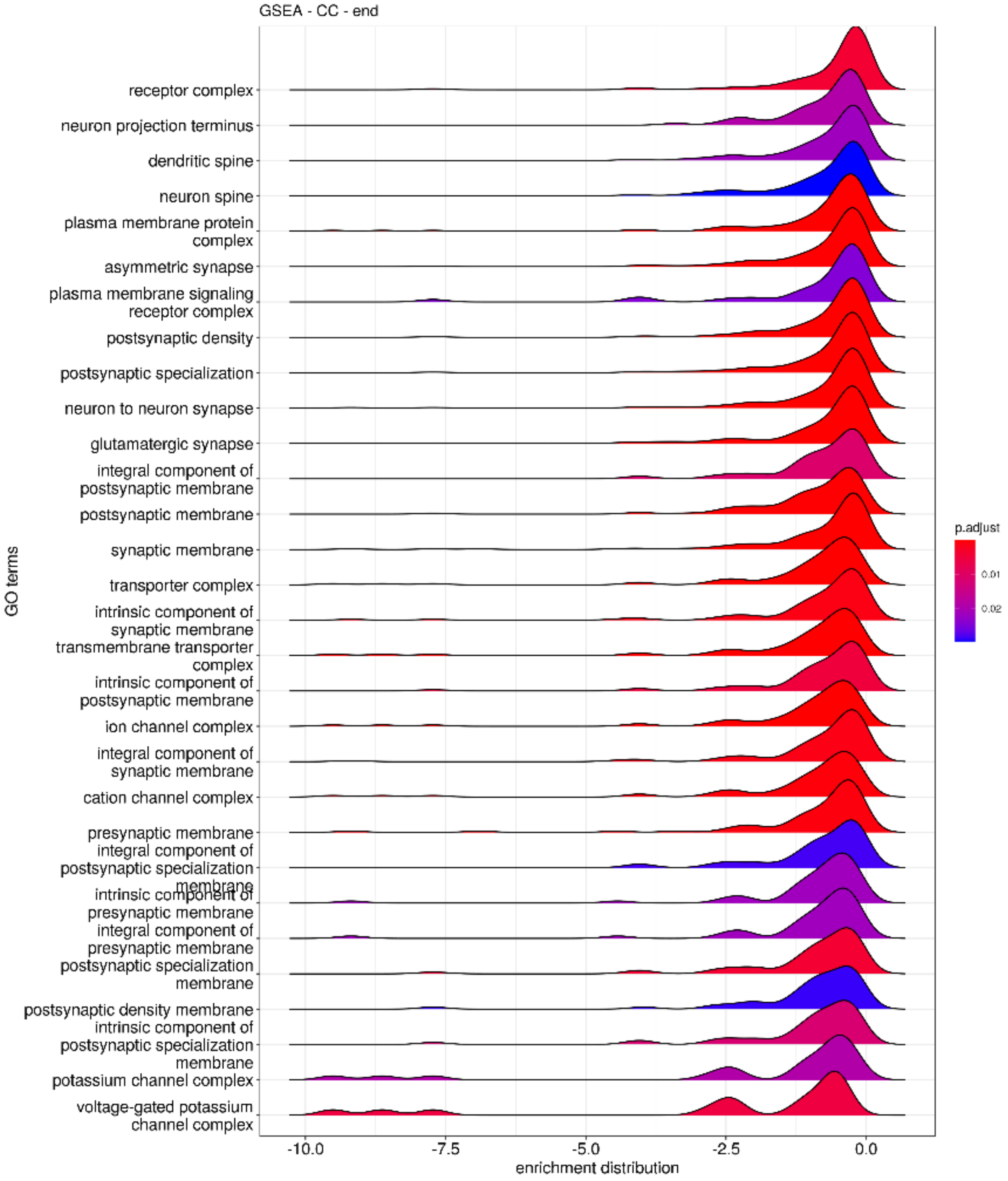
c


GSEA - BP - 120dpi
GO terms
regulation of cation transmembrane transport
regulation of membrane potential
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
cell junction assembly
regulation of cell junction assembly
synapse organization
regulation of synapse structure or activity
regulation of synapse organization
cell-cell adhesion via plasma-membrane adhesion molecules
action potential
synapse assembly
regulation of synapse assembly
-10.0
-7.5
-5.0
-2.5
0.0
enrichment distribution
p.adjust
0.01
0.02
0.03
0.04


GSEA - CC - 120dpi
receptor complex
neuron projection terminus
dendritic spine
neuron spine
plasma membrane protein complex
integral component of postsynaptic membrane
postsynaptic density
asymmetric synapse
postsynaptic specialization
transporter complex
Schaffer collateral - CA1 synapse
glutamatergic synapse
transmembrane transporter complex
neuron to neuron synapse
intrinsic component of postsynaptic membrane
ion channel complex
anchored component of membrane
integral component of synaptic membrane
integral component of presynaptic membrane
integral component of postsynaptic specialization membrane
GABA-ergic synapse
synaptic membrane
postsynaptic membrane
intrinsic component of presynaptic membrane
intrinsic component of synaptic membrane
cation channel complex
presynaptic membrane
postsynaptic specialization membrane
potassium channel complex
voltage-gated potassium channel complex
GO terms
-6
-4
-2
0
enrichment distribution
p.adjust
0.01
0.02
0.03
0.04


GSEA - MF - 120dpi
GO terms
inorganic cation transmembrane transporter activity
metal ion transmembrane transporter activity
passive transmembrane transporter activity
channel activity
ion channel activity
cell adhesion molecule binding
cation channel activity
gated channel activity
voltage-gated channel activity
voltage-gated ion channel activity
channel regulator activity
ligand-gated channel activity
ligand-gated ion channel activity
-6
-4
-2
0
enrichment distribution
p.adjust
0.01
0.02
0.03
0.04


d
GSEA - BP - end
signal release
regulation of nervous system development
cognition
learning or memory
dendrite development
regulation of membrane potential
regulation of neurotransmitter levels
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
regulation of transporter activity
locomotory behavior
regulation of transmembrane transporter activity
neurotransmitter transport
regulation of cation transmembrane transport
regulation of ion transmembrane transporter activity
potassium ion transport
cell junction assembly
synapse organization
regulation of cell junction assembly
regulation of synapse structure or activity
potassium ion transmembrane transport
regulation of synapse organization
regulation of nervous system process
cell-cell adhesion via plasma-membrane adhesion molecules
neuromuscular process
multicellular organismal signaling
synapse assembly
regulation of synapse assembly
GO terms
-12
-8
-4
0
enrichment distribution
p.adjust
0.01
0.02
0.03
0.04


GSEA - CC - end
GO terms
receptor complex
neuron projection terminus
dendritic spine
neuron spine
plasma membrane protein complex
asymmetric synapse
plasma membrane signaling receptor complex
postsynaptic density
postsynaptic specialization
neuron to neuron synapse
glutamatergic synapse
integral component of postsynaptic membrane
postsynaptic membrane
synaptic membrane
transporter complex
intrinsic component of synaptic membrane
transmembrane transporter complex
intrinsic component of postsynaptic membrane
ion channel complex
integral component of synaptic membrane
cation channel complex
presynaptic membrane
integral component of postsynaptic specialization membrane
intrinsic component of presynaptic membrane
integral component of presynaptic membrane
postsynaptic specialization membrane
postsynaptic density membrane
intrinsic component of postsynaptic specialization membrane
potassium channel complex
voltage-gated potassium channel complex
-10.0
-7.5
-5.0
-2.5
0.0
enrichment distribution
p.adjust
0.01
0.02

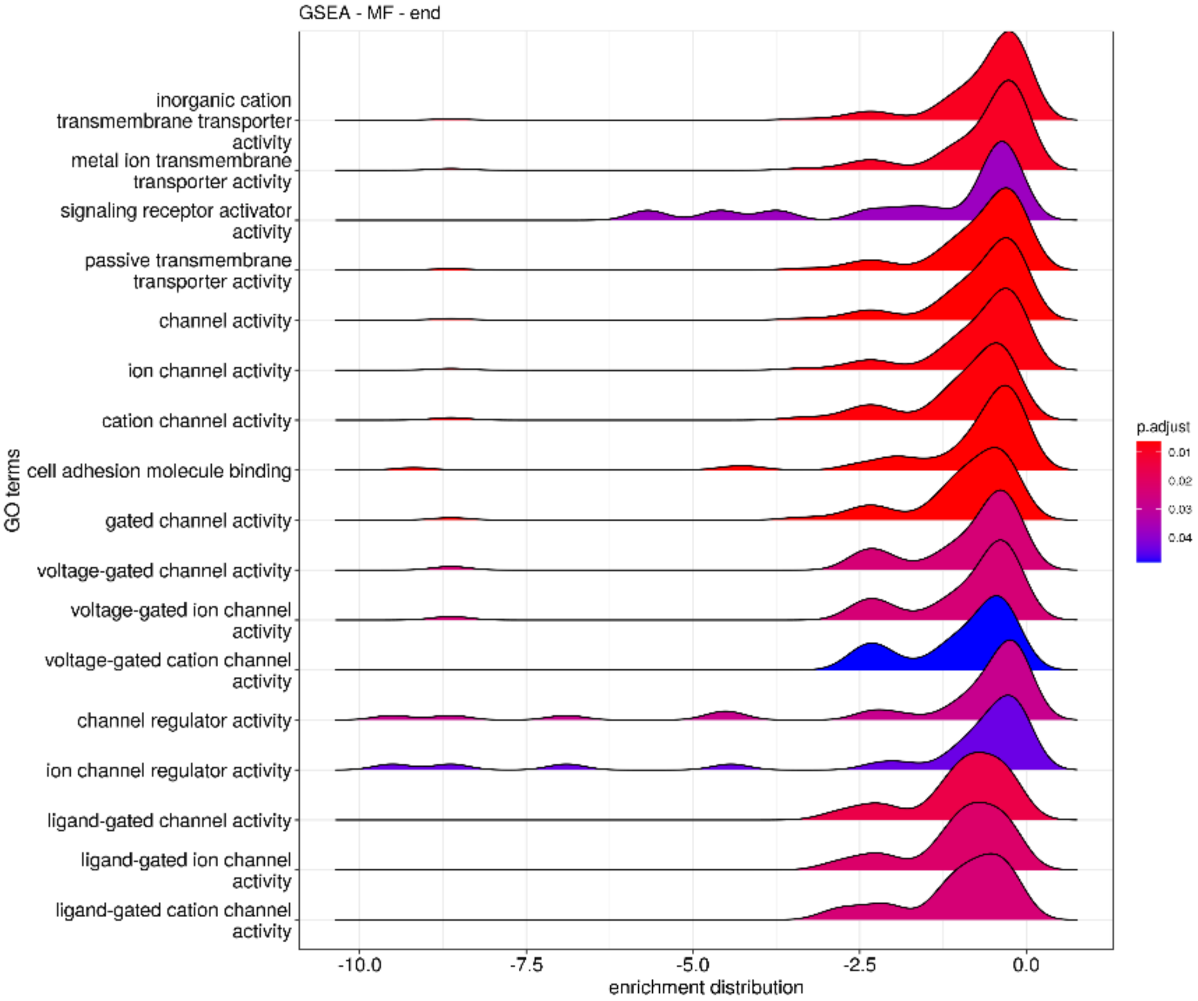
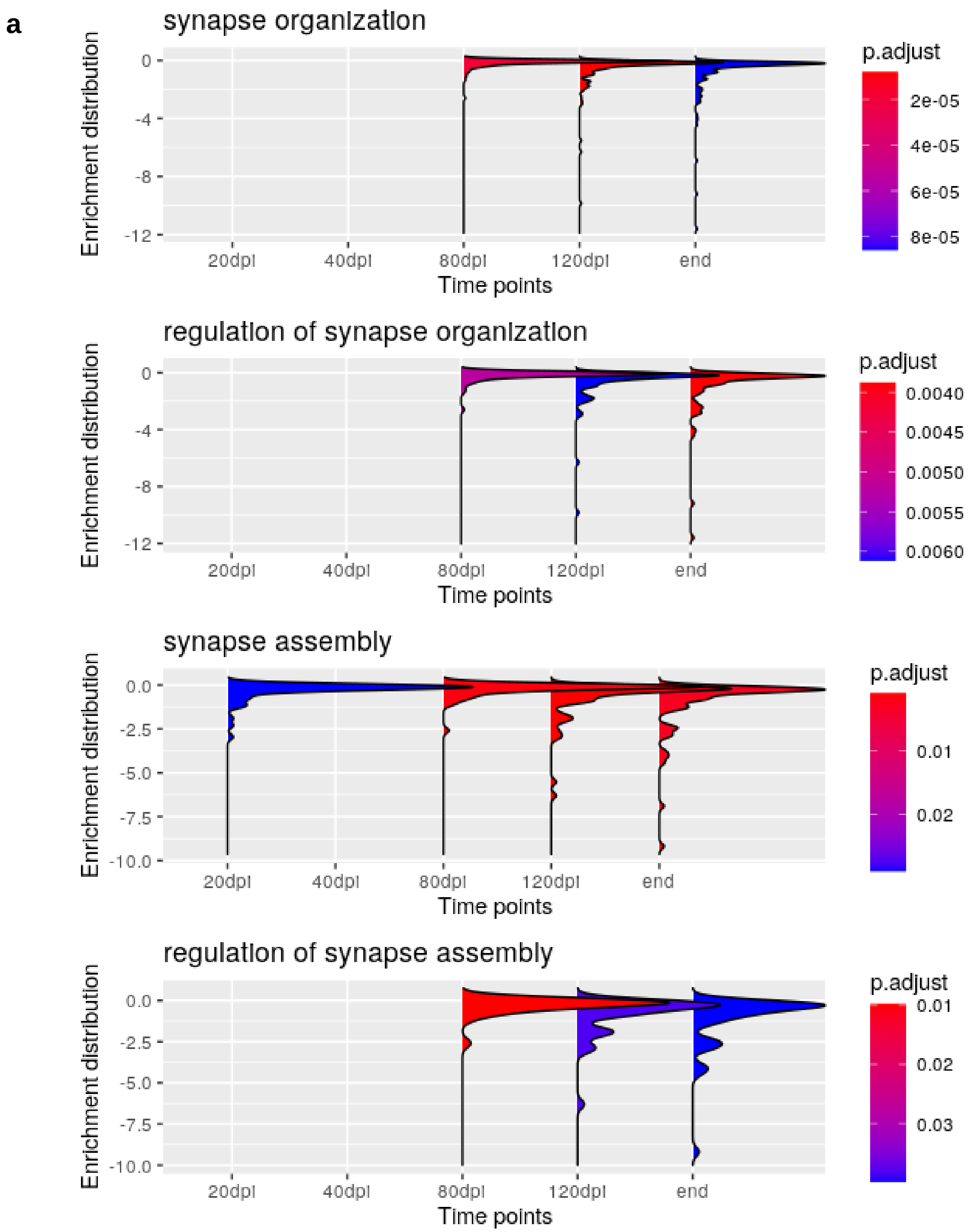
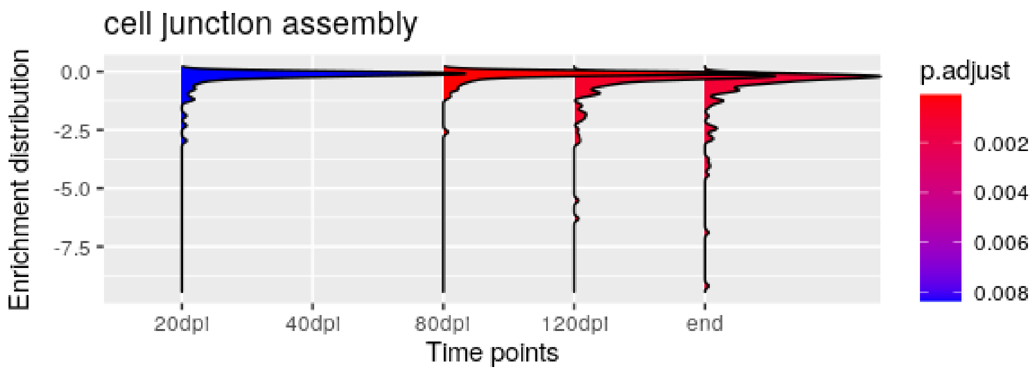
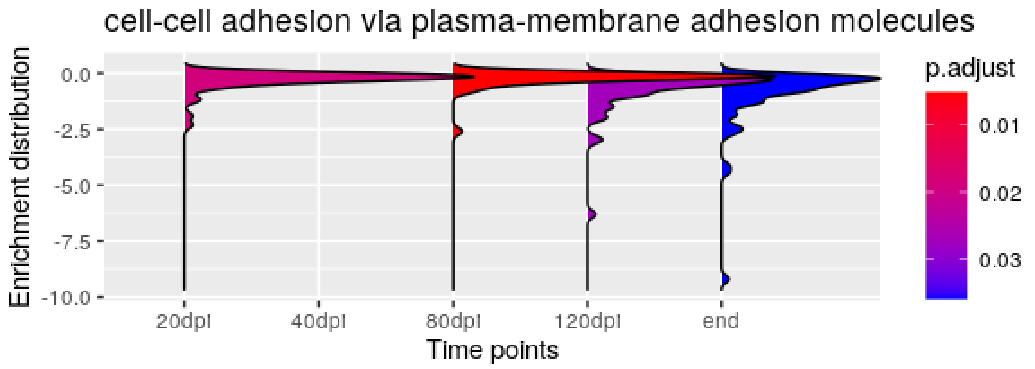


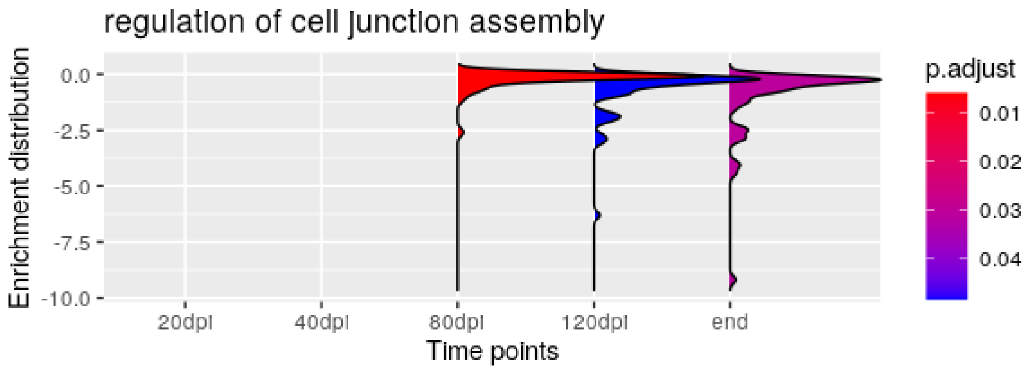
***Figure 4.17: A gene set enrichment analysis suggests global downregulation of all identified GO gene sets across all time points at 20, 80, 120 dpi and end-stage, while there were no enriched terms at 40 dpi.*** *The ridge plots show the distribution of the expression of genes (x-axis) associated with a particular GO term (y-axis). Negative values of the enrichment distribution signify a downregulation of associated genes. The analysis identified synaptic and cell adhesion dysregulation as early as the 20 dpi* ***(a – BP)****, and localisation of gene products in the synapses and ion channel complexes* ***(a - CC)****.* ***(b)*** *Synaptic perturbations were identified as early as the 80 dpi time point, possibly due to the higher sensitivity of the analysis.* ***(c and d)*** *Similar pathways relevant to synaptic function and cell adhesion were also found to be dysregulated at the last two time points. The MF classification corroborated the results of the ORA, identifying perturbations in systems of ion homeostasis. No enriched gene sets were identified for the 40 dpi time point. The colour of each ridge plot represents the BH-adjusted p-value.*

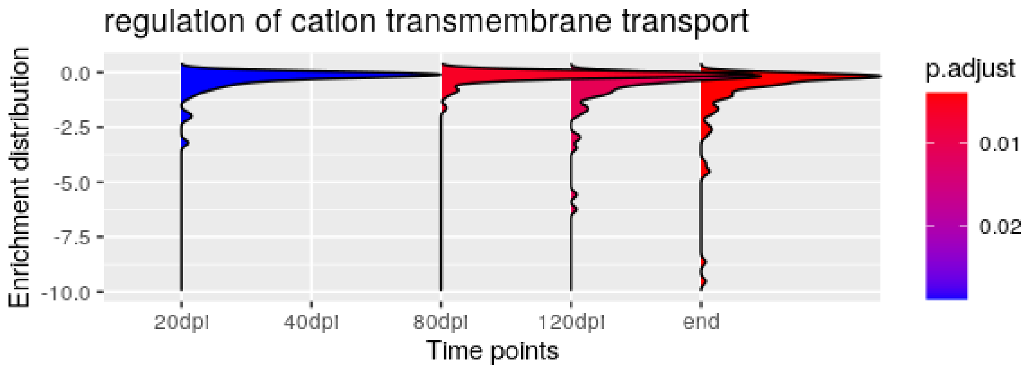
A close examination of the previous plots (Figure 4.17) revealed gene sets that were dysregulated in more than one time point. Since the ridge plots allow us to track the magnitude of expression of gene sets, we investigated whether there were gene sets that become increasingly dysregulated following disease progression. We highlight 8 pathways from the BP classification and 2 pathways from the CC and MF classifications that showed progressive dysregulation, with gene expression becoming increasingly downregulated approaching the disease end-stage (Figure 4.18). Focusing on the BP classification (Figure 4.18a), synapse assembly and organisation pathways, also underlined by previous analyses, were found to be progressively downregulated. The same trend followed the cell adhesion, cell junction and cation transmembrane transport gene sets. Interestingly, a similar pattern was observed for some of those gene sets, where there was enrichment at 20 dpi, which was followed by no enrichment at 40 dpi, then the gene set was enriched again at 80 dpi with an enrichment distribution similar to that of the 20 dpi time point before the distribution shifted to the left at 120 dpi and even further at the end-stage. Shifting our focus to the CC classification (Figure 4.18b), we highlight two closely related gene sets relevant to the potassium channel complex that exhibited the same interesting pattern of progressive downregulation, which was even more substantial than the previous examples. Finally, we highlight two gene sets of the MF classification (cell adhesion molecule binding and channel regulator activity) which showed progressive downregulation at the later time points only (Figure 4.18c).


a
synapse organization
Enrichment distribution
0
-4
-8
-12
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
2e-05
4e-05
6e-05
8e-05
regulation of synapse organization
Enrichment distribution
0
-4
-8
-12
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
0.0040
0.0045
0.0050
0.0055
0.0060
synapse assembly
Enrichment distribution
0.0
-2.5
-5.0
-7.5
-10.0
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
0.01
0.02
regulation of synapse assembly
Enrichment distribution
0.0
-2.5
-5.0
-7.5
-10.0
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
0.01
0.02
0.03

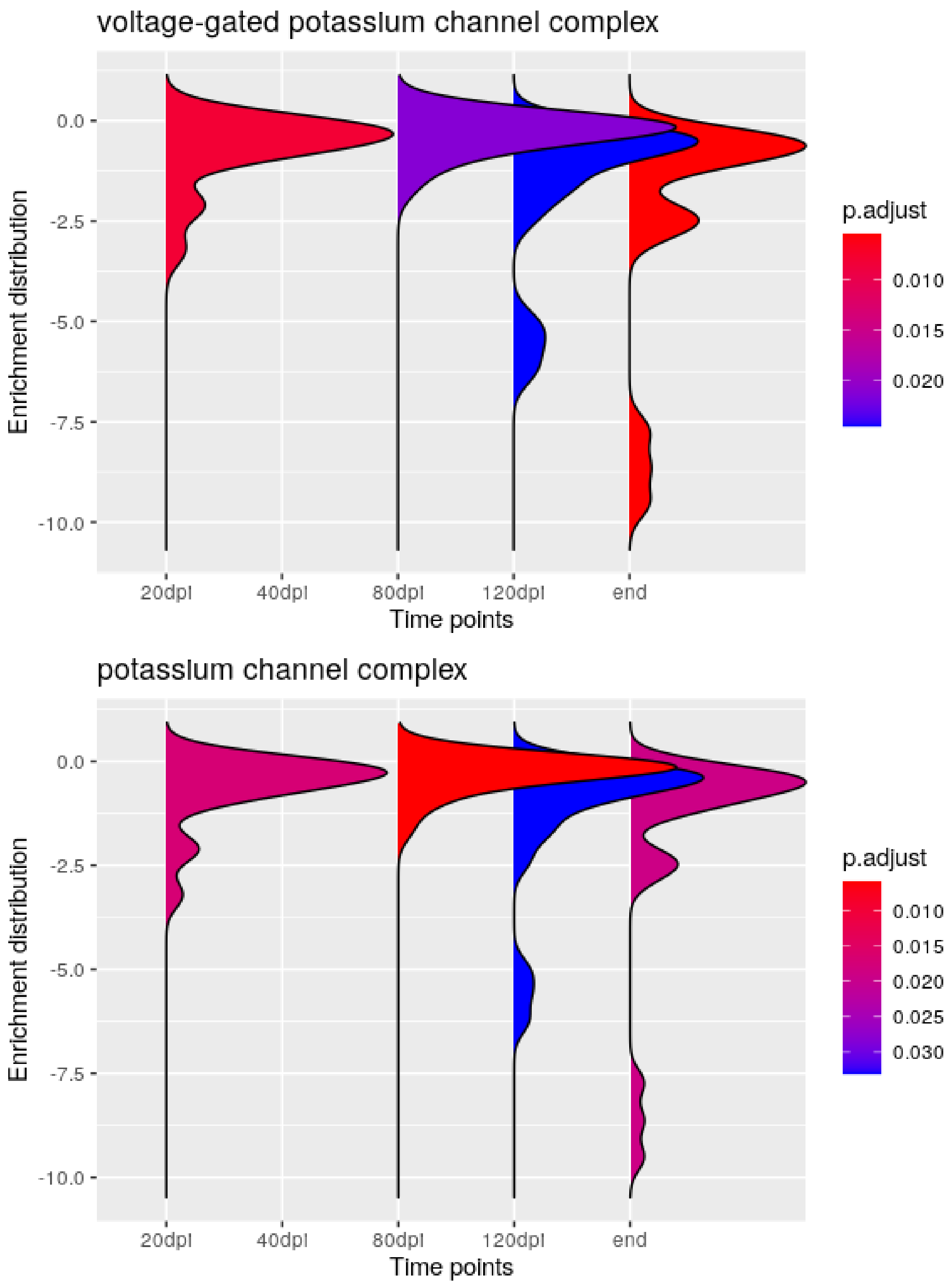

cell junction assembly
Enrichment distribution
0.0
-2.5
-5.0
-7.5
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
0.002
0.004
0.006
0.008


cell-cell adhesion via plasma-membrane adhesion molecules
Enrichment distribution
0.0
-2.5
-5.0
-7.5
-10.0
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
0.01
0.02
0.03


regulation of cell junction assembly
Enrichment distribution
0.0
-2.5
-5.0
-7.5
-10.0
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
0.01
0.02
0.03
0.04


regulation of cation transmembrane transport
Enrichment distribution
0.0
-2.5
-5.0
-7.5
-10.0
20dpi
40dpi
80dpi
120dpi
end
Time points
p.adjust
0.01
0.02

**b**



voltage-gated potassium channel complex

potassium channel complex

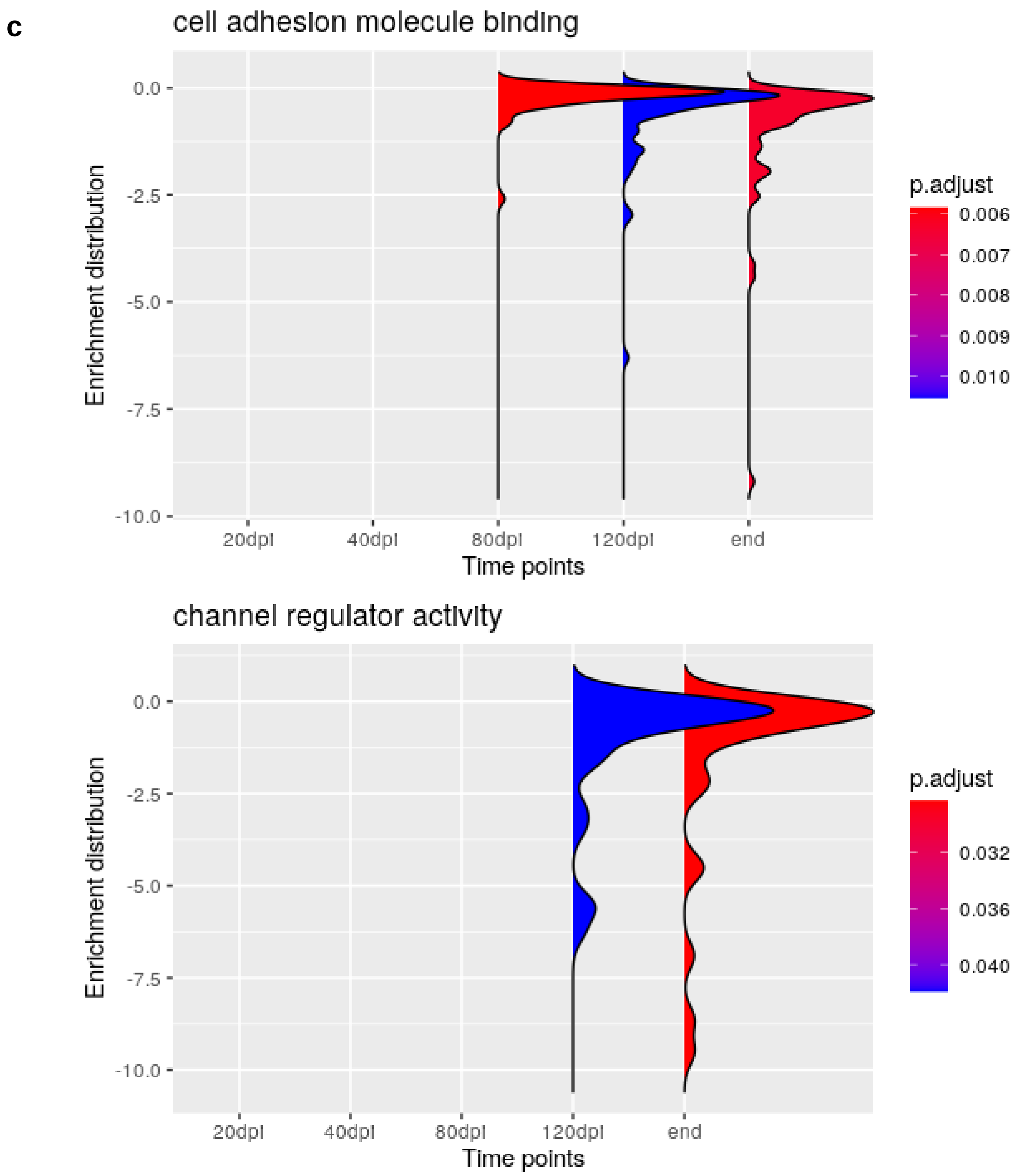


***Figure 4.18: Synapse, cell junction, cell adhesion, and ion homeostasis GO gene sets exhibited gradual dysregulation that follows disease progression.*** *Combining the ridge plots from multiple time points in a single plot allows us to visualise the gradual downregulation of specific gene sets using the* ***(a)*** *BP,* ***(b)*** *CC, and* ***(c)*** *MF classification.* ***(a)*** *Synapse assembly and organisation pathways, also underlined by previous analyses, were found to be progressively downregulated. The same trend followed the cell adhesion, cell junction and cation transmembrane transport gene sets. Interestingly, a similar pattern was observed for some of those gene sets, where there was enrichment at 20 dpi, which was followed by no enrichment at 40 dpi, then the gene set was enriched again at 80 dpi with an enrichment distribution similar*

*to that of the 20 dpi time point before the distribution shifted to the left at 120 dpi and even further at the end-stage.* ***(b)*** *Two closely related gene sets relevant to the potassium channel complex exhibited the same interesting pattern of progressive downregulation under the CC classification.* ***(c)*** *Two gene sets of the MF classification showed progressive downregulation at the later time points only. Each plot represents a single GO term. The x-axis represents disease progression (left to right). Time points without plots indicate that the specific GO term was not enriched at that specific time point.*

Overall, gene enrichment analyses allowed us to get an overview of dysregulated biological processes, as well as their cellular location and associated molecular functions. Initially, an over-representation analysis identified a global dysregulation of synaptic pathways, especially at the later time points. Cell adhesion and myelination pathways were found to be dysregulated in mature oligodendrocytes and oligodendrocyte precursor cells. In addition, we identified dysregulation in systems responsible for the homeostasis and transport of ions. The subsequent gene set enrichment analysis strengthened these findings and suggested synaptic dysregulation as early as 80 dpi. Importantly, the data suggest the existence of an early/late dysregulation signature that is manifested by enriched gene sets at 20 dpi, then disappears at 40 dpi, and then re-emerges at 80 or 120 dpi and becomes stronger at the end-stage.

The differential expression patterns and the enrichment analyses suggest a diverging transcriptomics profile between clusters of neurons, oligodendrocytes, and astrocytes. The following three sections will focus exclusively on these three broad cell types, aiming to dissect the transcriptomics of each population more finely.

### 4.2.5 Transcriptomics of neurons

Neurons represented the largest identified population in all our datasets, while they were also associated with the highest number of DEGs. Given that prion disease is a neurodegenerative disorder, we decided to first focus on the transcriptomics of neurons.

Based on the previous data, we decided to investigate the direction of differential expression of neuronal genes. Heatmap plots corroborated the findings of the enrichment analyses and suggested that most DEGs of neuronal clusters were, indeed, downregulated in disease (Figure 4.19). Starting at 20 dpi, we observed a diffuse pattern of differential expression characterised by small numbers of DEGs without obvious polarisation between up and downregulation in all neuronal clusters, except cortical neurons cluster 11, which showed a higher number of downregulated genes. As previously underlined, the 40 and 80 dpi time points had very small numbers of DEGs. Moving on to the 120 dpi time point, we saw an increase of downregulated DEGs, while the numbers of the upregulated DEGs remained low. Finally, at the end-stage, we observed a jump in the numbers of downregulated DEGs accompanied by a minute increase of upregulated genes. This discrepancy was so pronounced that for some clusters the numbers of downregulated genes were more than ten times higher than those of the upregulated genes. Interesting patterns also emerged when we tracked the trajectories of specific clusters throughout disease progression. Clusters 12, 13, 15, 18, 46, 47, and 50 consistently showed little evidence of dysregulation across all time points, which could be attributed to low numbers of cells and insufficient power of the analysis. Most clusters that displayed a strong DE signature at the end-stage, such as clusters 7, 9, 10, 14, 17, and 44, exhibited a gradual increase of downregulated DEGs which started from the 120 dpi time point. Some clusters showed a more abrupt increase in the numbers of DEGs at the end-stage, such as clusters 4 and 11. Interestingly, cluster 11, which had the higher number of DEGs at 20 dpi, did not exhibit substantial dysregulation until the end-stage. In addition, clusters 7 and 10 with relatively few cells had the highest number of DEGs at the end-stage, suggesting that the analysis is powered enough to identify these changes, and the lack of DEGs in other clusters with few cells might not be as a result of an underpowered study, and could instead reflect the underlying biology.

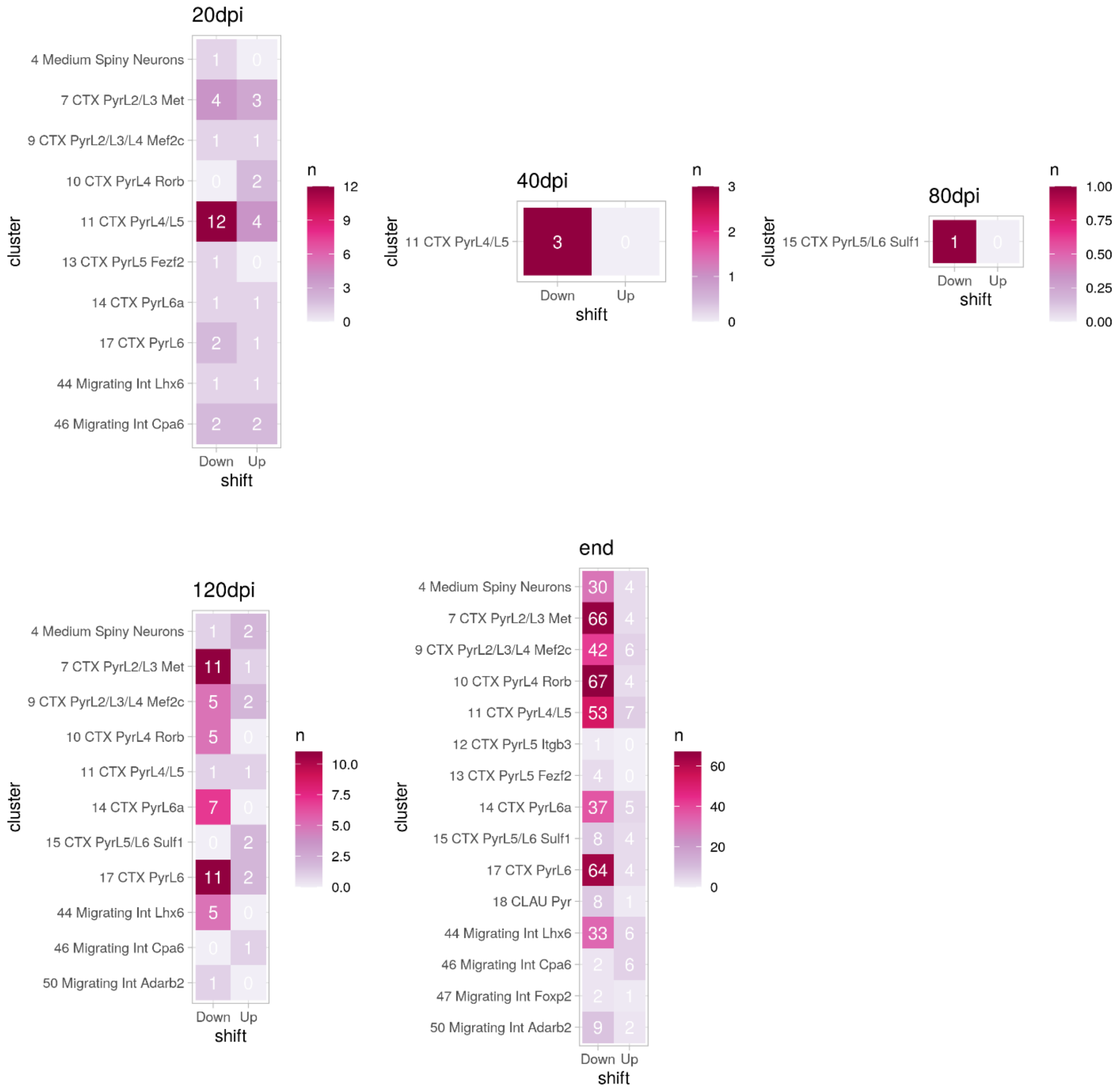
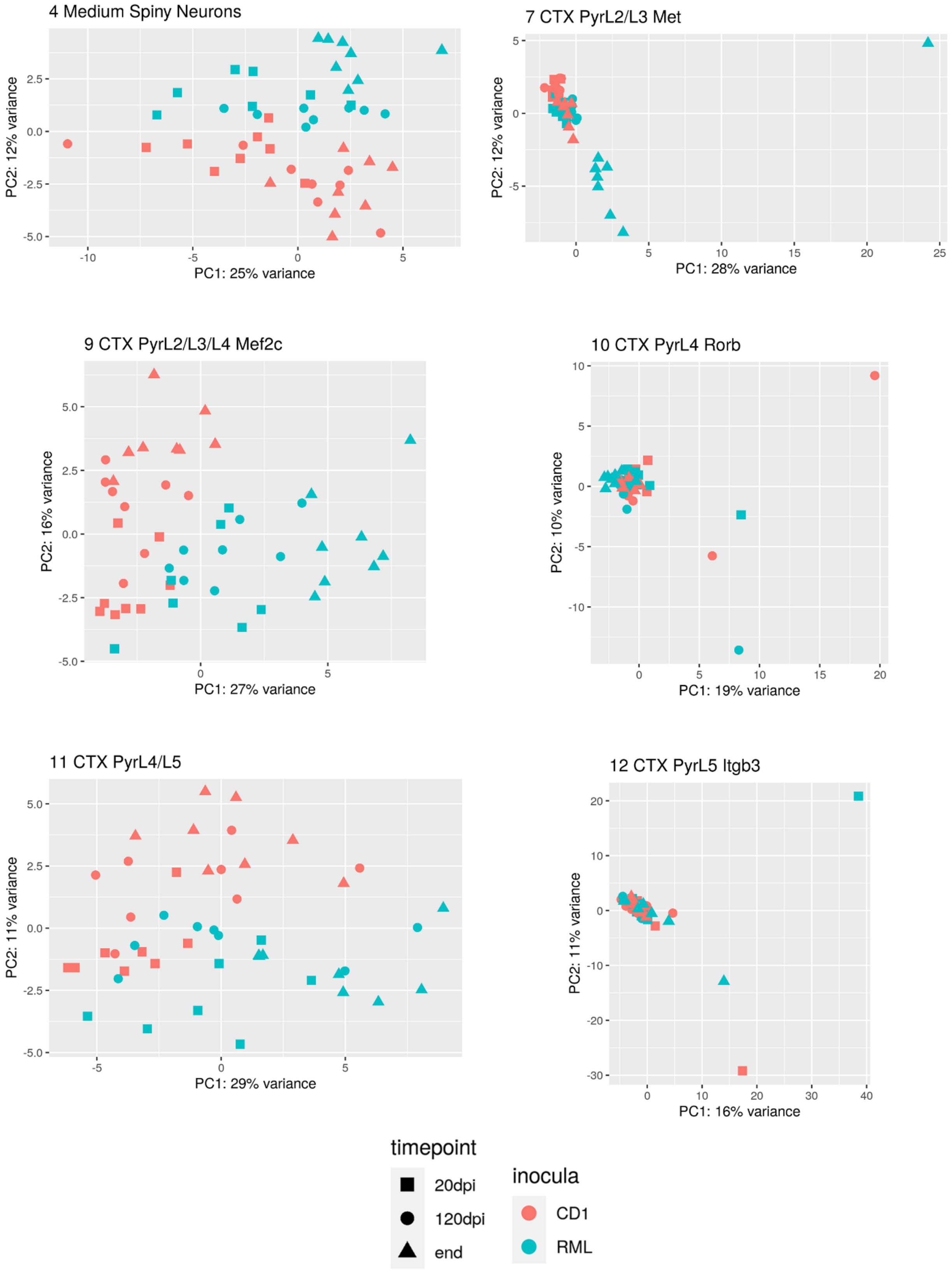
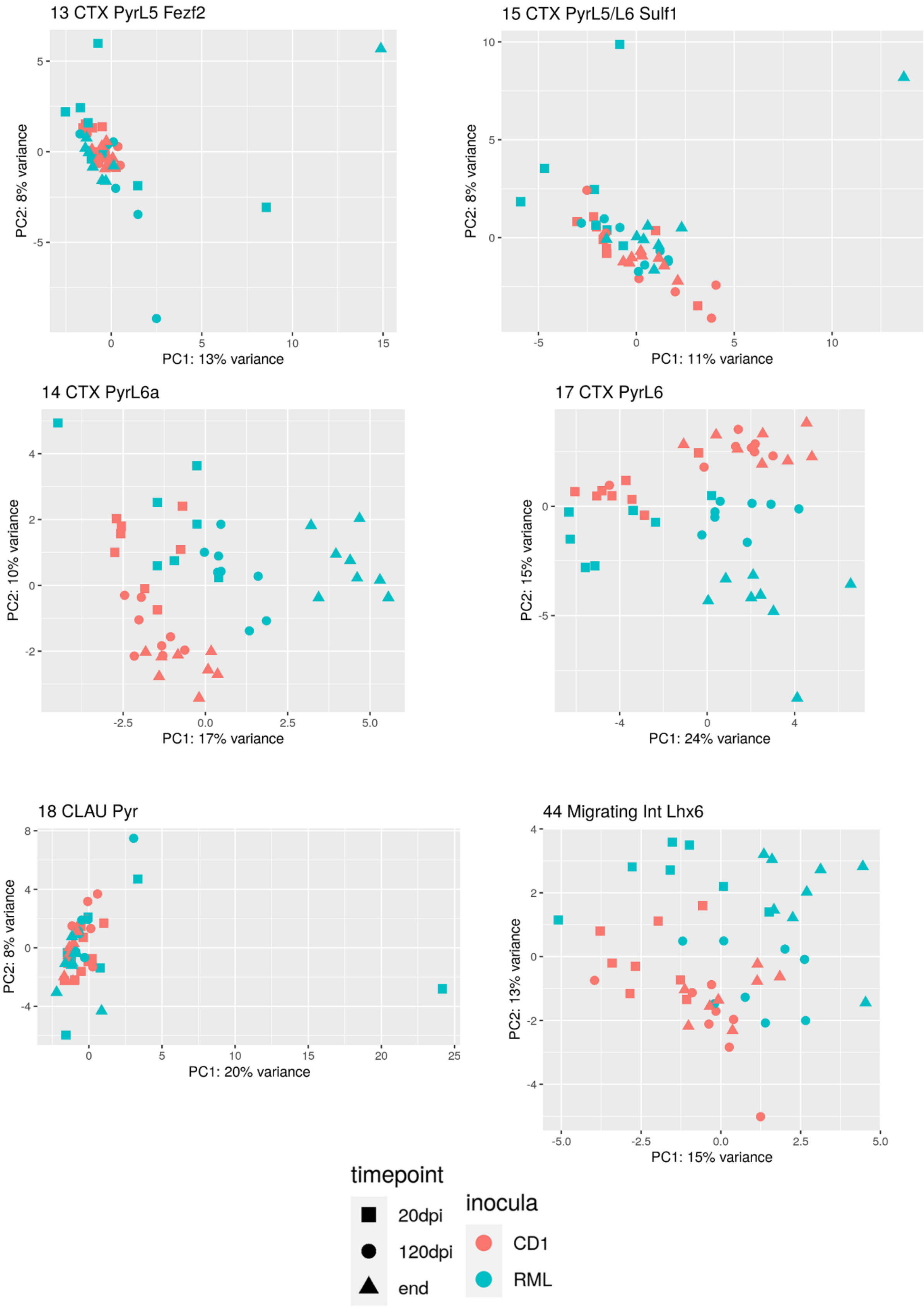


***Figure 4.19: Heatmap plots corroborated the findings of the previous enrichment analyses suggesting a downregulatory disease signature in most neuronal clusters.*** *We observed a diffuse pattern of differential expression at 20 dpi characterised by small numbers of DEGs without obvious polarisation between up and downregulation in all neuronal clusters, except cortical neurons cluster 11, which showed a higher number of downregulated genes. The 40 and 80 dpi time points had very small numbers of DEGs. We saw an increase of downregulated DEGs at the 120 dpi time point, while the numbers of the upregulated DEGs remained low. Finally, at the end-stage, we observed a jump in the numbers of downregulated DEGs accompanied by a minute increase of upregulated genes.*

We then investigated the sample distances in the two-dimensional space. We used the per-sample aggregated counts data to perform a PCA analysis of each neuronal cluster separately. We observed substantial differences between the patterns arising suggesting selective transcriptomic perturbation of neuronal subtypes (Figure 4.20). For some clusters, such as clusters 10, 12, and 13 we observed a tight clustering of all samples with no discrimination between the CD1 and RML groups. These clusters seem to not be affected transcriptionally by the disease. In contrast, the cluster of medium spiny neurons (cluster 4), clusters 9, 11, 14, and 17 of cortical neurons, and, to a lesser extent, cluster 44 of migrating interneurons all exhibited a pattern that allowed visual discrimination between disease and controls. Upon closer inspection of the plots for clusters 4, 9, 11, 14, 17, and 44 we identified an interesting and recurring pattern where while the samples associated with the RML group were spread out (teal points), the end-stage samples (teal triangles) were positioned the furthest away from the CD1 controls (red points). There was no evident discrimination between the 20 dpi (teal squares) and 120 dpi (teal circles) samples.


4 Medium Spiny Neurons
PC2: 12% variance
2.5
0.0
-2.5
-5.0
-10
-5
0
5
PC1: 25% variance
7 CTX PyrL2/L3 Met
PC2: 12% variance
5
0
-5
0
5
10
15
20
25
PC1: 28% variance
9 CTX PyrL2/L3/L4 Mef2c
PC2: 16% variance
5.0
2.5
0.0
-2.5
-5.0
0
5
PC1: 27% variance
10 CTX PyrL4 Rorb
PC2: 10% variance
10
5
0
-5
-10
0
5
10
15
20
PC1: 19% variance
11 CTX PyrL4/L5
PC2: 11% variance
5.0
2.5
0.0
-2.5
-5.0
-5
0
5
PC1: 29% variance
12 CTX PyrL5 Itgb3
PC2: 11% variance
20
10
0
-10
-20
-30
0
10
20
30
40
PC1: 16% variance
timepoint
■ 20dpi
● 120dpi
▲ end
inocula
● CD1
● RML


13 CTX PyrL5 Fezf2
PC2: 8% variance
5
0
-5
0
5
10
15
PC1: 13% variance
15 CTX PyrL5/L6 Sulf1
PC2: 8% variance
10
5
0
-5
0
5
10
PC1: 11% variance
14 CTX PyrL6a
PC2: 10% variance
4
2
0
-2
-2.5
0.0
2.5
5.0
PC1: 17% variance
17 CTX PyrL6
PC2: 15% variance
0
-5
-4
0
4
PC1: 24% variance
18 CLAU Pyr
PC2: 8% variance
8
4
0
-4
0
5
10
15
20
25
PC1: 20% variance
44 Migrating Int Lhx6
PC2: 13% variance
4
2
0
-2
-4
-5.0
-2.5
0.0
2.5
5.0
PC1: 15% variance
timepoint
■ 20dpi
● 120dpi
▲ end
inocula
● CD1
● RML

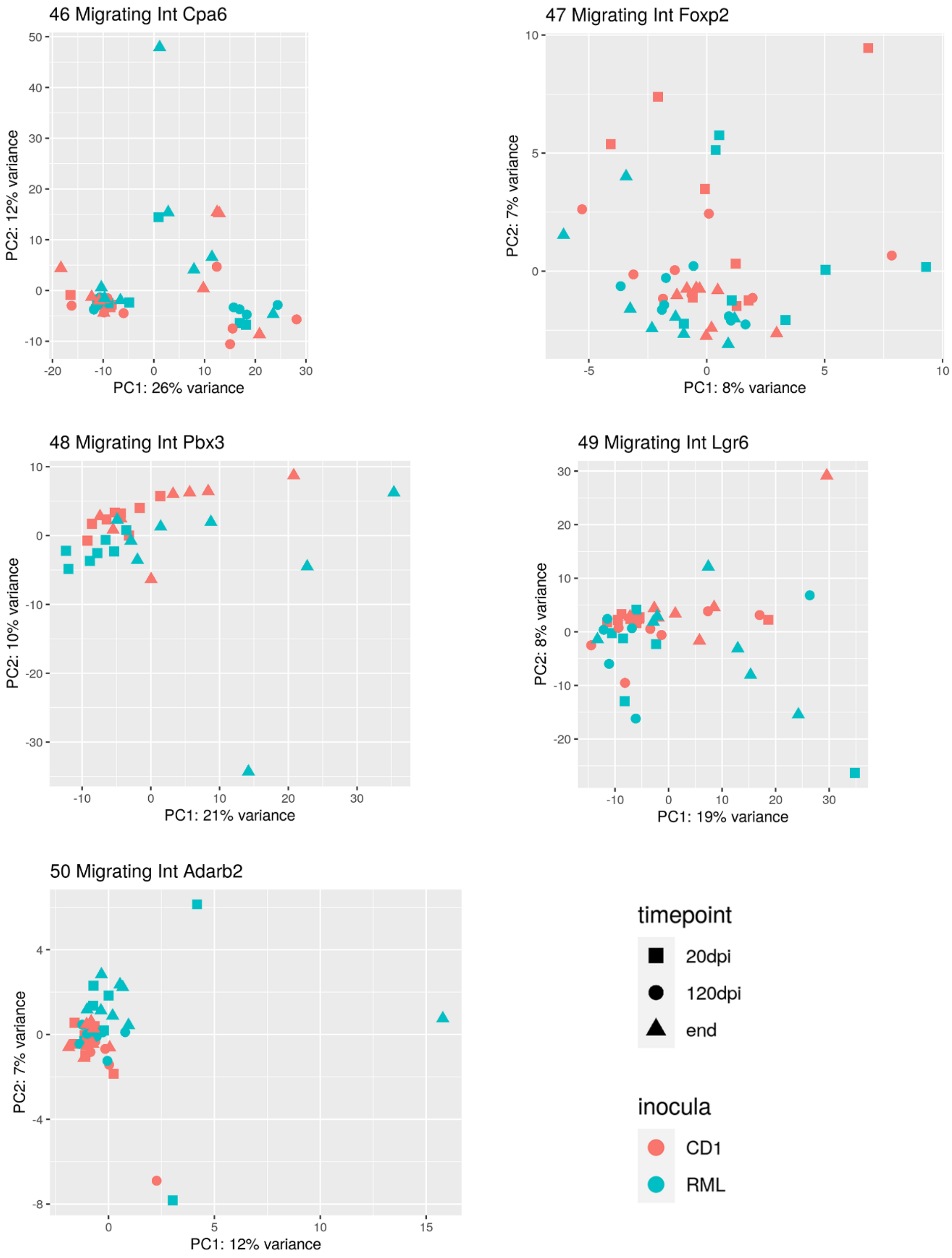


***Figure 4.20: Segregation patterns of PCA plots of 20 and 120 dpi and end-stage samples suggest selective transcriptomic perturbation of specific neuronal subtypes.*** *The plots visualise the first two principal components of a PCA performed on the per-sample aggregated counts of each neuronal cluster*

*separately. 40 dpi and 80 dpi time points are not shown due to a lack of interesting transcriptomic differences. A tight clustering of all samples with no discrimination between the CD1 and RML groups was observed for some clusters, such as clusters 10, 12, and 13. In contrast, the cluster of medium spiny neurons (cluster 4), clusters 9, 11, 14, and 17 of cortical neurons, and, to a lesser extent, cluster 44 of migrating interneurons all exhibited a pattern that allowed visual discrimination between disease and controls. An interesting and recurring pattern was identified when examining the plots of clusters 4, 9, 11, 14, 17, and 44 where while the samples associated with the RML group were spread out (teal points), the end-stage samples (teal triangles) were positioned the furthest away from the CD1 controls (red points). There was no evident discrimination between the 20 dpi (teal squares) and 120 dpi (teal circles) samples. The spurious sample originating from mouse 828719 has been removed from the plots because it was found to be at large distances from all other samples and changed the scale of the plots.*

### 4.2.6 Transcriptomics of astrocytes

Astrocytes also exhibited interesting transcriptional patterns and were one of the populations that were studied in more detail. Even though there were two astrocytic clusters, one of those, cluster 69, was much more abundant in cell numbers and showed interesting gene expression perturbations (Figure 4.21). When studying the direction of differential expression of the astrocytic clusters we observed a higher number of upregulated genes at the 20 dpi time point (5 upregulated genes versus 1 downregulated), then 1 and 4 upregulated genes at the 40 and 80 dpi time points, respectively, followed by an abrupt increase in the numbers of downregulated genes at the last two time points. At 120 dpi there was a sudden downregulation of 73 genes, while at the end-stage this number increased slightly to 91. At the same time, the number of upregulated genes remained approximately the same (16 at 120 dpi and 17 at the end-stage). These numbers were only relevant to cluster 69, while only one downregulated gene was associated with cluster 68.

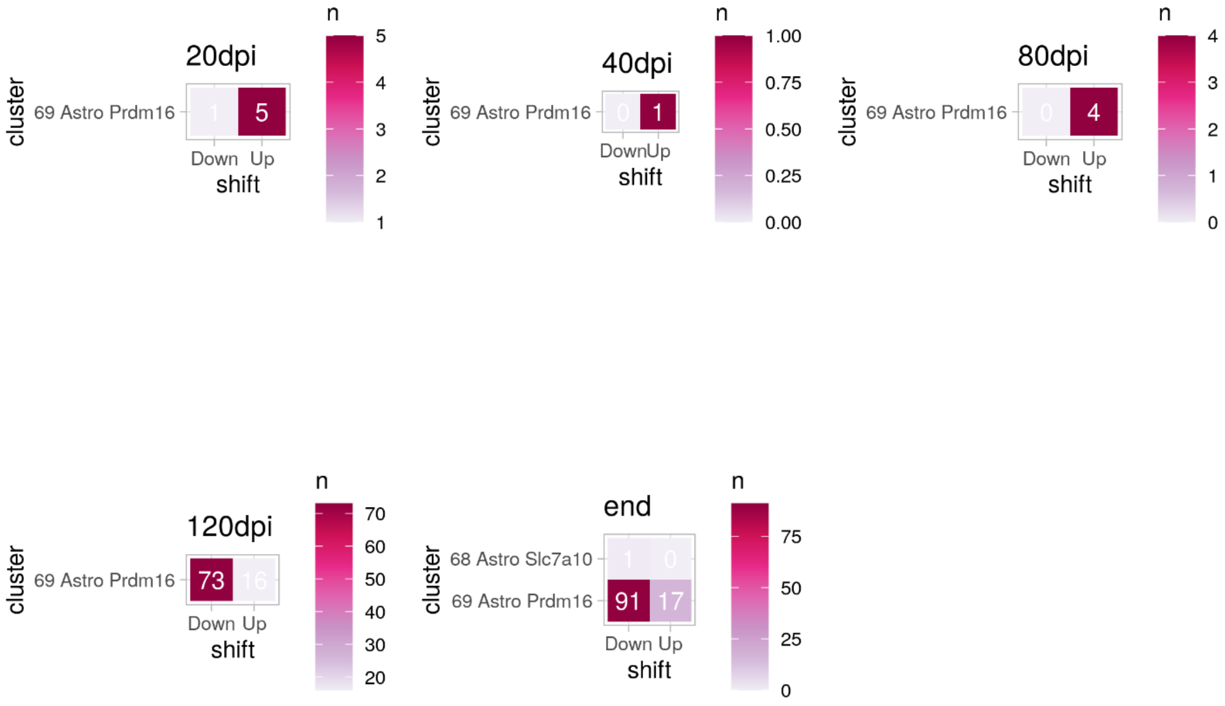


***Figure 4.21: A small number of astrocytic genes were upregulated at the early time points, while a strong downregulation of gene expression was observed at the last two time points of the experiment.*** *Cluster 69 of astrocytes was the most abundant in cell numbers and showed evidence of transcriptomic dysregulation, while cluster 68 only had 1 associated DEG at the end-stage. At 20 dpi only 6 genes were found to be DE, 1 being downregulated and 5 with increased levels. Then 1 and 4 genes were found to be upregulated at 40 and 80 dpi, respectively. A substantial down-regulatory trend was observed at the last two time points, with 73 and 91 genes exhibiting reduced levels of expression at 120 dpi and the end-stage, respectively. The number of upregulated genes remained approximately the same (16 at 120 dpi and 17 at the end-stage).*

We then generated per-sample aggregated datasets of the two clusters of astrocytes to investigate sample distances in the two-dimensional space. Following a PCA, the visualisations of the first two principal components suggest a diverging transcriptomic profile associated with astrocytes of cluster 69, while sample distances were short regarding cluster 68, apart from one outlier sample (Figure 4.22). Focusing on cluster 69, the first principal component nicely separated the CD1 and RML groups for the 120 dpi and end-stage time points. Interestingly, the right panel of Figure 4.22 suggests a gradual change in the transcriptomic landscape. Samples of the RML group and 20 dpi time point (teal squares) are separated from the control group (red points) by the second principal component, however, they are in similar positions on the x-axis (first principal component). Following disease progression, we observed a gradual shift of the RML samples towards the right of the x-axis with the samples at 120 dpi (teal circles) being further apart than the controls, and the samples at the experimental end-stage (teal triangles) exhibiting even larger distances from the CD1 group, suggesting an amplification of the transcriptomic perturbations.

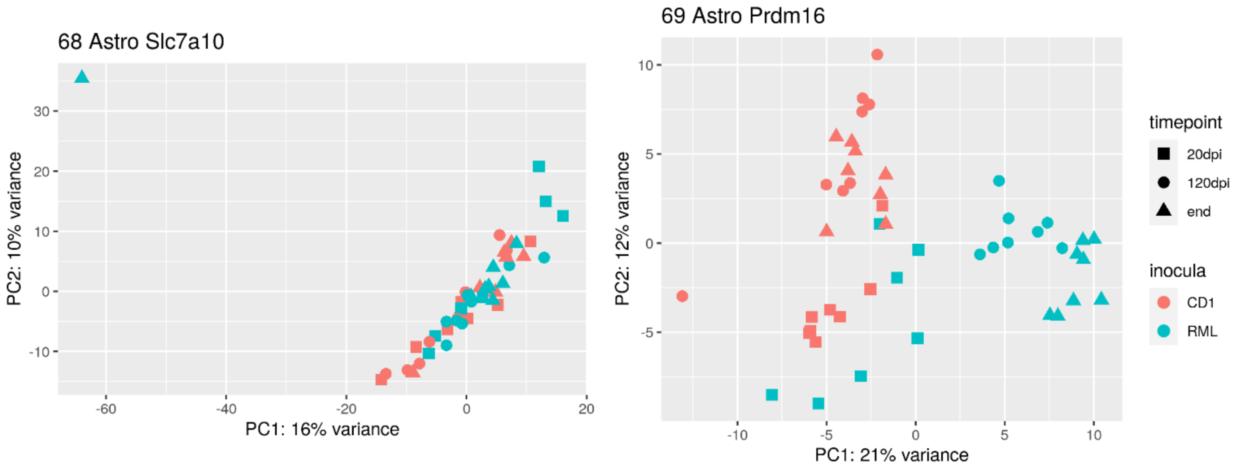


***Figure 4.22: A PCA plot of astrocyte cluster 69 suggests a gradual transcriptomic dysregulation during the last two time points, while cluster 68 astrocytes from both RML and CD1 groups cluster together.*** *The plots show the first 2 principal components of PCA performed on the per-sample aggregated gene counts. Most of the samples associated with cluster 68 cluster together in the low-dimensional space suggesting little transcriptomic difference, apart from one outlier sample. Astrocytes belonging to cluster 69 exhibited an interesting pattern of gradual perturbations in gene expression during the disease progression. While samples from the RML 20 dpi time point (teal squares) were localised in similar coordinates on the x-axis as the control samples (red points), there was a noticeable shift towards the higher values of the x-axis associated with the samples of the 120 dpi time point (teal circles), which was further amplified at the end-stage (teal triangles), suggesting a continuum of transcriptomic dysregulation that follows disease progression. The spurious sample originating from mouse 828719 has been removed from the plots because it was found to be at large distances from all other samples and changed the scale of the plots.*

To identify activated astrocyte populations and classify them based on the A1/A2 classification, we performed real-time quantitative PCR analysis on N = 4 biologically independent samples per time point and inoculum. The material used was bulk brain nuclei suspension, which included astrocytes among other cell populations. We assayed 5 A1 signature (*C3, Fkbp5, Gbp2, Ggta1, Serping1*), 3 A2 signature (*Cd109, S100a10, Tm4sf1*) and 2 pan-astrocytic activation genes (*Hspb1, Vim*). Both pan-astrocyte signature genes were found to be significantly upregulated in the last time point and *Vim* was also found to be significantly upregulated at 20 dpi, suggesting astrocyte activation. 2/5 A1 signature genes were significantly upregulated at the last time point (*Fkbp5, Ggta1*), suggesting the existence of A1 astrocytes. *Cd109* A2 signature gene was significantly upregulated at the disease end-stage, while *Tm4sf1* was significantly downregulated at 80 dpi and the end-stage (Supplementary Figure 12).

Another classical marker that can be used to quantify the presence of astrocytes is *Gfap* (Yang & Wang, 2015). We quantified astrogliosis by performing RNAscope on fixed mouse brain slices to visualise *Gfap* expression in all brain regions throughout disease progression. Our differential expression analysis had already identified *Gfap* to be one of the transcripts with the highest increase in expression in disease, and RNAscope data confirmed this finding (Figure 4.23 and Figure 4.24). *Gfap* levels were found not to increase in control (CD1-inoculated) mice during ageing, while there was a statistically significant decrease of *Gfap* expression in the hippocampus (N = 3 independent biological replicates per time point; two-way ANOVA; Sidak's multiple comparisons test). In contrast, its levels were visually elevated starting at 80 dpi, increasingly affecting all brain regions throughout disease progression. Quantification of the percentage of positive pixels in each anatomical area which corresponds to transcript expression suggested a statistically significant increase of *Gfap* expression in all brain regions at 120 dpi and the end-stage in RML-inoculated animals, compared to the first two time points, 20 and 40 dpi (N = 3 independent biological replicates per time point; two-way ANOVA; Sidak's multiple comparisons test). For the hippocampus, thalamus and brain stem, this significant increase was evident as early as 80 dpi. These results suggest generalised astrogliosis and are in accordance with our transcriptomic findings and published literature (Manuelidis et al., 1987).

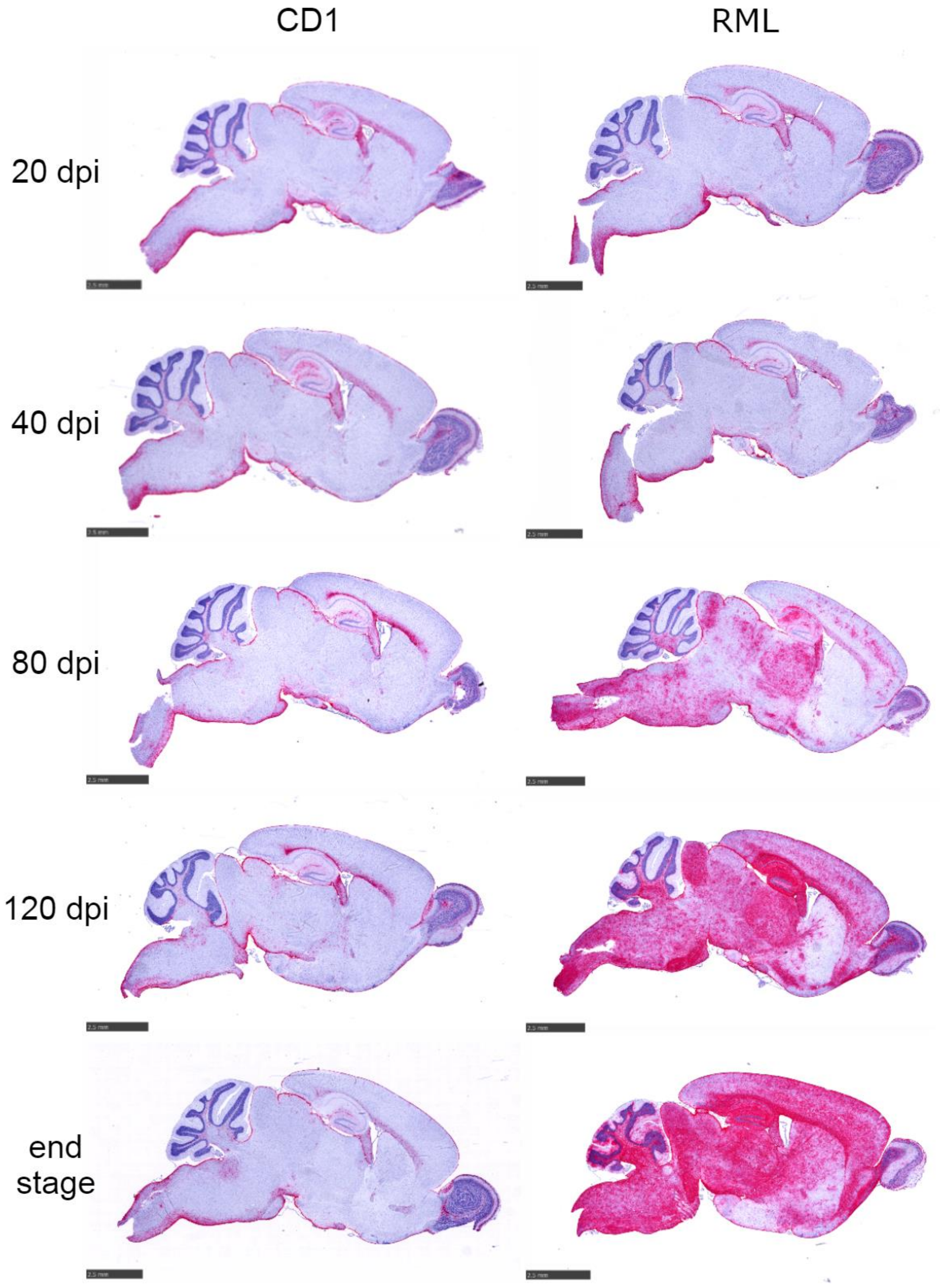
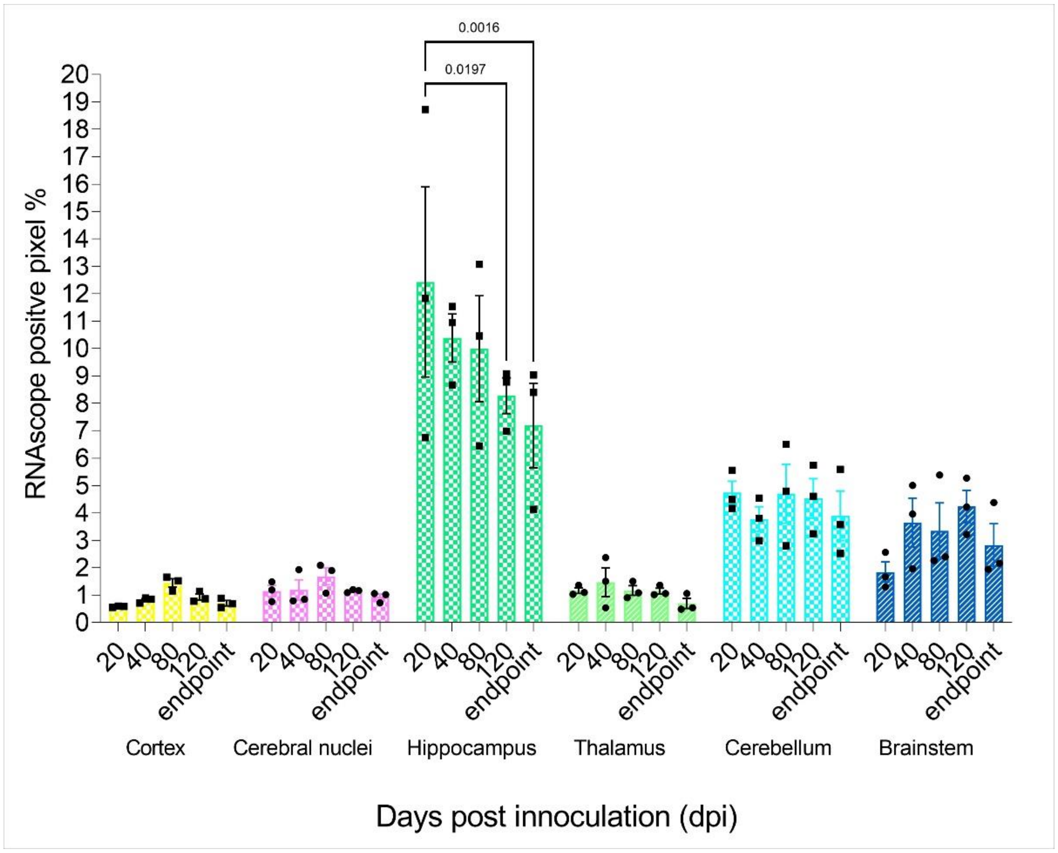

CD1
RML
20 dpi
40 dpi
80 dpi
120 dpi
end stage

***Figure 4.23: Representative images of stained mouse brains using RNAscope showing increased Gfap expression in RML-inoculated mice during the later time points.*** *Fixed mouse brains were sliced, processed using the RNAscope protocol and probed for Gfap expression. Each column corresponds to a different experimental group (left: CD1-inoculated mice, right: RML-inoculated mice), while each row corresponds to a different time point. Gfap transcript abundance is depicted in red, while nuclei are depicted in blue. Red staining at the edges of the tissue is a known artefact of the methodology and does not correspond to gene transcripts. Gfap levels were shown not to increase in control (CD1-inoculated) mice during ageing. In contrast, Gfap levels were visually elevated starting at 80 dpi, increasingly affecting all brain regions throughout disease progression. The black scale bars correspond to a length of 2.5 mm.*


0.0016
0.0197
RNAscope positve pixel %
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
20
40
80
120
endpoint
Cortex
Cerebral nuclei
Hippocampus
Thalamus
Cerebellum
Brainstem
Days post innoculation (dpi)

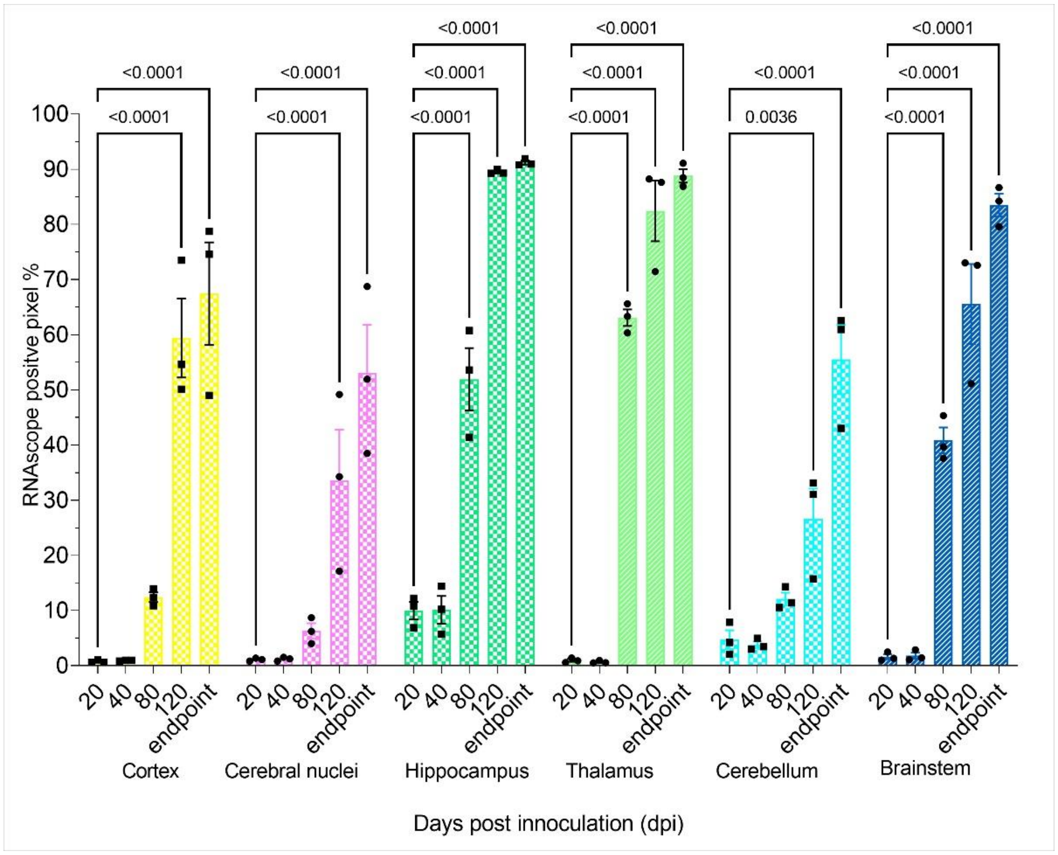

<0.0001
<0.0001
<0.0001
<0.0001
0.0036
<0.0001
RNAscope positve pixel %
0
10
20
30
40
50
60
70
80
90
100
20
40
80
120
endpoint
Cortex
Cerebral nuclei
Hippocampus
Thalamus
Cerebellum
Brainstem
Days post innoculation (dpi)

***Figure 4.24: Quantification of the RNAscope signal suggested an increase of Gfap expression in all brain regions throughout disease progression in RML-inoculated mice.*** *Gfap levels were shown to remain stable in control mice, except in the hippocampus, where Gfap levels decreased with ageing. In contrast, all brain regions of the CD1-inoculated mice showed a statistically significant increase of Gfap expression in the later time points (120 dpi and end-stage, and 80 dpi for the areas of the hippocampus, thalamus, and brain stem), compared to the first two (20 and 40 dpi). A Shapiro-Wilks Normality Test was performed to ensure normality before calculating p-values using a two-way ANOVA. P-values were corrected using Sidak's multiple comparisons test. Numbers on top of the bars represent the calculated p-values. N = 3 independent biological replicates per inoculum per time point. Points represent biological replicates.*

### 4.2.7 Transcriptomics of mature oligodendrocytes and oligodendrocyte precursor cells

The two final cell populations that were studied in more detail comprised mature oligodendrocytes and oligodendrocyte precursor cells. Transcriptomic perturbations of OPCs (cluster 61) were only evident during the last two time points (Figure 4.25). Interestingly, at 120 dpi, most of the DEGs associated with OPCs were upregulated (11 upregulated versus 1 downregulated), while this trend was reversed at the last time point, where only 7 genes were found to be upregulated and 31 downregulated. OPCs did not have a strong signature of differential expression at the 20 dpi time point, as only one DEG was identified. In contrast, transcriptomic differences were observed in mature oligodendrocytes as early as 20 dpi. 8 of the DEGs were found to be downregulated, while 4 upregulated. This signature was lost at 40 and 80 dpi and re-emerged at 120 dpi, where 15 genes were found to be differentially expressed, with most of them being downregulated. The trend continued until the end-stage when 41 DEGs were downregulated and 12 upregulated.

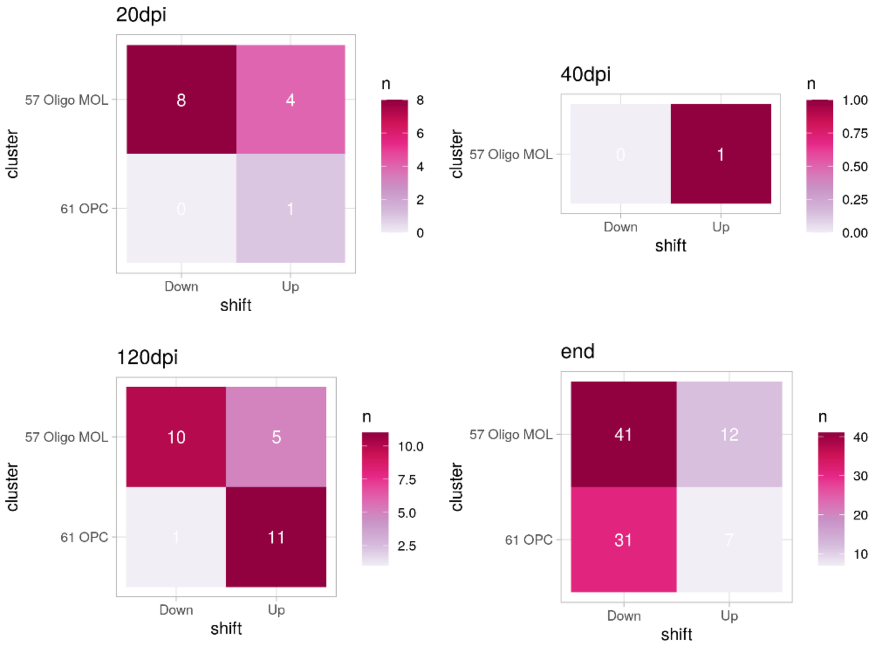


***Figure 4.25: Transcriptomic perturbations associated with mature oligodendrocytes became evident as early as 20 dpi, while the transcriptomic landscape of OPCs only began to change at 120 dpi.*** *Mature oligodendrocytes exhibited a downregulatory signature at 20 dpi, which was then absent at 40 and 80 dpi and re-emerged at 120 dpi and was amplified at the end-stage. 11 upregulated genes were associated with OPCs at the 120 dpi time point, while the trend reversed at the end-stage with most DEGs being downregulated. The 80 dpi time point is not present in this figure because no DEGs existed for these cell populations.*

We then assayed the distances between biological samples in the low-dimensional space by performing a PCA on the per-sample aggregated gene counts. Only small segregation of the samples from the two experimental groups was observed for the mature oligodendrocytes, which was evident only for the 120 dpi and end-stage and was based on the first principal component (Figure 4.26). Regarding the OPCs, there was no clear separation, with most of the samples occupying the same area in the two-dimensional space. Overall, the PCA plots suggest more pronounced transcriptomic differences associated with the mature oligodendrocytes at the 120 dpi and end-stage time points.

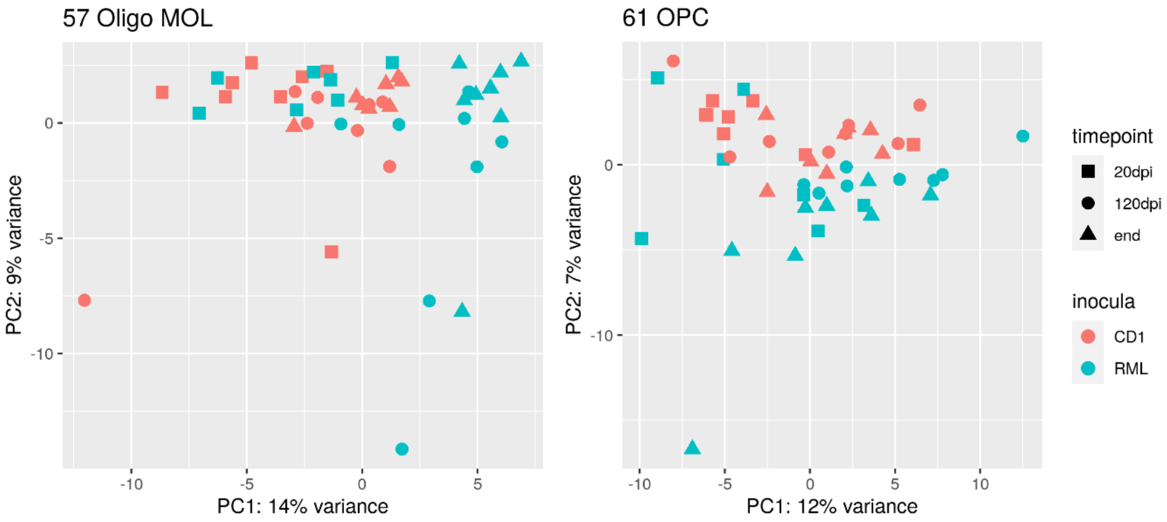


***Figure 4.26: A PCA of the per-sample aggregated counts suggested only small transcriptomic differences between all the OPC samples, while more pronounced perturbations were associated with the mature oligodendrocytes at the 120 dpi and end-stage time points.***

## 4.3 Discussion

### 4.3.1 Experimental design and pathophysiological characterisation of animal samples

Following the successful validation of our experimental protocol as described in the previous chapter, we proceeded to study murine prion disease under tightly controlled experimental settings. Prion disease is associated with a progressing pathology and most scientific studies have focused on the later stages of its course. This rapid nature of the disease could reflect underlying transcriptomic dynamicity, which would only be observable by monitoring the experimental model through time. In addition, pivotal work by Sandberg and colleagues suggested the existence of two distinct mechanisms of prion propagation that manifest during the early and late stages of the disease (Sandberg et al., 2011). We designed a time-course experiment that would allow us to investigate transcriptomic alterations during the mechanistic shift of prion propagation and query the earliest stages of the disease. In addition, the nature of our study would require a system that can faithfully and accurately recapitulate previous findings, so that we minimise the uncertainty associated with the pathophysiology and natural history of the model and focus on transcriptomics instead. Finally, the model selected would have to be devoid of genetic manipulation that can introduce artefacts and hinder the interpretation and generalisation of the results. After careful consideration of those factors we decided to proceed with a model system that comprised wild-type FVB mice inoculated with RML prions; a system that had been extensively studied in the past and was well-characterised and suitable for the needs of our study (Sandberg et al., 2011, 2014). Further discussions with M. Sandberg led to the selection of 5 time points when samples would be collected. The selection was purposefully made to include representative samples from the different mechanistic stages of the disease: the 20 dpi time point would provide insight into the earliest disease state when infectivity is low, the 40 dpi time point would be located in the exponential phase, the 80 dpi time point would be located just after the mechanistic shift, at the beginning of the plateau phase, followed by the 120 dpi time point in the plateau phase before the onset of symptoms. The last time point would be at the onset of symptoms when scrapie sickness was confirmed and would designate the disease end-stage. Since adequate controlling of the experiment is of paramount importance, we decided to include two control groups. One of the groups of mice would be inoculated with sterile PBS only, and the other with uninfected CD1 brain homogenate, which is also the same homogenate used to dilute the RML inoculum. The main comparison would be performed between the RML and CD1 groups, however, a comparison between CD1 and PBS groups could also be used to identify technical noise or transcriptomic variation relevant to the introduction of external brain homogenate and the process of the intracerebral inoculation, which could later be removed from the data. To exclude variation relevant to the normal ageing of the mice, we inoculated mice of approximately the same age and cullings from each group were performed on the same day. The experimental groups would include 15 mice for the RML and CD1 groups so that enough mice would survive to the experimental end-stage and enough samples would be available for the assays planned (snRNA-seq, IHC, infectivity titration etc.). The PBS group would be smaller comprising 5 mice since inoculum toxicity was expected to be low and fewer samples would be required from this group (only for snRNA-seq). Overall, the number of time points and mice in each time point was selected as a balance between collecting the appropriate number of samples to reliably generate adequate datasets based on previous studies, and the need to reduce the number of laboratory animals used and align with the ethical principles of the 3Rs (Replace, Reduce, Refine).

Subsequent to the conclusion of the animal experiments, brain samples were assayed to ensure disease progression and prion infectivity were comparable to previous studies and selected time points corresponded to the theoretical disease course. To that end, we performed immunohistochemical analyses (standard pathology staining with H&E and anti-PrP antibodies) and infectivity titration using the SCA. Expert review of the ICH staining confirmed a match to expected disease progression, while the infectivity plots recapitulated the two phases of infectivity. All evidence suggested that time point selection had been successful and did satisfy our experimental design criteria, so we shifted our focus to the transcriptomic studies.

### 4.3.2 Single-cell transcriptomics of murine prion disease

We set out to perform an unbiased whole-transcriptome single-cell study of murine prion disease using the previously generated brain samples. We decided to only focus on a specific anatomical area to minimise tissue heterogeneity and selected the frontal lobe as it is heavily affected in disease. For the selection of the number of nuclei to be sequenced, we based our calculations on previous studies (Mathys et al., 2019; Rosenberg et al., 2018) and a consideration of the balance between the number of nuclei sequenced and the depth of sequencing of each nucleus. Taking into consideration the technical limitations of the SPLiT-seq protocol and the results of preliminary test experiments, we aimed at sequencing 45,000 nuclei per time point, for a total of 225,000 total nuclei sequenced across all samples. The distribution of samples in each time point was designed to prioritise the RML and CD1 groups (N = 8 per group), which would the main comparison, versus the PBS group (N = 3), which would only be used as an additional control. The experimental design also mandated the parallel processing and sequencing of the RML and CD1 groups of each time point to minimise batch effects.

We generated nuclei suspensions from 95 samples in total and sequenced the 5 time points in 5 separate batches. The 20 dpi time point was the first one sequenced and included fewer cells in the final library due to an unexpected nuclei loss. Nevertheless, all libraries resulted in similar numbers of sequencing reads and following pre-processing, and a similar number of identified nuclei. Sequencing quality was consistently high for all libraries and no filtering of the raw sequencing data was deemed necessary. In contrast, stringent filtering of two stages was applied to the feature count matrices. In the first instance, the splitseq-tools algorithm automatically identified and discarded cells with a low amount of information associated with background noise resulting from ambient RNA. This is usually performed by identifying the negative concavity coordinates, or “knees”, of a “knee plot”, which represents the sorted number of cell barcodes (x-axis) vs the number of UMI counts detected per cell barcode (y-axis). The plot is expected to contain two such knees, and the mid-point of the first knee is usually used as a cut-off to differentiate real cells from the background. In some cases, when the slope is not pronounced enough, the algorithm can omit the first knee of the plot and identify the second one, substantially inflating the number of identified cells and decreasing the mean number of UMIs per cell (representative examples can be found in Supplementary Figure 13). This was the case with the 20 dpi time point data, where the number of cells identified far exceeded the number of cells in the input material. These numbers included low-quality nuclei that contained high proportions of background noise (ambient RNA) due to the failure of the algorithm in some samples. Fortunately, our second filtering round could easily remedy this phenomenon, as the filtering criteria were manually selected and were equal for all datasets. We based our filtering on published studies and empirical data and decided to be as stringent as possible, sacrificing some of the available information for a higher-quality dataset overall. This resulted in approximately 200,000 transcriptomes after filtering with a median of 943 and a mean of 1050 features/nucleus, which is not as high as typical single-cell high-throughput sequencing (usually a mean of around 2,000-3,000 features/cell for commercial methods), however, this is justified because of the nature of the starting material. Here we have used single-nucleus sequencing (instead of single-cell); the nuclei include much lower amounts of available RNA and single-nucleus methods typically generate less than 2,000 features/nucleus, even on commercial platforms. Interestingly, our dataset was more information-rich than the original SPLiT-seq study (Mdn = 677 features/nucleus for Rosenberg et al., 2018), or some studies using commercial solutions (Thrupp et al., 2020 identified a mean of 879 features/nucleus using the 10X Genomics single-cell gene expression v2 kit) while being comparable to or lower than other studies (Nagy et al., 2020 had a mean of 2,144 features in neurons and 1,144 features in glia). Overall, the dataset quality was on par with published research and as expected by the methodology used.

The next step in our analysis pipeline was to annotate cell identities, which usually requires prior unsupervised clustering. While clustering is of paramount importance for single-cell studies as it is used to identify putative cell types, it poses significant technical, biological, and computational challenges, while no consensus exists in the scientific community regarding standard methodological practices (Kiselev et al., 2019). The infancy of the single-cell transcriptomics field combined with the plethora of laboratory protocols, computational tools, and algorithms complicates the selection of appropriate approaches and makes the matter more of a subjective choice based on previous experience or adequately satisfying results. The broad aim of clustering is to discover the natural groupings of objects, and, when applied to transcriptomic studies, provide an unbiased approach that can — in theory — categorise different cell types based on their gene expression profile (Jain, 2010). In reality, technical (low initial amounts of RNA which leads to a high dropout rate, batch effects, the presence of ambient RNA or cell doublets), biological (tissue heterogeneity, transient biological states), and computational challenges (high levels of dropout and noise, increasing scale of transcriptomic data, manual selection of algorithm parameters) can hinder biological interpretation. At the same time, cell-type annotation would still require manual review of highly expressed transcripts the matching of this information to previous studies and published literature, a subjective process that requires decision making, therefore impeding automation, and lowering reproducibility. Nevertheless, the process has become more efficient with the introduction of new tools that can streamline data analysis and the availability of accumulated knowledge in the form of cell atlases.

Our study is characterised by additional complexity as it included multiple time points which were sequenced independently and essentially constitute individual sub-datasets. The first apparent challenge was related to the size of the combined dataset, which made a collective analysis impractical. Additionally, each time point was expected to represent a different stage in the disease, so biological heterogeneity was anticipated. Finally, to track the disease transcriptomics between different time points, cell clustering and annotation would have to be consistent in all sub-datasets, which severely limited the suitability of manual methods. Even though we tried using the more traditional approach of unsupervised cell clustering specifying a range of different parameters to the clustering algorithms of Seurat, and then manually attempting to annotate the data based on the expression of the most characteristic genes for each cluster and using reference cell atlases, we quickly realised that this method was impossible to automate and reliably reproduce the same clusters across all time points to facilitate data interpretation and comparisons (data not shown). In order to overcome these difficulties, we decided to proceed with a data-driven approach of cluster annotation which was first introduced in version 3 of Seurat (Stuart et al., 2019). This strategy based on "anchoring" datasets together allowed us to collectively analyse diseased and control biological states, transfer cluster labels from thoroughly annotated reference datasets, and more importantly, enabled the comparison of all time points cluster-by-cluster, as it minimised subjective decisions regarding cluster annotation, ultimately leading to increased reproducibility and automation. Cluster annotation was based on a label transfer algorithm that can effectively match query populations to annotated reference datasets. Since this matching is based on transcriptomic information, the algorithm can perform best when the reference and query datasets contain the same cell populations and are generated in the same manner. We selected the mouse brain dataset generated by Rosenberg et al. (2018) to be used as a reference since it was produced using the same single-nucleus method (SPLiT-seq), it included the same tissue, and it was thoroughly annotated. A caveat of this approach was that the reference data was generated from very young mice (postnatal days 2 and 11), compared to our adult mice; however, we could not find evidence that this negatively affected the downstream analysis. Some preprocessing of the dataset was essential to increase the concordance between the two datasets, such as keeping only the cells from the frontal lobe and olfactory bulb (in case it was not completely removed during dissection) and merging the data from the postnatal day 2 and 11. After label transfer, the few cells putatively originating from the olfactory bulb were filtered out, while after discussions with external advisors we decided not to set any additional filters. Dimensionality reduction was performed independently from annotation (which was conducted on a per-cell level), and UMAP plots layered with transferred cluster information and labels reassuringly demonstrated that cluster identities corresponded well with the visual separation of clusters. Finally, another line of reassuring evidence was provided by quantifying the expression levels of known marker genes for broad cell types which corroborated cluster identities.

This data-driven approach identified a maximum of 25 different clusters (some time points had a lower number) that mostly comprised neuronal sub-clusters, astrocytes, and oligodendrocytes. Clusters of ependymal, endothelial, vascular, and leptomeningeal cells included only small numbers of cells (as expected) and did not show any interesting transcriptomic differences, possibly due to their low abundance, so they were not the focus of this study. Microglia, even though their relevance to disease is appreciated, are commonly found in small numbers in the brain, and were identified in only the last two time points in small numbers, so no meaningful information could be extracted (although this could indicate that their numbers are increased in the later stages of the disease. However, this hypothesis was not further investigated). In addition, evidence suggests that single-nucleus studies are not well suited for the investigation of microglia transcriptomics in disease, since technical bias leads to depletion of a small set of genes that are enriched for microglial activation markers (Thrupp et al., 2020). Studies focusing on microglia necessitate the use of population enrichment protocols and single-cell sequencing.

Following consistent cell annotation across all time points, we set out to investigate the hypothesis that specific neuronal sub-clusters are more vulnerable to the toxic effects of the disease (selective toxicity) leading to a more pronounced decrease in their numbers. Neurons are the only cells that are known to be led to cell death due to prion infection, while glial cells might replicate prions but do not suffer toxicity (Krejciova et al., 2017; Lakkaraju et al., 2021; Prinz et al., 2004). The pathophysiological hallmarks of prion disease include neuronal loss and gliosis, so we expected to see a decrease in the number of neurons and an increase in the abundance of astrocytes, oligodendrocytes, and microglia (even though microglia numbers were too low to provide sufficient information, as previously discussed). To get an overview of the behaviour of broadly defined populations, we grouped cells into 10 broad groups (migrating interneurons, cortical neurons, medium spiny neurons, astrocytes, OPCs, oligodendrocytes, VLMCs, ependymal, immune, and vascular cells) and calculated their relative proportions in each of the 5 time points. This approach precludes the use of canonical statistical tests, such as the t-test, since they are not designed for relative proportion data, and, more importantly, does not consider the effects of sampling variation between the different biological replicates (Aitchison, 2008). Therefore, we employed a permutation test to calculate p-values for each cluster and confidence intervals for the magnitude of the difference via bootstrapping. This analysis gave inconsistent results, with numbers of migrating interneurons increasing and medium spiny neurons increasing at the 20 dpi time point. The same picture was evident at the last time point (end-stage). Oligodendrocytes were found to be reduced at the last time point only, while OPCs were reduced at the first and last time points. Astrocytes were found to be reduced at the first time point and then increased at 120 dpi, with their levels not affected at the end-stage (even though astrocytosis is expected in prion disease). Immune cells (microglia) displayed a more consistent trajectory, being increased in numbers in the last two time points.

Next, we focused specifically on neurons, where the changes in cell proportions were relatively small. Most neuronal clusters showed to be reduced in numbers, however, this reduction was evident from the first time point, after which the abundance of neuronal cell types became comparable again between disease and controls. A more pronounced decrease in numbers was once more evident at the end-stage. Interestingly, the 20 dpi and end-stage time points were similar in both broad populations and neuronal sub-clusters, while neuronal cell reduction was not validated at 20 dpi by histopathology. This suggested that identified differences in the cell populations are more likely attributed to underlying transcriptomic changes that affect cell cluster determination. The number of cells in each cluster requires a prior cluster annotation, which can be inconsistent between time points, even when automated data-driven approaches are used. The existence of multiple time points hinders data interpretation since it complicates consistent cluster identity assignment. Overall, even though we observed a reduction in some neuronal clusters at the end-stage of the disease, our results were inconsistent when all time points were concerned so no conclusions regarding selective neuronal toxicity can be drawn. Our observations highlight the importance of the use of alternative techniques that do not rely on transcriptomic changes, such as cell sorting based on specific markers, for more reliable quantification of cell numbers and investigation of selective toxicity in complicated time-course experiments.

The abundance of RNA species informs on and determines the state of cells and tissues, and the quantification of mRNA transcripts opens a window to the underlying molecular processes. Differential gene expression analyses aim to identify quantitative differences in transcript abundance between two biological states and, even though they constitute an integral part of RNA-seq data analysis, accurate detection of DE genes has proved to be a challenging task when single-cell sequencing experiments are concerned. Due to the nature of single-cell methodologies, scRNA-seq datasets are highly heterogeneous and have a higher level of noise due to biological and technical reasons, requiring, thus, specifically designed statistical approaches that can efficiently handle the zero-inflated distribution of the gene counts and the sparsity of the data (Mou et al., 2019; T. Wang et al., 2019). The most widely used methods employ the Wilcoxon rank-sum test, which has become the de facto statistical method for single-cell studies and the default option of many analytical pipelines, including the popular Seurat toolkit. In fact, a recent study by (Squair et al., 2021) suggested that the Wilcoxon rank-sum test has been used to such an extent that it accounted for as many recent single-cell studies as all other statistical methods combined. The same study, though, also underlined the poor performance and high false-positive rate of the Wilcoxon rank-sum test and similar methods that do not account for variation between biological replicates (cells from the same sample are not independent replicates), while highlighting the importance of per-sample data aggregation and the use of pseudo-bulk analyses.

Since a robust DE analysis is of paramount importance and the central focus of our study, we opted to employ three different approaches and compare the results: the widely-used —though criticised— Wilcoxon rank-sum test, as well as two alternative pseudo-bulk approaches based on the Wald test and the quasi-likelihood ratio test. The Wilcoxon test is the default and recommended test of the Seurat toolkit, which meant that the analysis was easy and seamless to perform, while we opted to use DESeq2 and glmGamPoi for the pseudo-bulk approaches, which necessitated data wrangling to manually aggregate and prepare the gene counts in the appropriate formats. We opted not to perform imputation of missing single-cell data, since it has shown that it does not improve the performance of downstream analyses (Hou et al., 2020).

Our differential gene expression analysis between the RML and CD1 groups identified around 1000 genes when using Seurat, while, surprisingly, the pseudo-bulk methods identified around 5000 DEGs. This comes in contrast to previous observations that considered pseudo-bulk approaches as more conservative (Squair et al., 2021; T. Wang et al., 2019). Reassuringly, when we tested the concordance between Seurat and the pseudo-bulk approaches, we found that Seurat hits are mostly replicated by both DESeq2 and glmGamPoi, however, the agreement was much lower for the 20 dpi time point, where Seurat identified perturbations in few clusters of cortical neurons, while DESeq2 also identified clusters of migrating interneurons as being dysregulated. GlmGamPoi only identified DEGs associated with most of the clusters of cortical neurons. Perturbations in oligodendrocytes and astrocytes were only suggested by Seurat and were not replicable by any other method. These differences were expected since comprehensive studies have highlighted that agreement between different analysis methods is generally low (Soneson & Robinson, 2018; T. Wang et al., 2019).

When correlating the number of cells per cluster with the number of DEGs identified, we uncovered a very high positive correlation associated with the pseudo-bulk methods only, suggesting that the abundance of identified DEGs is mostly driven by cluster size, complicating the biological interpretation of the results. This was not true for Seurat, which also identified a smaller number of dysregulated transcripts. These were the main reasons that guided our decision to proceed with downstream analyses focusing on the gene lists generated by Seurat, as they appeared to be more stringent, they were not extensively affected by cluster size, and they included high percentages of genes that were also deemed DE by the pseudo-bulk approaches.

Regardless of the methodology used, our data suggested a selective transcriptomic response of individual cell clusters to disease, only partially attributed to differences in cluster size. More interestingly, we were able to identify a pattern of oscillating transcriptomic perturbations commencing at 20 dpi, when infectivity was low, then subsiding at 40 and 80 dpi even though infectivity proceeded exponentially, before re-emerging during the infectivity plateau at 120 dpi and being amplified at the end-stage. Additionally, the majority of DEGs at the 20 dpi time point (25 out of 42 unique genes) also exhibited this oscillating pattern of early/late dysregulation indicating an early transcriptomic response to toxicity which is reinstated in the late stages of the disease. The DEG heatmaps suggest the existence of three phases: the first one includes the 20 dpi time point, where some transcriptomic perturbations were identified, the second is a phase of transcriptomic silence that spans the 40 and 80 dpi time points, while the third phase starts at 120 dpi and proceeds until the disease end-stage. In contrast, the infectivity assays demonstrate the existence of two mechanistic phases of prion propagation, as described by previous thorough studies of the RML-FVB mouse model (Sandberg et al., 2011). Taken together, our findings demonstrated that prion infectivity does not elicit a transcriptomic response in vivo, which is supported by previous studies that demonstrated that infectious prions are not directly toxic (Benilova et al., 2020).

We hypothesise that the three transcriptomic phases correspond to fluctuations in the concentration of a toxic PrP species. The notion of the existence of such a protein, which has been named PrP$^{L}$ (PrP lethal), was first formulated by (A F Hill et al., 2000), and further discussed when more supporting evidence was collected by (Sandberg et al., 2011). This hypothesis suggests that neurotoxicity is mediated by PrP$^{L}$, which is a separate entity from PrP$^{Sc}$, however, its formation is catalysed by it (Andrew F Hill & Collinge, 2003; A F Hill et al., 2000). The model specifies that toxicity becomes evident only when the concentration of PrP$^{L}$ surpasses a local threshold (John Collinge & Clarke, 2007). Nevertheless, its existence is debated, and alternative hypotheses suggest that toxicity could be caused by PrP$^{Sc}$ (Aguzzi & Falsig, 2012; Chakrabarti & Hegde, 2009; Kristiansen et al., 2007; Moreno et al., 2012; Solomon et al., 2010). Since the existence of PrP$^{L}$ and its characterisation is not an object of our study, we will use the general term “toxic PrP species” to uncouple prion infectivity and toxicity.

We attribute the triphasic DGE pattern to an underlying mechanism of toxic PrP species clearance following the external introduction of toxic material where a subset of more vulnerable cells responds more aggressively, and clearance mechanisms are activated. Toxic species from the inoculum (which is prepared from an end-stage mouse brain) are introduced intracerebrally at inoculation and elicit a transcriptomic response. Our data from the 20 dpi time point could represent the tail of this response, which might have been even stronger earlier than that. Unfortunately, we did not have earlier time points to evaluate this hypothesis. Following inoculation, clearance mechanisms are activated, and the toxic species are gradually depleted, while the infectious species are either not affected or quickly replaced by replication. Transcriptomic alterations remain minimal at 40 and 80 dpi, even though the production of the infectious prion species increases exponentially, until reaching the second mechanistic phase which catalyses the production of the toxic species once again. When a critical concentration is reached, cell clearance mechanisms are overwhelmed, and toxic pathways are irreversibly triggered (120 dpi and end-stage). Some cell types were shown to respond more aggressively, with more pronounced changes in their transcriptomic profiles. This hypothesis of selective toxicity is further substantiated by the observation that cell clusters that responded early to toxicity also showed a stronger DE signature at 120 dpi and the end-stage.

We argue that the transcriptomic signature at 20dpi and 120dpi/end-stage are caused by the same toxic PrP species and represent similar responses, but with substantially different amplitudes. The titre of toxic PrP species is expected to be low at 20dpi since the RML inoculum used only contained 30 μL of a 1% dilution of the end-stage brain. In addition, sampling at 20 dpi might not represent the peak of the transcriptomic response, especially since brain homogenate has been shown to be cleared out between 4 days to 2 weeks post-inoculation (Büeler et al., 1993). The low titres and quick clearance would suggest that the transcriptomic response might have been even stronger at earlier time points and may have had more common genes with the end-stage. In contrast, the last two time points are associated with high titres of toxic PrP species and sustained exposure to the toxic agent. This would dictate a more pronounced transcriptomic response, especially since cell clearance mechanisms are expected to be saturated (Goold et al., 2015; López-Pérez et al., 2020; Mays & Soto, 2016; McKinnon et al., 2016). Further evidence from our study that supports this hypothesis is the existence of the early/late oscillating gene signature. Approximately half of the DEGs at the 20 dpi time point (25 out of 42 unique genes) reappear at the 120 dpi and end-stage time points. However, an RT-qPCR experiment was able to validate this signature for only 2 out of 6 genes assayed, underlying the fact that these transcriptomic differences are subtle and traditional validation approaches might not be statistically powerful enough when used with small sample sizes, like the one we used for validation (N = 4 biological replicates).

We also investigated the existence of common genes between our study and two pivotal studies in the field by Hwang et al., 2009 and Scheckel et al., 2020, which were thoroughly discussed in the introduction. We started by intersecting our sets of DEGs with the “prion signature” of 333 DEGs mentioned in Hwang et al., 2009. We were only able to identify 8 common genes, which were nevertheless associated with multiple clusters (*Gfap, Hexb, C4b, Clu, Plce1, Abca1, Pbxip1, Apod*). *Hexb* was the only gene identified at the 20 and 40 dpi time points. *Gfap* and *Hexb* were identified at the 80 dpi time point. *Gfap, Clu, C4b,* and *Hexb* were identified at the 120 dpi time point. Finally, the same genes were identified at the end-stage, with the addition of *Plce1, Abca1, Pbxip1*, and *Apod*. Even though the concordance between the two datasets is very low, this can partially be explained by the differences in the experimental and analytical methodology used. The 2009 study used microarray technology to analyse whole mouse brains, while our data was generated using protocols based on next-generation sequencing and only profiled the frontal cortex. In addition, the different analytical pipelines that the raw data was subjected to could also introduce bias. However, studies have shown that concordance between microarray and next-generation sequencing technology is usually high, so these differences might be attributed to the experimental design —the 333 DEGs reported by Hwang et al. are found at the intersection of multiple mouse and prion strains— or the lower sensitivity of snRNA-seq (Rao et al., 2018; S. Zhao et al., 2014).

When we compared our main findings with the more recent and more similar time-course experiment by (Scheckel et al., 2020) we identified similar and contrasting results. The major difference was that our study suggested that clusters of cortical neurons were associated with the highest number of DEGs, followed by medium spiny neurons and migrating interneurons of the neuronal clusters, and astrocytes, oligodendrocytes and OPCs of the glial clusters, while the study from Aguzzi’s group identified minimal changes in the expression levels of neuronal transcripts. Furthermore, we identified a signature of toxicity as early as 20 dpi, which is absent from the previous study. Both studies agree on the extensive glial involvement in the end-stage of the disease and have identified numerous common genes being differentially expressed (Figure 4.27). When we compared the lists of DEGs we identified common patterns between the two studies — most DEGs were found at the end-stage; there were numerous DEGs common between the last two time points — and 134 shared dysregulated genes. These represented approximately 30% of the total unique genes identified in our study (134 shared genes / 438 total unique genes across all time points) and were mostly found to be dysregulated at the last two time points. Differences between the two studies could be attributed to the different methodology used (ribosome profiling versus snRNA-seq) and the experimental design (transgenic mouse model on a C57BL/6 background versus wild-type FVB mice).

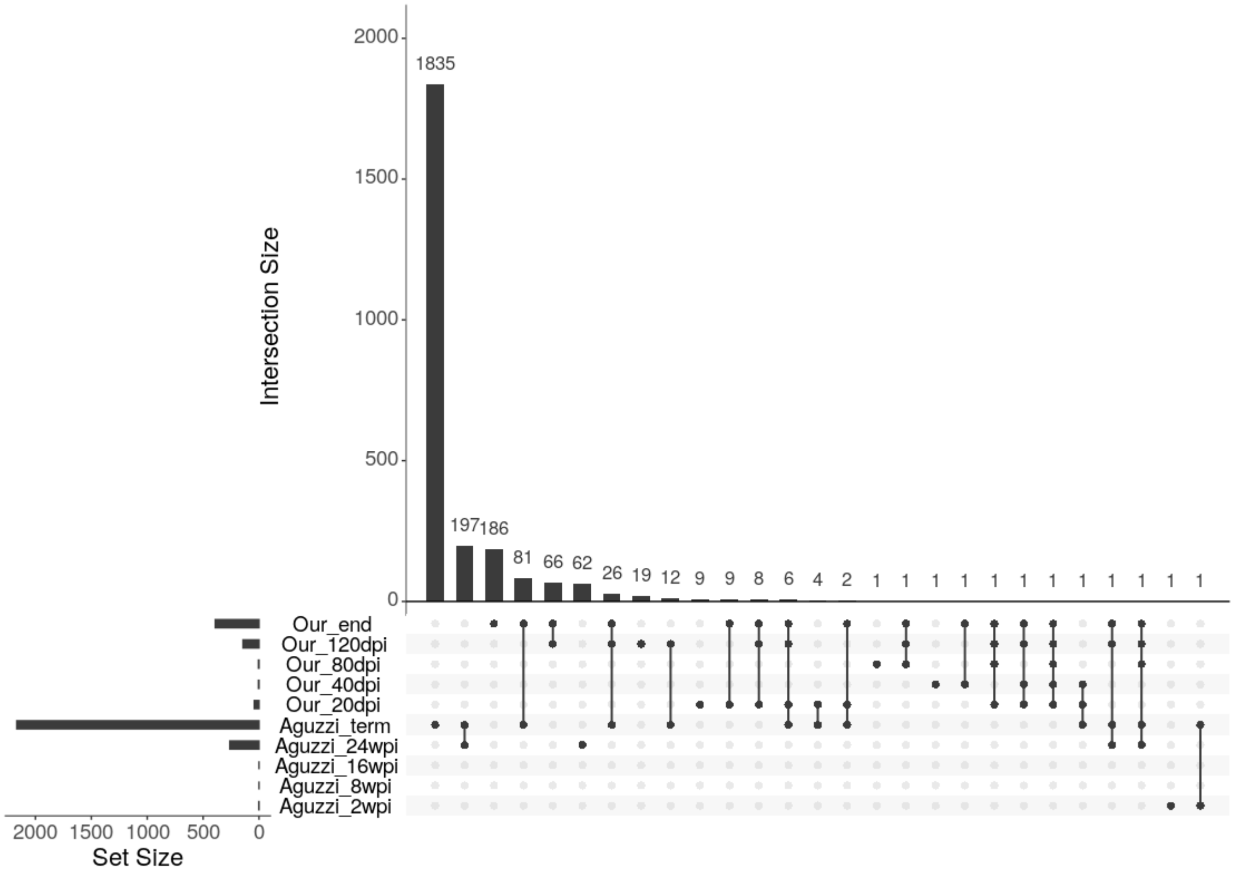


***Figure 4.27: Relationship of gene sets between our study and the study by Scheckel et al. (2020).*** *We identified common patterns and shared genes between the two studies. Both studies identified the highest number of dysregulated genes at the last time point followed by the second to last. In addition, there were shared genes between the last two time points in each study independently. 134 genes were shared between the two studies across all time points. Gene lists were downloaded from the supplementary material of the online eLife publication. Genes were filtered based on criteria selected by the authors to include only those that were deemed to be differentially translated (|log2FC| > 1 and FDR < 0.05) and different cell types were then aggregated per time point. Labels "Our" and "Aguzzi" on the vertical axis represent gene sets of our study and the study by Aguzzi's group, respectively. Wpi: weeks post-inoculation.*

Zooming out of the lists of individual genes, we employed a gene set enrichment analysis (GSEA) to identify perturbed biological processes. The main advantage of our approach is that it can identify perturbed gene networks even when each gene might not be significantly differentially expressed. For this type of analysis, we aggregated the expression data of all genes across all cell types at each time point. Expectedly, most of the enriched terms were associated with synaptic processes. Previous studies have consistently highlighted the central role of synaptic dysfunction in prion diseases, as well as other protein misfolding neurodegenerative disorders (Mallucci, 2009; Soto & Satani, 2011). It is widely accepted that synaptic perturbation precedes cell death, and the two processes are separately regulated. Since our experiment is terminated when scrapie sickness is confirmed, which is earlier than the disease end-stage if mice were allowed to reach their end of life, the alterations that become apparent are likely to reflect the mechanisms of synaptotoxicity, while cell death processes might not have been activated yet. Indeed, we encountered no evidence of activated molecular mechanisms of cell death and were able to isolate similar numbers of nuclei when preparing nuclear suspensions from all time points (even though the number of nuclei isolated for a given volume of brain sample was not formally quantified). Furthermore, we observed an enrichment of pathways associated with ion homeostasis, and, more specifically, the regulation of potassium channels, which is a common feature of neurodegenerative diseases and neurological disorders, especially in astrocytes and neurons (Kumar et al., 2016; Lee et al., 2022; S. Wang et al., 2022; Xiao Zhang et al., 2018). Most of those processes were found to be dysregulated at the 20 dpi time point and the last two or three time points, without being enriched at the 40 dpi time point. The GSEA, which uses a different analytical methodology and does not rely on the list of DEGs, provides additional support for our three phases hypothesis and evidence that the system can indeed recover transcriptionally when the externally introduced toxic species have been cleared.

Of interest was also the identification of one gene, *Hexb,* which was the only found to be consistently differentially expressed across all time points, and consistently upregulated in 19 clusters in total. *Hexb* encodes the beta subunit of two related enzymes, the beta-hexosaminidases A and B. These enzymes are mainly found within lysosomes and are involved in the catabolism of sphingolipids, oligosaccharides, and glycoproteins. Loss-of-function mutations cause Tay-Sachs and Sandhoff diseases in humans which are metabolic disorders associated with neurodegeneration and motor regression due to the accumulation of GM2 ganglioside in neurons (Mahdieh et al., 2018; Maier et al., 2003; Myerowitz et al., 2002). Upregulation of *Hexb* has been reported by previous studies in prion diseases and provides more evidence to the hypothesis of an activated clearance mechanism and the involvement of sphingolipid metabolism and lysosomal pathways in disease pathogenesis (Carroll et al., 2020; de Melo et al., 2021; Hwang et al., 2009; H. O. Kim et al., 2008; Mahfoud et al., 2002; D. R. Taylor & Hooper, 2006).

We will now focus on the three populations of interest separately and discuss changes in their transcriptomic landscape as well as the results of the over-representation analyses.

### 4.3.3 Transcriptomics of neurons

Neurons are the only cell type that is known to suffer the ill effects of prion infection and prion neurotoxicity is cell-autonomous, i.e., the expression of normal host prion protein is essential for the manifestation of toxicity (S Brandner et al., 1996). For these reasons, studying the transcriptomic perturbations in neurons is essential. However, given the fact that we are only able to assay living cells, a caveat of our approach is that it is possible that highly affected neurons are lost before sample collection and therefore not included in the dataset which would then be enriched with resilient populations that have survived. This effect might be especially true for the later time points, where vacuolation and neuronal loss is evident from the histopathological analysis. Still, this phenomenon might not be as profound since no major differences were identified in the cell proportions as discussed in the previous section.

Keeping this potential caveat under consideration, we examined the DGE lists for all neuronal populations identified. Firstly, we identified numerous neuronal populations that behaved differently throughout the course of the disease, highlighting that changes in gene expression can be conditional on cell type. Most of the DEGs were found in the last two time points, while the dysregulation pattern of three phases is only evident in two clusters (7 and 11), which represent cortical neurons that show perturbations at 20 dpi which are reversed at 40 and 80 dpi, before becoming evident again at 120 dpi and amplified at the last time point. Overall, most of the DEGs were found to be downregulated, with a universal suppression of transcription especially apparent at the disease end-stage.

The cell-specific response to disease is further demonstrated when looking at the segregation patterns of the different cell clusters in the low-dimensional space, where we observed that transcriptomic response is not only cell-type-specific but also cell-subpopulation-specific, with neuronal subpopulations showing distinct grouping patterns.

The medium spiny neurons and some of the subpopulations of cortical neurons were found to be more affected by the disease with clear separation of cell clusters originating from RML-inoculated and CD1-inoculated mice, while the effect was less pronounced for the migrating interneurons. In contrast, some subpopulations of cortical neurons and migrating interneurons were shown to not be affected by the disease, with all samples clustering tightly together. These results once again highlighted the importance of single-cell resolution transcriptomics.

When reviewing the list of DEGs, we observed that the gene *Maml3* is downregulated at 120 dpi and end-stage in neuronal clusters only. In addition, the percentage of cells expressing the gene was found to be reduced. These results were concordant with the pseudobulk analysis, with p-values among the lowest in each cell cluster. The levels of *Maml3* follow a downward curve with expression being suppressed following disease progression during the last two time points. These observations led us to investigate the involvement of the Notch signalling pathway, where Maml3, the Mastermind Like Transcriptional Coactivator 3, is a transcriptional coactivator. In addition, Maml3 is linked with positive regulation of transcription by RNA polymerase II creating a connection between the decreased levels of its expression and the generalised transcriptional suppression that we observed. The Notch pathway is important in development but has also been implicated in neurodegenerative and other diseases that cause cognitive impairment, notably Alzheimer's disease, multiple sclerosis, and amyotrophic lateral sclerosis (Ables et al., 2011; Ho et al., 2020). Previous studies have also identified links between the Notch pathway and prion diseases. A 2005 study showed that Notch-1 expression was higher in RML-inoculated mice compared to healthy controls, with expression levels increasing concomitantly with PrP$^{Sc}$, while the levels of the Notch intracellular domain transcription factor (NICD), a cleavage product, were also higher in prion-infected ScN2a cells compared to uninfected N2a cells (Ishikura et al., 2005). A few years later it was demonstrated that inhibition of the Notch pathway and introduction of quinacrine that inhibits the formation of PrP$^{Sc}$ in cultured cells can diminish PrP$^{Sc}$ levels in the brains of RML-inoculated mice (Spilman et al., 2008). While the published literature suggests an increase in expression of Notch in prion disease, our data demonstrated the opposite trend. In addition, no other gene of the Notch pathway was found to be DE, indicating that the dysregulation of *Maml3* could be an independent event that could affect Malm3 target genes, but does not necessarily prove the downregulation of the complex Notch pathway.

In addition, we identified transcriptomic changes in groups of genes encoding phosphodiesterases. These enzymes hydrolyse cyclic nucleotides, regulating the levels of the second messengers cAMP and cGMP and, thus, cell function (Boswell-Smith et al., 2006). These did not exhibit a uniform change in expression: Pde7b was found to be downregulated at 120dpi and end-stage, Pde10a upregulated at 20dpi and downregulated at 120dpi and end-stage, and Pde4a and Pde4b downregulated at the end-stage only. This irregular pattern suggests that these perturbations likely represent secondary effects and compensation to prior dysfunction, as supported by the literature (Bollen & Prickaerts, 2012).

Finally, *Sox5* and *Sox6,* two genes of the SoxD family, were found to be downregulated in neurons at the disease end-stage. This gene family called SoxD is important for neural development, while Sox5 has also been implicated in schizophrenia in human single-cell transcriptomics studies (Ruzicka et al., 2020).

We then focused on interpreting the results of the over-representation analysis, which provided information about perturbed gene networks in each neuronal subcluster. Cluster 11 of cortical neurons was the only cell cluster that had associated pathways at 20 dpi (it was also the cluster with the higher number of DEGs). The ORA identified terms related to the metabolism of glycosylated proteins, and the function of transferases, including enzymes that direct oligosaccharide processing. No neuronal clusters were identified by the ORA for the time points from 40 dpi to 120 dpi, however, in the disease end-stage, the majority of the perturbed clusters were of the neuronal type (29/37 clusters in total in MF, CC, and BP classifications), which is due to the allocation of numerous neuronal subclusters, but also highlighting the fact that prion disease disproportionally impacts neuronal populations. Even though the dysregulation profiles were different for distinct neuronal types (e.g., cortical neurons versus migrating interneurons), the most perturbed pathways were relevant to synaptic function (see also discussion in the previous section

4.2.4). Finally, we were still not able to identify pathways related to cell death, highlighting that synaptic perturbations and cell death are two, separately regulated processes.

We proceeded to validate our findings by quantifying the levels of genes associated with synaptic pathways using a quantitative real-time PCR assay, however, we were only able to validate a similar pattern for only 2 out of 5 genes assayed. We argued that the gene expression changes were small and the real-time PCR methods might not have enough statistical power to distinguish them when used with such small sample sizes (N = 4), especially since suspensions of mixed cell populations were used, where neurons were only a subset and transcriptomic changes were expected to be diluted.

We suggest that future experiments planned to validate these transcriptomic changes specifically in neurons could be based on single-population sequencing. In short, populations of neurons could be isolated, then lysed and the RNA could be extracted. Library preparation and sequencing would then be possible following well-established bulk sequencing methods. These techniques are expected to have enough power to identify very small changes in gene expression. More details regarding future experiments will be discussed in section 6.2.

### 4.3.4 Transcriptomics of astrocytes

The role of astrocytes in prion disease has long been debated. Early studies had shown that mice with astrocyte-specific PrP expression could develop prion disease after inoculation with infectious material (Jeffrey et al., 2004; Raeber et al., 1997). However, future work questioned these findings, highlighting that the transgene constructs used had some activity in neurons as well (Marino et al., 2000). Further studies demonstrated that astrocytes can, indeed, replicate prions, however, they do not suffer from prion toxicity, and glial activation is non-autonomous (i.e., it requires neuronal PrP, not glial) (Krejciova et al., 2017; Lakkaraju et al., 2021).

In our data, the astrocytes cluster 69 was the cluster with the higher number of identified differentially expressed genes — cluster 68 had a small number of cells and did not show interesting transcriptomic variation, so we will focus on cluster 69. Like the neuronal signature, most of the astrocyte-related genes were found to be suppressed during the disease end-stage. In addition, a PCA analysis of the different biological samples indicated that transcriptomic changes in astrocytes followed disease progression during the last two time points. The ORA identified astrocytes as the only population with perturbed biological pathways at 120 dpi, while these perturbations were amplified in the end-stage. The affected pathways pertained to synaptic function, cell junctions, and cell adhesion.

Indeed, astrocytic dysfunction has been implicated in prion disease through the activation of the unfolded protein response (UPR), while astrocytes have synaptogenic functions and there is evidence that astrocytes take an active part in the synapse as a third member — abnormal astrocytic function can cause or contribute to synaptic imbalances and cognitive impairment (Santello et al., 2019; Smith et al., 2020). Furthermore, the cellular component classification indicated the involvement of specifically glutamatergic synapse pathways at both 120 dpi and disease end-stage. Glutamate homeostasis is one of the fundamental functions of astrocytes, essential to protecting neuronal cells from glutamate build-up and excitotoxicity (Chung et al., 2015; Mahmoud et al., 2019). Our data indicated that all 20 genes that are involved in the glutamatergic synapse pathway (*Nrxn1, Grm3, Plcb1, Dgkb, Cadm1, Gpm6a, Mdga2, Gpc6, Nlgn1, Rgs7, Adgrl3, Grid2, Magi2, Eps8, Dlgap1, Ncam1, Abr, Shisa9, Tnik, Ephb1*) were downregulated at both 120 dpi and end-stage, suggesting suppression of glutamate reuptake in prion disease and pointing towards reported mechanisms of neuronal toxicity (Goniotaki et al., 2017; Khosravani et al., 2008).

Gap junctions allow astrocytes to form dynamic networks and, even though their exact role has not been extensively studied, there is evidence that they are essential for modulating inflammatory response, buffering ions and neurotransmitters, and distributing energetic substrates throughout the brain (Santello et al., 2019; Wallraff et al., 2004). Their malfunction has been implicated in numerous diseases and neurological disorders, including Charcot-Marie-Tooth disease, hereditary deafness, and uncorrelated motor neuron firing (Dong et al., 2018). It would, thus, be safe to assume that these networks could be affected in prion diseases, without it being clear, though, whether their role is causal or a secondary response to disease.

The Molecular Function classification identified cell adhesion as the top pathway for both 120 dpi and end-stage time points. Cell adhesion molecules (CAMs), a subset of cell surface proteins involved in the binding of cells with the extracellular matrix or other cells, have previously been implicated in prion disease. Studies have shown that CAMs can bind $PrP^C$, and stipulate that the function of normal PrP is related to the recruitment of signalling molecules that control the stability of the adhesion complexes on the plasma membrane (Martins et al., 2010; Petit et al., 2013; Schmitt-Ulms et al., 2001). Furthermore, more recent transcriptomic studies have shown that cell adhesion and extracellular matrix organisation genes were enriched among astrocyte-specific genes (Scheckel et al., 2020). Nevertheless, the exact interplay between cell adhesion molecules and prion disease remains elusive.

While astrocytes are indispensable to the maintenance and integrity of the central nervous system, certain conditions can cause a phenotypic shift, making these cells assume toxic phenotypes that contribute to neurotoxicity (Liddelow et al., 2017). This activation is mediated by the microglia and can have different outcomes in neurodegenerative conditions, leading to a crude separation of two astrocytic phenotypes: A1, the neurotoxic astrocytes induced by neuroinflammation, and A2, the neuroprotective astrocytes induced by ischemia. Even though we were interested in assaying the phenotype that astrocytes in our dataset assume, our attempt to quantify the A1 and A2 signatures using the single-cell RNA-seq data was hindered due to low sequencing sensitivity. We noticed that most of the A1/A2 gene sets were missing from our data, hindering the extraction of meaningful conclusions. This prompted us to use quantitative real-time PCR to assay these specific genes in nuclei suspensions. We found evidence of astrocytic activation in the disease end-stage, while the exact phenotype of those astrocytes was not clear (2/5 A1 genes were upregulated, A2 genes were up, and downregulated). These experiments highlight the need for more sensitive approaches, such as single-population sequencing or low-throughput high-sensitivity single-cell sequencing of sorted astrocytes (such as Smart-seq2).

To address this limitation of our dataset, we used a different approach to identify the presence of astrocytes and quantify astrogliosis in the mouse brain. We performed

RNAscope, a special genomics technique based on in situ hybridisation for the detection of target RNA molecules of interest. We probed the mouse brain for *Gfap* mRNA, which is considered to be a highly specific marker for astroglia. Even though more recent studies have identified lower *Gfap* expression in neurons in the human hippocampus (Hol et al., 2003), its expression is expected to be much higher in astrocytes and by comparing the two experimental groups we were able to quantify gliosis in mouse prion disease. Our findings demonstrated astrocytic involvement initiating as early as 80 dpi, which mirrored the transcriptomic results. Interestingly, *Gfap* was found to be elevated as early as 80 dpi in the hippocampus, thalamus and the brainstem, while its levels in the cortex (which was the tissue used for the transcriptomic study) remained lower until the 120 dpi time point. These results suggest that the astrocytic activation might have been even more pronounced had another anatomical region been used for single-cell RNA sequencing. Finally, no evidence of any effect of the residual inoculum in the activation pattern of the astrocytes was found, as *Gfap* was found to be increased in expression in regions further away from the inoculation site.

#### 4.3.5 Transcriptomics of oligodendrocyte precursor cells and mature oligodendrocytes

The final populations on which we focused were the oligodendrocyte precursor cells (OPCs) and mature oligodendrocytes (MOLs). Oligodendrocytes are the myelinating glia of the central nervous system and there is evidence that they are incapable of replicating prions (Prinz et al., 2004). Even though demyelination is a common hallmark of neurodegenerative diseases such as multiple sclerosis or prion disease, these cells are generally understudied, and little is known in the context of prion diseases (Domingues et al., 2016). Rodent models lacking $PrP^C$ expression have been shown to develop a chronic demyelinating phenotype, highlighting the importance of axonal prion protein to peripheral myelin maintenance (Bremer et al., 2010; Nishida et al., 1999); however, these results could not be reproduced in non-rodent mammalian models (Richt et al., 2007; Yu et al., 2009).

Our transcriptomic analysis identified an increased expression of *C4b* in MOLs in disease during the last two time points. C4b is part of the complement system, a system of plasma proteins and part of the immune system that is activated by pathogens or pathogen-bound antibodies (Charles A Janeway et al., 2001). These results indicate that at least a proportion of the MOLs in our study could be activated. Indeed, there is evidence suggesting that oligodendroglial cells may be a source of complement proteins in the brain, contributing to the pathogenesis of inflammatory and neurodegenerative diseases such as Alzheimer's disease, multiple sclerosis, and Parkinson's disease (Hosokawa et al., 2003; Rus & Niculescu, 2001). C4d-immunoreactive complement-activated oligodendrocytes have been described in progressive supranuclear palsy, multiple system atrophy, amyotrophic lateral sclerosis, Parkinson's disease, Alzheimer's disease, and multiple sclerosis (Schwab & McGeer, 2002; Yamada et al., 1990, 1991).

Another interesting, overexpressed gene identified in MOLs was *Apod*. The gene encodes apolipoprotein D (apoD), a lipocalin with antioxidant and neuroprotective functions (Dassati et al., 2014; He et al., 2009). ApoD has been shown to be upregulated in astrocytes during ageing and in neurological disorders including bipolar disorder, schizophrenia, Alzheimer's disease, and Parkinson's disease (Bhatia et al., 2013; de Magalhães et al., 2009; Glöckner & Ohm, 2003; Loerch et al., 2008; Mahadik et al., 2002; Ordoñez et al., 2006; Thomas et al., 2001). In normal conditions, ApoD is expressed in low levels by the myelinating glia and its expression is rapidly increased in response to trauma or neurodegeneration. Evidence suggests that the increased production of ApoD constitutes an endogenous mechanism of protection (Corraliza-Gomez et al., 2019; Dassati et al., 2014). In summary, our data suggested that homeostatic mechanisms could be activated as a response to prion disease and neurodegeneration, an observation that is further supported by studies in the prion field that have identified increased levels of *Apod* expression (Hwang et al., 2009; R. A. Moore et al., 2014; Scheckel et al., 2020).

Turning to the ORA, we identified a suppressed cell adhesion pathway in OPCs: 7 out of 8 genes associated with the pathway were found to be downregulated at the disease end-stage (*Nrxn1, Dscam, Ptprt, Dscaml1, Tnr, Ctnnd2, Nlgn1*). We interpreted this finding as a possible sign of increased OPC mobility as a response to disease, however little is known regarding these migratory mechanisms (Fok-Seang et al., 1995).

The biological process classification uncovered perturbed myelination and neuronal ensheathment pathways in MOL populations, an expected result in prion disease characterised by demyelination, as previously discussed.

Due to the small numbers of MOLs, validation of gene signatures was not attempted, as the effects were expected to be diluted in the nuclei suspensions. Future validation of perturbed biological networks could be possible in sorted glial populations, using single-population sequencing approaches.

# 5 Single-cell transcriptomics of human prion disease

## 5.1 Introduction

### 5.1.1 Chapter summary

Following the successful mouse experiments, we decided to apply the same methodology to profile human prion diseases, specifically sporadic CJD. We designed a case-control study which included post-mortem and biopsy brain samples of sCJD patients and controls. Our results indicated that RNA quality and quantity in these samples were not sufficient for single-cell sequencing using high-throughput methods. We explore the reasons that this might be the case and suggest alternative approaches for future experiments.

## 5.2 Results

**Nuclei extraction, library preparation, and sequencing**

We performed single-nucleus RNA sequencing on the post-mortem and biopsy frozen brain samples using the Parse Evercode WT protocol, a commercialised and improved version of the SPLiT-seq protocol that was used for the mouse samples. Human cortex samples were left to thaw and the grey matter of the superior frontal gyrus was hand-dissected and dissociated (see chapter 2.13.1 for sample selection criteria). Nuclei suspensions were fixed and examined under the microscope for quality assurance. The modified protocol that included density gradient centrifugation remarkably improved the quality of the resulting suspensions, substantially reducing the amount of visible debris (data not shown).

When all nuclei suspensions from all 26 samples were prepared, the samples were diluted and loaded on a single 96-well plate for the following split-pool barcoding rounds. Post-mortem sCJD samples, sCJD biopsies, control post-mortem samples and control biopsies were loaded on the same plate and processed in the same batch, in order to reduce possible batch effects. At the end of the barcoding protocol, we recovered a total of approximately 60,000 nuclei that were separated into 6 sub-libraries. Sub-libraries 1-5 included approximately 9,000 nuclei, and sub-library 6 included approximately 5,000 –

7,000 nuclei. The resulting sub-libraries were processed in parallel for the preparation of sequencing libraries.

Sequencing libraries were pooled in pairs and sequenced on the NextSeq 500 (Illumina) for a total of 3 high output sequencing runs, yielding approximately 50,000 expected transcriptomes. Sequencing generated approximately 850 million reads in total. The quality of the sequencing runs was assessed by running FastQC on the resulting fastq files and examining the statistics (External Supplementary File 2). After ensuring that sequencing was of adequate quality, the files were processed using the Parse pipeline to generate the count matrices, which were loaded into Seurat for further analysis.

### Quality control in Seurat

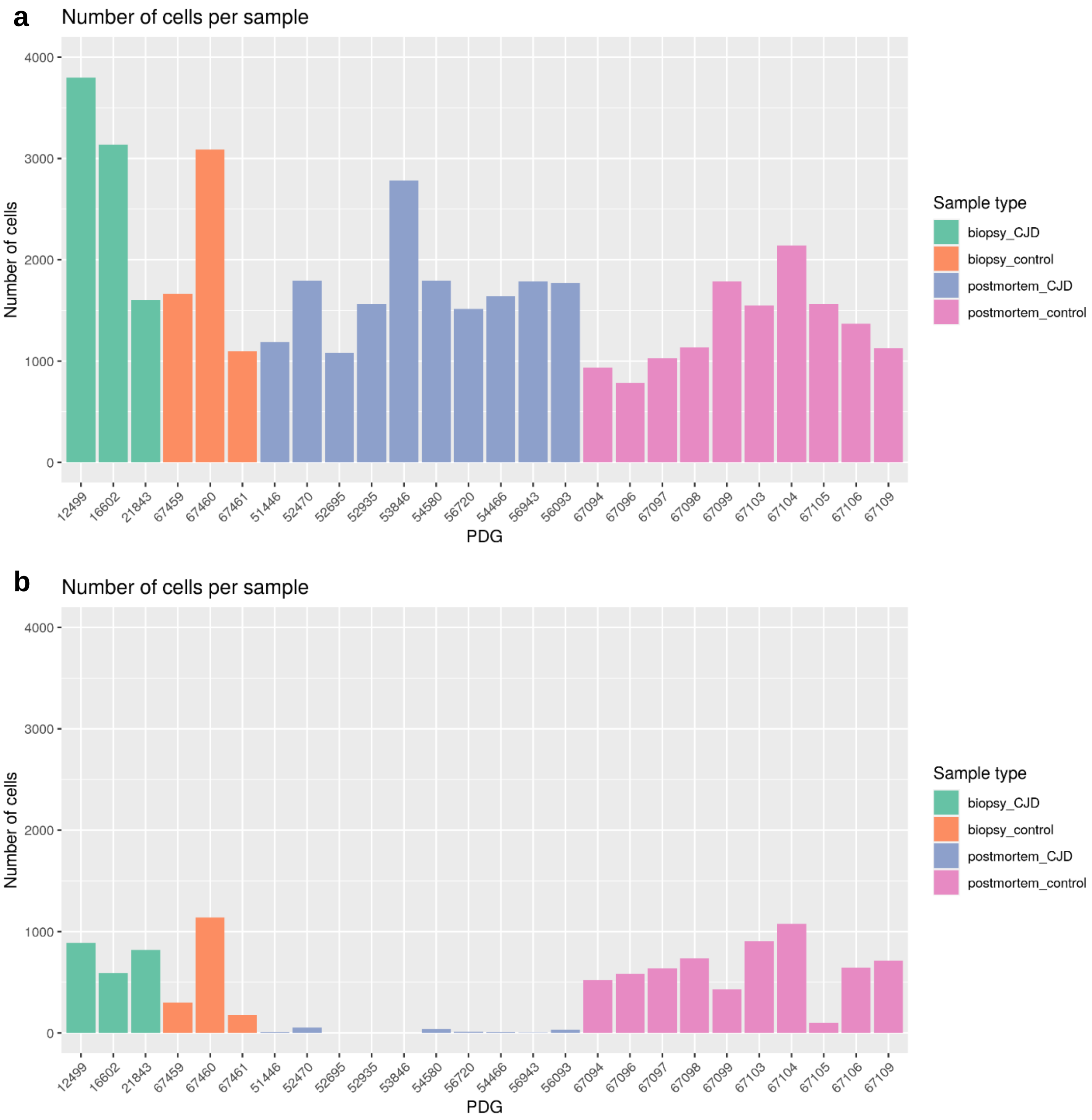
We followed the same filtering criteria as previously and filtered cells based on their feature count and the percentage of mitochondrial genes. Cells with a feature count between 250 and 2500 and a percentage of mitochondrial genes < 1% were retained. This filtering removed the majority of the data and impacted each group of samples differently (Table 10 and Figure 5.1 **a, b**). For the human biopsies, only 28% and 27% of the data passed the filtering criteria for the controls and disease, respectively. In regards to the post-mortem samples, 50% of the data generated from the post-mortem controls passed filtering criteria, while only 0.98% of the data associated with the post-mortem sCJD samples were of high enough quality.

| Sample group | No. of cells before QC | No. of cells after QC | Percentage of cells passing filters |
|---|---|---|---|
| CJD biopsy | 8536 | 2294 | 26.9% |
| Control biopsy | 5845 | 1611 | 27.6% |
| CJD post-mortem | 16916 | 158 | 0.9% |
| Control post-mortem | 13420 | 6341 | 47.3% |

***Table 10: Filtering the human data removed a large percentage of low-quality transcriptomes.*** *No.: number, QC: quality control.*

The number of features per nucleus was also found to be lower than in previous experiments (for the mouse data: M = 1050 features per nucleus, SD = 565; for the human data: M = 758 features per nucleus, SD = 496), even when only considering the cells that passed QC for the calculation of the means (Figure 5.1 **c, d**).

Based on the quality and the abundance of the data we decided to only proceed with analysing the groups of CJD and control biopsies since the post-mortem CJD samples were not usable after the extensive filtering. As part of the quality control process, we then clustered the 6 biopsy samples and generated a UMAP plot to assess sample distances in a low-dimensional space (Figure 5.2). We identified an underlying batch effect that drove cluster separation based on the sample identity and not cell type. Cells of each of the CJD biopsy samples mostly clustered together, with sample 16602 clustering further away from all other samples. The same phenomenon was evident to a lesser degree for the other two CJD biopsy samples (12499 and 21843) which did not seem to occupy the same space on the UMAP plot. This result highlighted the existence of some underlying technical bias with an effect strong enough to prohibit cluster separation due to biologically meaningful transcriptomic variation. The effect could also have been amplified due to the low depth of the transcriptomic data available and the small number of cells retained in the dataset after quality control.


a
Number of cells per sample
Number of cells
0
1000
2000
3000
4000
12499
16602
21843
67459
67460
67461
51446
52470
52695
52935
53846
54580
56720
54466
56943
56093
67094
67096
67097
67098
67099
67103
67104
67105
67106
67109
PDG
Sample type
biopsy_CJD
biopsy_control
postmortem_CJD
postmortem_control
b
Number of cells per sample
Number of cells
0
1000
2000
3000
4000
12499
16602
21843
67459
67460
67461
51446
52470
52695
52935
53846
54580
56720
54466
56943
56093
67094
67096
67097
67098
67099
67103
67104
67105
67106
67109
PDG
Sample type
biopsy_CJD
biopsy_control
postmortem_CJD
postmortem_control

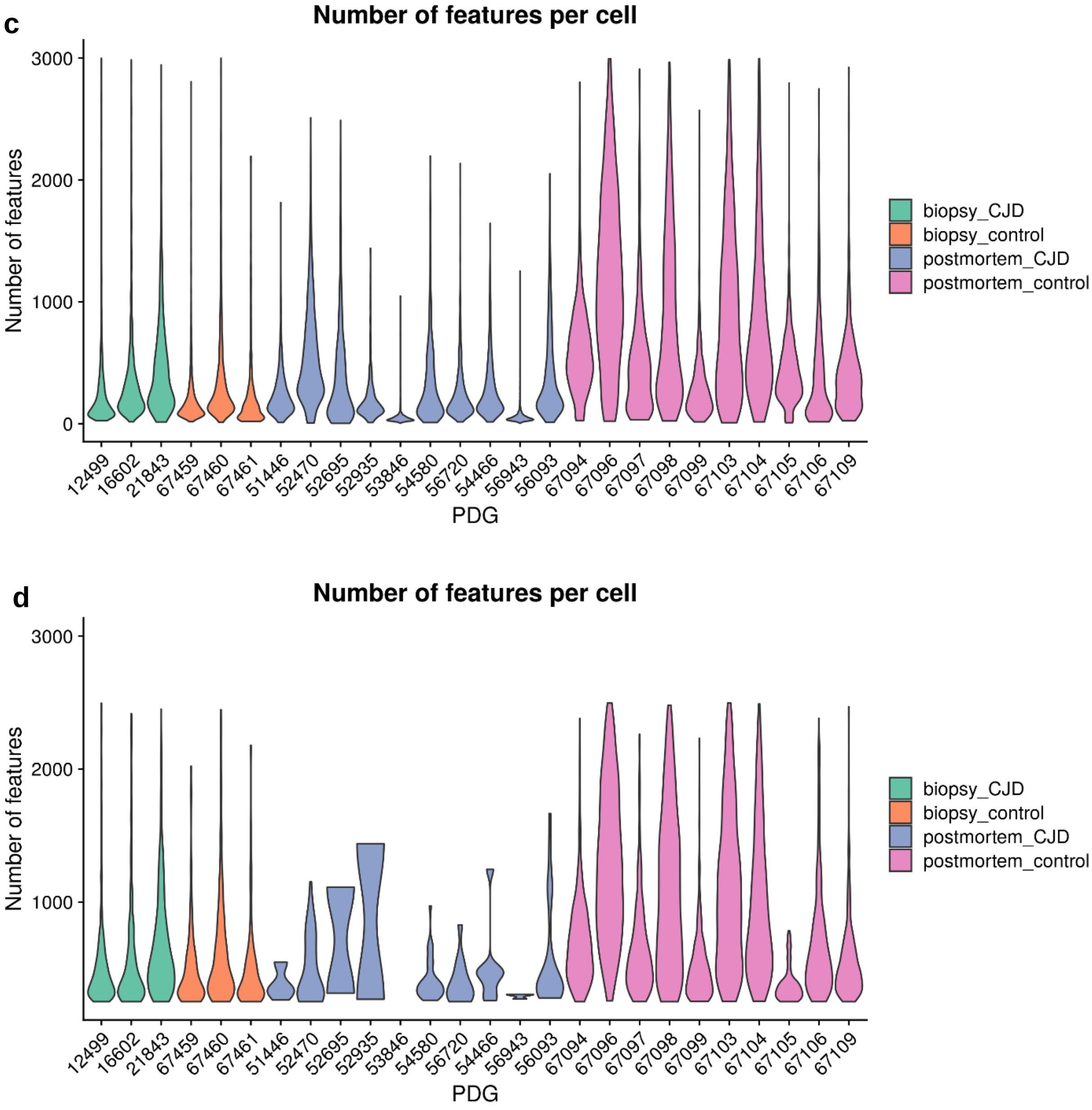


***Figure 5.1: A quality control step filters out most of the sequenced transcriptomes, highlighting the poor quality of the starting material.*** *(**a**, **b**) Bar plot of the number of cells identified per sample before and after quality control, respectively. While the filtering step removed a substantial percentage of cells from all samples, the filtering effect was more dramatic for the post-mortem sCJD brain samples, where only a very small number of cells passed the filtering criteria, rendering these samples unusable. In contrast, the post-mortem control brain samples were found to be of higher quality. Regarding the biopsy samples, both disease and control samples behaved similarly with approximately half of their transcriptomes passing the filtering criteria. (**c, d**) Violin plots show the distribution of the number of features per cell for each sample before and after quality control, respectively. The post-mortem control samples were found to have the*

*highest number of genes per nucleus, a measure of the quality of the original sample. The biopsies and post-mortem sCJD samples have a lower number of features per nucleus. The violin plots of the post-mortem sCJD samples in **d** appear irregular due to the small number of cells per sample.*

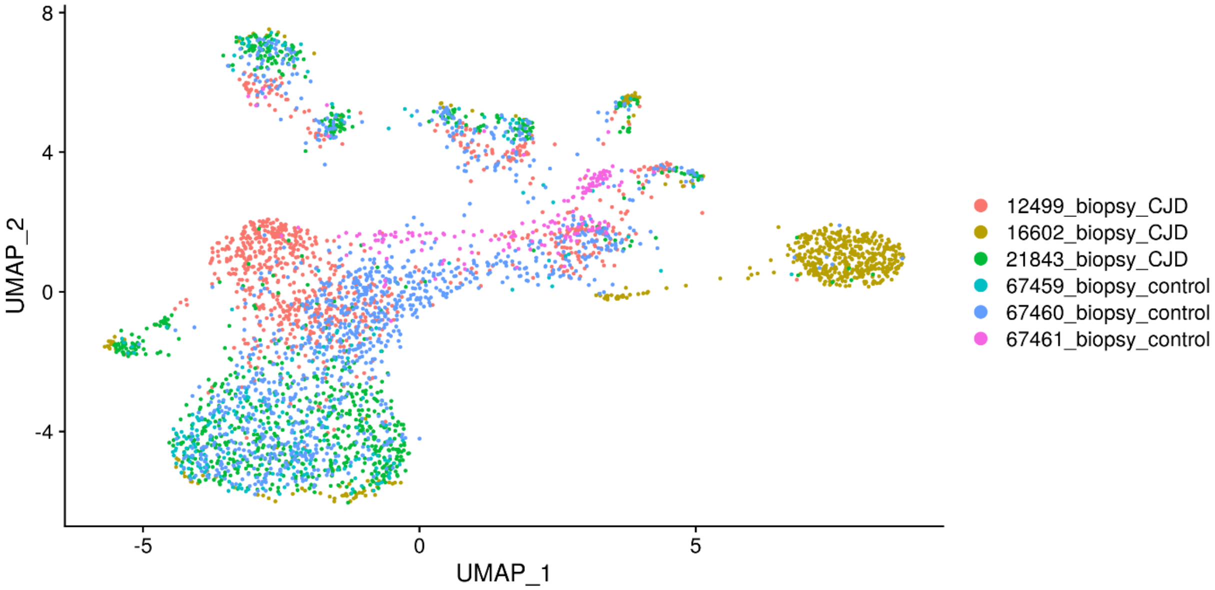


***Figure 5.2: A UMAP plot of the remaining biopsy samples visualises a non-uniform distribution of the cells.*** *The transcriptomes of the biopsy samples were visualised in the two-dimensional space using a UMAP plot. While cell clustering was expected to be driven by different cell types, we instead identified a pattern where cell clusters mostly comprise cells of the same sample. This is most evident in sample 16602, which clusters separately from all other samples, but can also be seen for CJD samples 21843 and 12499, or controls 67460 and 47461 which occupy different places on the UMAP plot even though they belong to the same experimental group. This clustering pattern indicates a strong transcriptomic bias which hinders further analysis.*

We decided not to proceed with the rest of the analysis, as any results generated would be biased and uninterpretable.

## 5.3 Discussion

Having concluded the mouse study, our subsequent scientific questions required the generation of similar single-cell RNA seq. datasets from human samples. We aimed to examine how previous findings in mouse models could also be relevant to human disease and identify similarities and differences and investigate human prion diseases in single-cell resolution for the first time.

The first challenge was faced when selecting the experimental cases and controls to be included in the study. We decided to introduce as few variables as possible, so we decided to only include codon 129 methionine homozygous cases of sporadic CJD. In addition, we decided to only focus on the frontal cortex, and more specifically the superior frontal gyrus, and only kept cases with samples stored in tissue cassettes. This last decision was important since tissue cassettes preserved the anatomy of the tissue slices and allowed us to easily identify the cortical region to be sampled. These cases had PrP type 2 or 3 (London classification), and we decided not to limit the selection to only one PrP type as this would decrease the number of available samples.

This stringent selection process narrowed down the possible samples to 10 sCJD cases with frozen samples that fitted all criteria. We were then able to request non-prion control samples kindly provided by the Queen Square Brain Bank, that matched the gender and anatomical region of our cases. However, we were not able to match the age between two groups, with controls having a mean age of 13.3 years higher than the cases. Neuropathological investigations revealed mild pathological ageing in most of the samples of both groups, however these were deemed to be non-contributory to the main cause of death. 20 post-mortem samples in total were used for this study.

In addition, we were expecting that the long post-mortem delays until sample collection could mean that transcriptomic information might be lost or altered, as RNA degradation and expression of ischemia-related gene patterns have previously been underlined (Ferreira et al., 2018; Heng et al., 2021; Highet et al., 2021). Indeed, the average post-mortem interval from death to sample collection was 4.7 days across all samples. To address this limitation, we designed a parallel study of non-dominant lobe biopsy material that had been acquired by the National Prion Clinic and archived in our Unit. Brain biopsies have a much smaller and less variable interval from sample collection to storage usually less than 30 minutes. We were able to source 3 precious biopsy samples with enough material to be used in our transcriptomics study. Accepting that biopsies of healthy individuals would be difficult to find, and post-mortem samples would not be an ideal control for these samples, we decided to use frontal lobe biopsies from non-neurodegenerative disease controls with mixed clinical diagnoses as the control group.

These 3 samples were kindly provided by BRAIN UK and included tissue with only non-specific minor histological changes (pathological non-diagnostic samples), sampled similarly to the biopsies. Exact age matching was not possible for these samples, however the difference in the means of ages between the two groups was small: the mean age of the controls was 4 years higher than the cases.

At the beginning of this research project, we selected SPLiT-seq on the basis that it can be used for infectious material and is compatible with BSL-3 procedures and working conditions. Indeed, we successfully applied this methodology for the mouse study, however by the time the human study was to commence, an updated and optimised commercial version of the sequencing pipeline was available. This recent version produced by Parse Biosciences promised higher sensitivity and eliminated the need to source all reagents separately, streamlining the library generation protocols. Since both methods were fundamentally similar, we decided to proceed with the optimised protocol and the Parse Evercode WT kit, to harness the increased sensitivity that is claimed to offer.

Nuclei suspensions and sequencing libraries were prepared in a BSL-3 laboratory as they contained human prions. Importantly, control and case samples were processed in pairs in parallel to minimise batch effects. A decontamination step was introduced to eliminate prion infectivity in all samples (including controls), and final libraries underwent quality control steps and were deemed to be of high quality before sequencing. Cases and controls were sequenced in the same run. Sequencing generated lower output than expected from high-output kits, which was attributed to a lower number of amplifiable molecules in each library. Nevertheless, the total number of reads was satisfactory for the projected number of cells sequenced.

Initial quality control metrics on the raw sequencing data were satisfactory, however, when the data were demultiplexed and loaded into Seurat for quality control, it became evident that the quality of the dataset was lower than the previous mouse study. We decided not to relax our filtering criteria in order to keep the high-quality transcriptomes. We noticed that filtering affected differently the sample groups. Approximately half of the cells originating from biopsy samples were excluded, while approximately all cells from the CJD group were deemed to be of unsuitable quality based on our filtering criteria.

The decontamination step could have uniformly reduced the number of transcripts available for sequencing but could not explain the specific reduction in the number of high-quality transcriptomes of the sCJD samples. In addition, CJD samples and controls were processed in parallel, so we could exclude batch effects arising due to sample handling.

A possible explanation for this discrepancy could be that the post-mortem delay was significantly longer for CJD patients compared to the non-prion controls (p-value = 0.0094; unpaired two-tailed t-test). While the mean post-mortem delay was approximately 6.5 days for our sCJD samples, it was only 3 days for the non-prion controls, a difference in means of 3.5 days.

Another possibility was that RNA quality was affected by prion disease. Studies have shown that total RNA quality was lower in the post-mortem AD human brain, and this affected mRNA quantification; however, this was not true for Parkinson’s disease or Huntington’s disease cases (Highet et al., 2021). In addition, a careful observation of the RIN values of samples from the Norsworthy et al. study reveals that RIN numbers for blood RNA from sCJD patients is consistently lower than age-matched controls (discovery phase RIN: sCJD samples = 5.6, SD = 1.3; control samples = 6.5, SD = 1.2. Replication phase RIN: sCJD samples = 5.8, SD = 1.8; control samples = 6.8, SD = 1) (Norsworthy et al., 2020). This evidence is not conclusive since there could be multiple confounding factors like differences in handling control and CJD samples, however, the possibility that the lower sample quality is a result of prion disease needs to be entertained.

This lack of well-preserved samples might be one of the factors that hindered the generation of high-quality single-cell datasets in human prion diseases and could partially explain the lack of relevant publications among other factors such as sample scarcity and human prion infectivity. In contrast, there have been single-cell studies on other human neurodegenerative diseases such as Alzheimer’s and Parkinson’s disease (Agarwal et al., 2020; Bryois et al., 2020; Mathys et al., 2019), indicating that single-nucleus sequencing of the post-mortem human brain is achievable. These studies used droplet-based approaches for library preparation raising the possibility that our methodology might have exacerbated existing sample quality differences. These studies did not provide information regarding RIN numbers for the samples used.

Since the data generated from the sCJD samples were of unusable quality, we decided to proceed with the analysis of the biopsies only. However, when we performed dimensionality reduction and drew the UMAP plots we identified a strong bias in the cell transcriptomes. Cells were clustering together not based on their cell type, but based on their biological sample of origin, indicating the existence of some strong transcriptomic bias that drives cell clustering. This phenomenon was especially evident for one of the biopsy samples, which occupied space further away from all other samples. We interpreted this finding considering the tissue quality when dissecting to prepare the suspensions. Samples originating from these human brain biopsies had no visible grey matter areas, instead, they consisted mostly of fat and white matter. Since the amount of sample used was small and biopsy samples could not have been collected from the same brain regions, we attributed the UMAP representation to sampling bias. This spurious sample was ultimately removed from the dataset.

Based on these preliminary results we decided not to proceed with further analysis and interpretation of our data. The small number of samples and cells remaining — 2 sCJD and 3 control biopsies and fewer than 3000 cells — combined with an overall biased dataset would prohibit the interpretation of the data and could lead to erroneous conclusions. Based on these experiments, we believe that the way forward with human post-mortem sCJD brain samples necessitates more sensitive single-nucleus sequencing protocols, or even single-population or bulk sequencing approaches.

# 6 Conclusions and future directions

## 6.1 Conclusions

This thesis set forth to profile the transcriptional landscape of prion disease in three different systems — cell lines, mouse and human brain — dissecting disease progression and aiming to characterise the heterogeneity of cellular response to prion infection, assess overlapping gene expression patterns in different organisms, and uncover biological mechanisms of prion toxicity. Previous studies in neurodegenerative diseases suggested that cellular response to disease is, indeed, heterogeneous, with each cell type — and their subtypes — assuming distinct phenotypes the function and impact of which we have just started to understand.

To address the limitations of previous transcriptomics studies, we investigated the suitability of contemporary single-cell and single-nucleus sequencing approaches. We established two fundamentally different methodologies in our Unit, based on droplet encapsulation and split-pool barcoding of single nuclei. Along with the practical experiments, we also developed and tested in silico analytical pipelines that harnessed the power of our Unit's computational infrastructure to deliver reproducible results and software to allow us to explore, visualise, and query sizable scRNA-seq datasets to answer our scientific questions.

We put our physical and computational methods to the test by profiling uninfected and chronically prion-infected cell lines. Even though we were not able to identify any transcriptomic effects of prion infection, these preliminary experiments allowed us to assess the suitability of the methodologies for future studies. We concluded that SPLiT-seq was more suitable for use with infectious material and compatible with BSL-3 working practices, while it also allowed the processing of multiple frozen samples in parallel.

We then designed a tightly controlled time-course mouse experiment and applied our previous knowledge to transcriptionally profile the frontal lobe of 95 mice in single-cell resolution. This was the first single-cell transcriptomics study of RML prions in rodent models and generated a breadth of information that can be used as a reference for future experiments. We found evidence supporting our hypothesis that cellular response to disease is heterogenous and identified cell populations that differently respond to inoculation with infectious material and prion propagation. We did not find evidence of selective toxicity contemplating that single-cell studies might not be the most sensitive tool for the quantification of small fluctuations in the numbers of cell populations. In accordance with previous research and published data, we were able to identify activated glial populations and an especially strong astrocytic signature, as well as activated homeostatic mechanisms, especially at the disease end-stage.

Some of our more interesting and unexpected findings included the observation that prion infectivity does not elicit a transcriptomic response in vivo, which supports the hypothesis that the infectious and the toxic prion species are different entities. In our rodent model, we described a triphasic transcriptomic response to prion infection where the early disease stages mirrored the response of the end-stage in lower amplitude, suggesting that the system could recover after the external introduction of toxic species until these species replicated and reached the threshold titres where cellular response became evident once again (at 80 and 120 dpi).

Our pathway analyses uncovered biological pathways perturbed across different cell types, with synaptic perturbations being the hallmark of cellular response to prion infection and toxicity. In addition, we identified dysregulated mechanisms of cell junction formation, ion transport, cell adhesion and pathways of excitotoxicity, while we did not find evidence of cell death.

We proceeded to apply our methodology to further characterise human prion disease, with, however, limited success. We tapped into our Unit’s resources and collaborators to acquire the best-kept post-mortem sporadic CJD brain samples and age and gender-matched controls, in an experiment carefully designed to control variables such as the *Prnp* genotype, sample handling batch effects, and brain region sampling differences. The highlight of the human study was the use of extremely precious archived sCJD human brain biopsies and non-neurological control brain biopsies. Even though we were not able to generate useful transcriptomic information, we did demonstrate that different methodological approaches are needed to assay these archived human brain samples and special considerations must be made to ensure sample quality.

Overall, we generated a rich and novel resource that includes transcriptomic and histopathological information which will be available to the scientific community. It is our hope that it will be further explored and utilised to provide answers to scientific questions, facilitate the design of future targeted experiments, act as an example of correctly controlled experimental design, and, more importantly, raise subsequent questions and stimulate curiosity in the exciting field of prion and prion-like diseases.

## 6.2 Future directions

As expected from novel, unbiased studies, our research has probably raised more questions than the ones it set out to answer. Our datasets can be used as a starting point for further exploration and the generation of interesting hypotheses that would require additional experimentation to be evaluated.

One of the points that became increasingly clear throughout our research was that frozen nuclei from archived brain tissue do not contain RNA of high enough quantity and quality to provide deep insights into biological mechanisms that involve genes expressed in lower levels. An interesting — although technologically challenging — follow-up of our work would be the deep sequencing of sorted single populations. One of the caveats is the successful dissociation of the frozen tissue to release nuclei without damaging their structure and nucleic acids, as well as the selection of protein markers for flow-cytometry-based cell sorting, especially when dealing with nuclei instead of whole cells. The ideal experiment would involve a freshly isolated mouse brain that is enzymatically and mechanically dissociated to get single live cells (not nuclei) that can then be sequenced in depth using sensitive, low-throughput protocols, like Smart-seq2. Unfortunately, the logistics of processing multiple samples as quickly as possible make these experiments complicated, while this methodology would not apply to any archived (i.e., frozen) brain tissue, excluding, thus, the use of human samples.

An alternative approach that could eliminate the need of using human brain tissue, while also providing a suitable model to study prion diseases would be the use of human brain organoids. Brain organoids are three-dimensional tissues generated from human embryonic stem cells that can recapitulate aspects of the in vivo physiology and architecture of the human brain (N. Sun et al., 2021). They can act as a useful tool to provide non-invasive access to patient-derived human tissue and enable studies of human brain development, brain cancer and neurodegenerative diseases. Proof-of-principle studies have demonstrated their utility in studying Alzheimer's and Parkinson's diseases and ALS (Choi et al., 2020; H. Kim et al., 2019; Osaki et al., 2018). In addition, some progress has also been made in the field of prion diseases, where human organoids have been shown to become infected with and accumulate human prions (Groveman et al., 2021). While caveats need to be taken into account (heterogeneity of the models, absence of non-neuronal linage cells, size constraints), infecting with prions and single-cell sequencing these models might provide a powerful alternative to profiling archived human brain tissue, especially important in rare diseases, such as sCJD.

Based on our quality control assays, the only feasible way to generate sequencing libraries from the archived human sCJD brain samples would be to perform bulk sequencing of specific anatomical regions of the brain, such as the superior frontal gyrus of the cortex. Even though such methods are unable to provide data at single-cell resolution, the gentler processing of the samples (no dissociation required) and the simultaneous sequencing of the RNA content of tens of millions of cells would probably allow the preparation of higher-quality sequencing libraries. Then, one could envisage those specific genes of interest identified in the mouse study relevant to specific cell populations could be examined in the bulk dataset, allowing to make some inferences regarding cell population behaviour. Furthermore, no next-generation sequencing data is available for the prion-infected human brain.

A caveat for all possible experiments discussed would be that they need to include ways to mitigate human prion infectivity. As we have found, decontamination methods are usually not compatible with sequencing protocols and health and safety guidelines do restrict the experimental freedom that is taken for granted in other neurodegenerative diseases.

Focusing on mouse experiments, which are easier to control and do not require BSL-3 work, one of the more interesting experiments would aim to transcriptionally characterise the activation state of microglia in RML-infected mice. Due to their small numbers, microglia were not detectable in our study, however, their percentage can be increased with fluorescence-activated or magnetic sorting methods. Then the enriched population can be sequenced either in bulk or using sensitive single-cell methods as discussed in (H. Wang, 2021) for Alzheimer's disease.

In addition, another option to study the effect of disease in a controlled way would be to assay cellular response in regions of the mouse brain that are more and less affected by the disease. As discussed in section 4.2.2, prion pathology spreads gradually throughout the mouse brain, so, especially at time points around 120 dpi, some brain regions are expected to have extensive abnormal PrP deposition and vacuolation, while others will be less affected. Microdissection of these regions and following dissociation and sequencing using bulk or single-cell methods could provide insights into early disease mechanisms and biological pathways that control cell susceptibility to prion disease. Similar studies in AD have identified differences in the phenotype and number of astrocytes around amyloid plaques (G. R. Frost & Li, 2017; Perez-Nievas & Serrano-Pozo, 2018).

Another way to assess prion spread in the mouse brain and accompanying transcriptomic response would be to perform spatially resolved transcriptomics (Nature method of the year 2020) ("Method of the Year 2020: spatially resolved transcriptomics.," 2021). These ground-breaking methods that empower large consortia such as the Human Cell Atlas allow capturing of both transcriptomic and anatomical information and, when coupled with immunohistochemical staining, could be used to investigate the cellular response to prion infection in multiple brain regions, and, possibly, the correlation of gene expression patterns with prion pathology across the whole mouse brain. One of the caveats of these methods is that the sensitivity is usually lower than single-cell approaches, so fewer transcripts may be identified.

Prion biology is further complicated due to the existence of multiple prion strains. In this thesis, we only focused on the RML murine prion strain to minimise the variables considered in our experiments, however, similar research could be performed for different mouse strains. These experiments could highlight the specificity of cellular response to each prion strain; however, the transcriptomic perturbations identified may be evident of a general response to neurodegenerative disease and not relevant to the RML prion strain. In addition, previous work has shown that the inflammatory response is similar in 3 murine prion strains (Carroll et al., 2015, 2016).

One of the central findings of our research was that transcriptomic perturbations proceed in three phases: they become evident early after inoculation, then follows a period of transcriptomic silence, before increasing again at 120 dpi and becoming more pronounced at the end-stage. Based on the infectivity data, we hypothesised that prion infectivity does not elicit a widespread transcriptomic response, which is caused by a toxic prion species. To assess whether this pattern is dependent on the host PrP levels, our time-course experiments could be replicated in PrP-overexpressing mouse lines. If these perturbations are caused by a toxic prion species we would expect these changes to appear more quickly, based on the research by (Sandberg et al., 2014). In addition, recent efforts in our Unit are expected to lead to the isolation of the lethal PrP species, termed $PrP^{L}$. An interesting experiment would involve direct intracerebral inoculation of mice with this purified lethal species in order to assay the transcriptomic response. These experiments could allow the discrimination between prion-specific transcriptomic perturbations and changes relevant to global neuroinflammation. A similar experiment that could help dissect mechanisms of toxicity would be to study the cellular response of *Prnp*-null mice to RML inoculation. Activated gene networks could also shed some light on prion clearance mechanisms, potentially contributing to the search for disease-modifying drugs.

Finally, some technological advances would be required to assay the correlation between single-cell transcriptomics and prion infectivity. As we demonstrated in our study, prion infectivity is lost when nucleus suspensions are prepared. Future experiments with fresh mouse brain tissue which can be more easily dissociated to prepare single-cell suspensions (not nucleus) could allow us to assay the prion infectivity of each sample. This would be especially interesting in single populations which could be sorted based on the surface PrP expression of each cell. A comparison of the PrP-rich and PrP-poor cell populations could uncover gene expression changes relevant to PrP expression. A more ambitious experiment could involve methodologies that allow assaying two modalities in single-cell resolution, namely infectivity and gene expression. A protocol that can generate transcriptomic and prion infectivity information for each cell in parallel is being developed in our Unit, and we are excited to explore all the different experimental possibilities that would be unlocked when this system is operational.

# Acknowledgements

With this PhD coming to an end, I am left with vivid memories of the last four years that included a lot of ups and downs, “aha!” moments, covid restrictions and uncertainty, gratification from working experiments, reunions with colleagues, laughs with new and old friends, travelling, picnics in the park, and lots of pub visits.

My journey could not have been possible if it wasn’t for my primary supervisor, Emmanuelle Viré, who took me in as an Erasmus student and made me the scientist that I am today through sharing her knowledge, mentoring, and jokes. Her enthusiasm for science is infectious and I am so grateful that our paths met because I now have a good new friend. I am also extremely grateful to my secondary supervisor, Simon Mead, who was there for me during the easy and challenging times, always ready to help and make things happen. Our conversations were the best academic lessons for me, and his knowledge deeply impacted this PhD. I want to thank the rest of the Human Genetics group, past and present, for always providing helpful advice and their time and making our office a great and fun workplace. Thanks to Tracy Campbell, Penny Norsworthy, Lee Darwent, Fernando Guntoro, Liam Quinn, Chiara Giordani, Helen Speedy, Elizabeth Hill, Nour Majbour, and Isabel Natho. Special thanks to Holger Hummerich who was always willing to troubleshoot our server’s issues and had to put up with my requests for new software, Fuquan Zhang whose skills and scientific thinking greatly contributed to this project, and our rotating students Tom Murphy and Tom Trainer who contributed to this research and embraced the project as their own. I would also like to give special thanks to our Unit’s director, Professor John Collinge, who believed in my project and always provided constructive feedback during our meetings.

This project immensely benefited from taking place in the collaborative environment of our Unit. I would like to thank the histology team and pathologists Tamsin Nazari, Florin Pintilii, Conor Preston, Fabio Argentina, Jackie Linehan, and Sebastian Brandner; Malin Sandberg, Huda Al-Doujaily, and Michael De Oliveira for advice about the design of the mouse experiments and sourcing mouse brain homogenates; Jonathan Wadsworth for his excellent training and keeping me safe from infectious prions; Stephanie Canning and Azy Khalili for their time and advice regarding cell sorting; Christian Schmidt, George

Thirlway, and Parvin Ahmed for their work on the scrapie cell assay; all of the staff at the animal facility for their expertise, work and care with the mice in this study, including Lucy Draper, Nick Kaye, Craig Fitzhugh, and Gavin Graham for their first-class work in the mouse experiment; our Unit's staff who took care of everything that I needed as a PhD student including Samantha MacLeod, David Ruegg, Amir Mayahi, Norie Chiba, Damian Johnson, Ryan Peter, Blanche Kollo. Everyone has been truly kind and generous with sharing their time and expertise. Special thanks to Zane Jaunmuktane for providing her expertise on prion histopathology and coordinating the receipt of human control samples and to the UCL genomics facility for helping me with sequencing and providing their expertise, especially Tony Brooks and Paola Niola.

Many thanks to our collaborators, who discussed early findings, provided their expertise and shaped this PhD, including Gonçalo Castelo-Branco, and Mark Kristiansen.

I am grateful to all patients and families who consented to their samples being used in the study, to the National Prion Clinic for their help acquiring samples and patients' meta-data, and to the Queen Square Brain Bank and BRAIN UK for kindly providing human control samples.

A huge thank you goes to current and previous students of our Unit who were always there to talk, play board games, enjoy picnics at the park, share stories and go for trips to the local pub. Fernando Guntoro, Chiara Giordani, Lee Darwent, Ines Whitworth, Liam Quinn, Alexandra Philiastides, Hazim Halim, Tom Murphy, Tom Trainer, and Luke Dabin thank you all for the laughs and for making me feel welcome since the first day I joined.

I've left the most important people for the end. The biggest thank you goes to my family, Konstantia Roumeliotou, Dimitris Dimitriadis, Aikaterini Roumeliotou, and Eleni Dimitriadou, that have been extremely supportive since my first memories, believe in me and have always done everything in their power to allow me to fulfil my dreams. Our family dog Truffle has been my antianxiety therapist whenever I visited my family, she is the best girl and deserves all the pets. Sofia Chrysikopoulou, you've been by my side through my whole career so far, and you've always been there to cheer me up when things didn't go well and never stopped believing in me, and for that, I will always be grateful. Thanks to George Sideris and Eleni Exarchou who have been our amazing friends navigating

London life and were always there to discuss science, exchange ideas, get through the pandemic, and have fun; to Sotiria Kalpachtsi, my dearest friend and London nightlife expert; to Stelios Iliadis, Aris Sionakidis, and Alexandros Pavlaras my best mates who are always there for me and I miss deeply. To all my friends and family, I could have not done this without you.

This work was funded by the UK Medical Research Council to the MRC Prion Unit at UCL with additional funding from Alzheimer’s Research UK. I would also like to thank the Onassis Foundation for supporting my work and making me a scholar and part of their prestigious community.

# References

Abate, A. R., Chen, C.-H., Agresti, J. J., & Weitz, D. A. (2009). Beating Poisson encapsulation statistics using close-packed ordering. *Lab on A Chip*, *9*(18), 2628–2631. https://doi.org/10.1039/b909386a

Ables, J. L., Breunig, J. J., Eisch, A. J., & Rakic, P. (2011). Not(ch) just development: Notch signalling in the adult brain. *Nature Reviews. Neuroscience*, *12*(5), 269–283. https://doi.org/10.1038/nrn3024

Acosta, C., Anderson, H. D., & Anderson, C. M. (2017). Astrocyte dysfunction in Alzheimer disease. *Journal of Neuroscience Research*, *95*(12), 2430–2447. https://doi.org/10.1002/jnr.24075

Adey, A., Kitzman, J. O., Burton, J. N., Daza, R., Kumar, A., Christiansen, L., Ronaghi, M., Amini, S., Gunderson, K. L., Steemers, F. J., & Shendure, J. (2014). In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Research*, *24*(12), 2041–2049. https://doi.org/10.1101/gr.178319.114

Agarwal, D., Sandor, C., Volpato, V., Caffrey, T. M., Monzón-Sandoval, J., Bowden, R., Alegre-Abarrategui, J., Wade-Martins, R., & Webber, C. (2020). A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nature Communications*, *11*(1), 4183. https://doi.org/10.1038/s41467-020-17876-0

Aguzzi, A., Barres, B. A., & Bennett, M. L. (2013). Microglia: scapegoat, saboteur, or something else? *Science*, *339*(6116), 156–161. https://doi.org/10.1126/science.1227901

Aguzzi, A., & Falsig, J. (2012). Prion propagation, toxicity and degradation. *Nature Neuroscience*, *15*(7), 936–939. https://doi.org/10.1038/nn.3120

Aguzzi, A., Heikenwalder, M., & Polymenidou, M. (2007). Insights into prion strains and neurotoxicity. *Nature Reviews. Molecular Cell Biology*, *8*(7), 552–561. https://doi.org/10.1038/nrm2204

Aguzzi, A., & Zhu, C. (2017). Microglia in prion diseases. *The Journal of Clinical Investigation*.

Ahlmann-Eltze, C., & Huber, W. (2021). glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics*, *36*(24), 5701–5702. https://doi.org/10.1093/bioinformatics/btaa1009

Aitchison, J. (2008). The single principle of compositional data analysis, continuing fallacies, confusionsand misunderstandings and some suggested remedies. *Undefined*.

Ajami, B., Samusik, N., Wieghofer, P., Ho, P. P., Crotti, A., Bjornson, Z., Prinz, M., Fantl, W. J., Nolan, G. P., & Steinman, L. (2018). Single-cell mass cytometry reveals distinct populations of brain myeloid cells in mouse neuroinflammation and neurodegeneration models. *Nature Neuroscience*, *21*(4), 541–551. https://doi.org/10.1038/s41593-018-0100-x

Alibhai, J., Blanco, R. A., Barria, M. A., Piccardo, P., Caughey, B., Perry, V. H., Freeman, T. C., & Manson, J. C. (2016). Distribution of misfolded prion protein seeding activity alone does not predict regions of neurodegeneration. *PLoS Biology*, *14*(11), e1002579. https://doi.org/10.1371/journal.pbio.1002579

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., & Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, *17*(2), 137–145. https://doi.org/10.1038/s41592-019-0654-x

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K. L., & Steemers, F. J. (2014). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature Genetics*, *46*(12), 1343–1349. https://doi.org/10.1038/ng.3119

Andrews, S. (2010). *FastQC* [Computer software].

Arellano-Anaya, Z. E., Savistchenko, J., Mathey, J., Huor, A., Lacroux, C., Andréoletti, O., & Vilette, D. (2011). A simple, versatile and sensitive cell-based assay for prions from various species. *Plos One*, *6*(5), e20563. https://doi.org/10.1371/journal.pone.0020563

Asante, E. A., Linehan, J. M., Desbruslais, M., Joiner, S., Gowland, I., Wood, A. L., Welch, J., Hill, A. F., Lloyd, S. E., Wadsworth, J. D. F., & Collinge, J. (2002). BSE prions propagate as either variant CJD-like or sporadic CJD-like prion strains in transgenic mice expressing human prion protein. *The EMBO Journal*, *21*(23), 6358–6366. https://doi.org/10.1093/emboj/cdf653

Asante, E. A., Smidak, M., Grimshaw, A., Houghton, R., Tomlinson, A., Jeelani, A., Jakubcova, T., Hamdan, S., Richard-Londt, A., Linehan, J. M., Brandner, S., Alpers, M., Whitfield, J., Mead, S., Wadsworth, J. D. F., & Collinge, J. (2015). A naturally occurring variant of the human prion protein completely prevents prion disease. *Nature*, *522*(7557), 478–481. https://doi.org/10.1038/nature14510

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Atarashi, R., Wilham, J. M., Christensen, L., Hughson, A. G., Moore, R. A., Johnson, L. M., Onwubiko, H. A., Priola, S. A., & Caughey, B. (2008). Simplified ultrasensitive prion detection by recombinant PrP conversion with shaking. *Nature Methods*, *5*(3), 211–212. https://doi.org/10.1038/nmeth0308-211

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M., & Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, *7*(1), 16878. https://doi.org/10.1038/s41598-017-17204-5

Barcia, C., Ros, C. M., Annese, V., Gómez, A., Ros-Bernal, F., Aguado-Llera, D., Martínez-Pagán, M. E., de Pablos, V., Fernandez-Villalba, E., & Herrero, M. T. (2012). IFN-γ signaling, with the synergistic contribution of TNF-α, mediates cell specific microglial and astroglial activation in experimental models of Parkinson's disease. *Cell Death & Disease*, *3*, e379. https://doi.org/10.1038/cddis.2012.123

Basu, A., Basu, A., Avraham-Davidi, I., Habib, N., Regev, A., Zhang, F., Shekhar, K., Hofree, M., Weitz, D., Rozenblatt-Rosen, O., Burks, T., Choudhury, S., Aguet, F., Gelfand, E., & Ardlie, K. (2017). DroNc-seq step-by-step. *Protocol Exchange*. https://doi.org/10.1038/protex.2017.094

Basu, U., Almeida, L., Olson, N. E., Meng, Y., Williams, J. L., Moore, S. S., & Guan, L. L. (2011). Transcriptome analysis of the medulla tissue from cattle in response to bovine spongiform encephalopathy using digital gene expression tag profiling. *Journal of Toxicology and Environmental Health. Part A*, *74*(2–4), 127–137. https://doi.org/10.1080/15287394.2011.529062

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*, 38–44. https://doi.org/10.1038/nbt.4314

Belay, E. D. (1999). Transmissible spongiform encephalopathies in humans. *Annual Review of Microbiology*, *53*, 283–314. https://doi.org/10.1146/annurev.micro.53.1.283

Benilova, I., Reilly, M., Terry, C., Wenborn, A., Schmidt, C., Marinho, A. T., Risse, E., Al-Doujaily, H., Wiggins De Oliveira, M., Sandberg, M. K., Wadsworth, J. D. F., Jat, P. S., & Collinge, J. (2020). Highly infectious prions are not directly neurotoxic. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(38), 23815–23822. https://doi.org/10.1073/pnas.2007406117

Berry, D. B., Lu, D., Geva, M., Watts, J. C., Bhardwaj, S., Oehler, A., Renslo, A. R., DeArmond, S. J., Prusiner, S. B., & Giles, K. (2013). Drug resistance confounding prion therapeutics. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(44), E4160-9. https://doi.org/10.1073/pnas.1317164110

Bessen, R. A., Kocisko, D. A., Raymond, G. J., Nandan, S., Lansbury, P. T., & Caughey, B. (1995). Non-genetic propagation of strain-specific properties of scrapie prion protein. *Nature*, *375*(6533), 698–700. https://doi.org/10.1038/375698a0

Betmouni, S., Perry, V. H., & Gordon, J. L. (1996). Evidence for an early inflammatory response in the central nervous system of mice with scrapie. *Neuroscience*, *74*(1), 1–5. https://doi.org/10.1016/0306-4522(96)00212-6

Bhatia, S., Jenner, A. M., Li, H., Ruberu, K., Spiro, A. S., Shepherd, C. E., Kril, J. J., Kain, N., Don, A., & Garner, B. (2013). Increased apolipoprotein D dimer formation in Alzheimer's disease hippocampus is associated with lipid conjugated diene levels. *Journal of Alzheimer's Disease*, *35*(3), 475–486. https://doi.org/10.3233/JAD-122278

Bian, J., Napier, D., Khaychuck, V., Angers, R., Graham, C., & Telling, G. (2010). Cell-based quantification of chronic wasting disease prions. *Journal of Virology*, *84*(16), 8322–8326. https://doi.org/10.1128/JVI.00633-10

Birkett, C. R., Hennion, R. M., Bembridge, D. A., Clarke, M. C., Chree, A., Bruce, M. E., & Bostock, C. J. (2001). Scrapie strains maintain biological phenotypes on propagation in a cell line in culture. *The EMBO Journal*, *20*(13), 3351–3358. https://doi.org/10.1093/emboj/20.13.3351

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bollen, E., & Prickaerts, J. (2012). Phosphodiesterases in neurodegenerative disorders. *IUBMB Life*, *64*(12), 965–970. https://doi.org/10.1002/iub.1104

Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, *18*(1), 77. https://doi.org/10.1186/s13059-017-1212-4

Bonnycastle, L. L., Gildea, D. E., Yan, T., Narisu, N., Swift, A. J., Wolfsberg, T. G., Erdos, M. R., & Collins, F. S. (2020). Single-cell transcriptomics from human pancreatic islets: sample preparation matters. *Biology Methods and Protocols*, *5*(1), bpz019. https://doi.org/10.1093/biomethods/bpz019

Booth, S., Bowman, C., Baumgartner, R., Sorensen, G., Robertson, C., Coulthart, M., Phillipson, C., & Somorjai, R. L. (2004). Identification of central nervous system genes involved in the host response to the scrapie agent during preclinical and clinical infection. *The Journal of General Virology*, *85*(Pt 11), 3459–3471. https://doi.org/10.1099/vir.0.80110-0

Boswell-Smith, V., Spina, D., & Page, C. P. (2006). Phosphodiesterase inhibitors. *British Journal of Pharmacology*, *147 Suppl 1*, S252-7. https://doi.org/10.1038/sj.bjp.0706495

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, *20*(18), 3710–3715. https://doi.org/10.1093/bioinformatics/bth456

Brandner, S, Isenmann, S., Raeber, A., Fischer, M., Sailer, A., Kobayashi, Y., Marino, S., Weissmann, C., & Aguzzi, A. (1996). Normal host prion protein necessary for scrapie-induced neurotoxicity. *Nature*, *379*(6563), 339–343. https://doi.org/10.1038/379339a0

Brandner, Sebastian, & Jaunmuktane, Z. (2017). Prion disease: experimental models and reality. *Acta Neuropathologica*, *133*(2), 197–222. https://doi.org/10.1007/s00401-017-1670-5

Branton, D. (2016). Fracture faces of frozen membranes: 50th anniversary. *Molecular Biology of the Cell*, *27*(3), 421–423. https://doi.org/10.1091/mbc.E15-05-0287

Bremer, J., Baumann, F., Tiberi, C., Wessig, C., Fischer, H., Schwarz, P., Steele, A. D., Toyka, K. V., Nave, K.-A., Weis, J., & Aguzzi, A. (2010). Axonal prion protein is required for peripheral myelin maintenance. *Nature Neuroscience*, *13*(3), 310–318. https://doi.org/10.1038/nn.2483

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., & Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, *10*(11), 1093–1095. https://doi.org/10.1038/nmeth.2645

Brown, P., Gibbs, C. J., Rodgers-Johnson, P., Asher, D. M., Sulima, M. P., Bacote, A., Goldfarb, L. G., & Gajdusek, D. C. (1994). Human spongiform encephalopathy: the National Institutes of Health series of 300 cases of experimentally transmitted disease. *Annals of Neurology*, *35*(5), 513–529. https://doi.org/10.1002/ana.410350504

Bruce, M. E., McConnell, I., Fraser, H., & Dickinson, A. G. (1991). The disease characteristics of different strains of scrapie in Sinc congenic mouse lines: implications for the nature of the agent and host control of pathogenesis. *The Journal of General Virology*, *72 ( Pt 3)*, 595–603. https://doi.org/10.1099/0022-1317-72-3-595

Bruce, M. E. (1993). Scrapie strain variation and mutation. *British Medical Bulletin*, *49*(4), 822–838. https://doi.org/10.1093/oxfordjournals.bmb.a072649

Bryois, J., Skene, N. G., Hansen, T. F., Kogelman, L. J. A., Watson, H. J., Liu, Z., Eating Disorders Working Group of the Psychiatric Genomics Consortium, International Headache Genetics Consortium, 23andMe Research Team, Brueggeman, L., Breen, G., Bulik, C. M., Arenas, E., Hjerling-Leffler, J., & Sullivan, P. F. (2020). Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson’s disease. *Nature Genetics*, *52*(5), 482–493. https://doi.org/10.1038/s41588-020-0610-9

Büeler, H., Aguzzi, A., Sailer, A., Greiner, R. A., Autenried, P., Aguet, M., & Weissmann, C. (1993). Mice devoid of PrP are resistant to scrapie. *Cell*, *73*(7), 1339–1347. https://doi.org/10.1016/0092-8674(93)90360-3

Büeler, H., Fischer, M., Lang, Y., Bluethmann, H., Lipp, H. P., DeArmond, S. J., Prusiner, S. B., Aguet, M., & Weissmann, C. (1992). Normal development and behaviour of mice lacking the neuronal cell-surface PrP protein. *Nature*, *356*(6370), 577–582. https://doi.org/10.1038/356577a0

Burns, T. C., Li, M. D., Mehta, S., Awad, A. J., & Morgan, A. A. (2015). Mouse models rarely mimic the transcriptome of human neurodegenerative diseases: A systematic bioinformatics-based critique of preclinical models. *European Journal of Pharmacology*, *759*, 101–117. https://doi.org/10.1016/j.ejphar.2015.03.021

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096

Butler, D. A., Scott, M. R., Bockman, J. M., Borchelt, D. R., Taraboulos, A., Hsiao, K. K., Kingsbury, D. T., & Prusiner, S. B. (1988). Scrapie-infected murine neuroblastoma cells produce protease-resistant prion proteins. *Journal of Virology*, *62*(5), 1558–1564. https://doi.org/10.1128/JVI.62.5.1558-1564.1988

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., & Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, *16*(1), 43–49. https://doi.org/10.1038/s41592-018-0254-1

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., & Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, *357*(6352), 661–667. https://doi.org/10.1126/science.aam8940

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

Carroll, J. A., Race, B., Williams, K., Striebel, J., & Chesebro, B. (2020). RNA-seq and network analysis reveal unique glial gene expression signatures during prion infection. *Molecular Brain*, *13*(1), 71. https://doi.org/10.1186/s13041-020-00610-8

Carroll, J. A., Striebel, J. F., Race, B., Phillips, K., & Chesebro, B. (2015). Prion infection of mouse brain reveals multiple new upregulated genes involved in neuroinflammation or signal transduction. *Journal of Virology*, *89*(4), 2388–2404. https://doi.org/10.1128/JVI.02952-14

Carroll, J. A., Striebel, J. F., Rangel, A., Woods, T., Phillips, K., Peterson, K. E., Race, B., & Chesebro, B. (2016). Prion Strain Differences in Accumulation of PrPSc on Neurons and Glia Are Associated with Similar Expression Profiles of Neuroinflammatory Genes: Comparison of Three Prion Strains. *PLoS Pathogens*, *12*(4), e1005551. https://doi.org/10.1371/journal.ppat.1005551

Chakrabarti, O., & Hegde, R. S. (2009). Functional depletion of mahogunin by cytosolically exposed prion protein contributes to neurodegeneration. *Cell*, *137*(6), 1136–1147. https://doi.org/10.1016/j.cell.2009.03.042

Chandler, R. L. (1961). Encephalopathy in mice produced by inoculation with scrapie brain material. *The Lancet*, *1*(7191), 1378–1379.

Chandler, R. L. (1963). Experimental scrapie in the mouse. *Research in Veterinary Science*, *4*(2), 276–285. https://doi.org/10.1016/S0034-5288(18)34870-7

Charles A Janeway, J., Travers, P., Walport, M., & Shlomchik, M. J. (2001). *The complement system and innate immunity*.

Chasseigneaux, S., Pastore, M., Britton-Davidian, J., Manié, E., Stern, M.-H., Callebert, J., Catalan, J., Casanova, D., Belondrade, M., Provansal, M., Zhang, Y., Bürkle, A., Laplanche, J.-L., Sévenet, N., & Lehmann, S. (2008). Genetic heterogeneity versus molecular analysis of prion susceptibility in neuroblasma N2a sublines. *Archives of Virology*, *153*(9), 1693–1702. https://doi.org/10.1007/s00705-008-0177-8

Chen, G., Ning, B., & Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in Genetics*, *10*, 317. https://doi.org/10.3389/fgene.2019.00317

Chen, G., Schell, J. P., Benitez, J. A., Petropoulos, S., Yilmaz, M., Reinius, B., Alekseenko, Z., Shi, L., Hedlund, E., Lanner, F., Sandberg, R., & Deng, Q. (2016). Single-cell analyses of X

Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Research*, *26*(10), 1342–1354. https://doi.org/10.1101/gr.201954.115

Chen, S.-K., Tvrdik, P., Peden, E., Cho, S., Wu, S., Spangrude, G., & Capecchi, M. R. (2010). Hematopoietic origin of pathological grooming in Hoxb8 mutant mice. *Cell*, *141*(5), 775–785. https://doi.org/10.1016/j.cell.2010.03.055

Cherry, J. D., Olschowka, J. A., & O'Banion, M. K. (2014). Neuroinflammation and M2 microglia: the good, the bad, and the inflamed. *Journal of Neuroinflammation*, *11*, 98. https://doi.org/10.1186/1742-2094-11-98

Choi, H., Kim, H. J., Yang, J., Chae, S., Lee, W., Chung, S., Kim, J., Choi, H., Song, H., Lee, C. K., Jun, J. H., Lee, Y. J., Lee, K., Kim, S., Sim, H.-R., Choi, Y. I., Ryu, K. H., Park, J.-C., Lee, D., … Mook-Jung, I. (2020). Acetylation changes tau interactome to degrade tau in Alzheimer's disease animal and organoid models. *Aging Cell*, *19*(1), e13081. https://doi.org/10.1111/acel.13081

Chung, W.-S., Allen, N. J., & Eroglu, C. (2015). Astrocytes control synapse formation, function, and elimination. *Cold Spring Harbor Perspectives in Biology*, *7*(9), a020370. https://doi.org/10.1101/cshperspect.a020370

Clarke, B. E., & Patani, R. (2020). The microglial component of amyotrophic lateral sclerosis. *Brain: A Journal of Neurology*, *143*(12), 3526–3539. https://doi.org/10.1093/brain/awaa309

Clarke, M. C., & Haig, D. A. (1970). Evidence for the multiplication of scrapie agent in cell culture. *Nature*, *225*(5227), 100–101. https://doi.org/10.1038/225100a0

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(21), 7426–7431. https://doi.org/10.1073/pnas.0500334102

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., & Yosef, N. (2019). Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*, *8*(4), 315-328.e8. https://doi.org/10.1016/j.cels.2019.03.010

Collinge, J, Palmer, M. S., & Dryden, A. J. (1991). Genetic predisposition to iatrogenic Creutzfeldt-Jakob disease. *The Lancet*, *337*(8755), 1441–1442. https://doi.org/10.1016/0140-6736(91)93128-v

Collinge, J, Sidle, K. C., Meads, J., Ironside, J., & Hill, A. F. (1996). Molecular analysis of prion strain variation and the aetiology of "new variant" CJD. *Nature*, *383*(6602), 685–690. https://doi.org/10.1038/383685a0

Collinge, John, & Clarke, A. R. (2007). A general model of prion strains and their pathogenicity. *Science*, *318*(5852), 930–936. https://doi.org/10.1126/science.1138718

Collinge, J. (2001). Prion diseases of humans and animals: their causes and molecular basis. *Annual Review of Neuroscience*, *24*, 519–550. https://doi.org/10.1146/annurev.neuro.24.1.519

Collin, J., Zerti, D., Queen, R., Santos-Ferreira, T., Bauer, R., Coxhead, J., Hussain, R., Steel, D., Mellough, C., Ader, M., Sernagor, E., Armstrong, L., & Lako, M. (2019). CRX Expression in Pluripotent Stem Cell-Derived Photoreceptors Marks a Transplantable Subpopulation of Early Cones. *Stem Cells*, *37*(5), 609–622. https://doi.org/10.1002/stem.2974

Comoy, E. E., Mikol, J., Jaffré, N., Lebon, V., Levavasseur, E., Streichenberger, N., Sumian, C., Perret-Liaudet, A., Eloit, M., Andreoletti, O., Haïk, S., Hantraye, P., & Deslys, J.-P. (2017). Experimental transfusion of variant CJD-infected blood reveals previously uncharacterised prion disorder in mice and macaque. *Nature Communications*, *8*(1), 1268. https://doi.org/10.1038/s41467-017-01347-0

Comoy, E. E., Mikol, J., Luccantoni-Freire, S., Correia, E., Lescoutra-Etchegaray, N., Durand, V., Dehen, C., Andreoletti, O., Casalone, C., Richt, J. A., Greenlee, J. J., Baron, T., Benestad, S. L., Brown, P., & Deslys, J.-P. (2015). Transmission of scrapie prions to primate after an extended silent incubation period. *Scientific Reports*, *5*, 11573. https://doi.org/10.1038/srep11573

Comoy, E. E., Mikol, J., Ruchoux, M.-M., Durand, V., Luccantoni-Freire, S., Dehen, C., Correia, E., Casalone, C., Richt, J. A., Greenlee, J. J., Torres, J. M., Brown, P., & Deslys, J.-P. (2013). Evaluation of the zoonotic potential of transmissible mink encephalopathy. *Pathogens (Basel, Switzerland)*, *2*(3), 520–532. https://doi.org/10.3390/pathogens2030520

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Corraliza-Gomez, M., Sanchez, D., & Ganfornina, M. D. (2019). Lipid-Binding Proteins in Brain Health and Disease. *Frontiers in Neurology*, *10*, 1152. https://doi.org/10.3389/fneur.2019.01152

Costa, F., Grün, D., & Backofen, R. (2018). GraphDDP: a graph-embedding approach to detect differentiation pathways in single-cell-data using prior class knowledge. *Nature Communications*, *9*(1), 3685. https://doi.org/10.1038/s41467-018-05988-7

Courageot, M. P., Daude, N., Nonno, R., Paquet, S., Di Bari, M. A., Le Dur, A., Chapuis, J., Hill, A. F., Agrimi, U., Laude, H., & Vilette, D. (2008). A cell line infectible by prion strains from different species. *The Journal of General Virology*, *89*(Pt 1), 341–347. https://doi.org/10.1099/vir.0.83344-0

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, *348*(6237), 910–914. https://doi.org/10.1126/science.aab1601

Dassanayake, R. P., Zhuang, D., Truscott, T. C., Madsen-Bouterse, S. A., O'Rourke, K. I., & Schneider, D. A. (2016). A transfectant RK13 cell line permissive to classical caprine scrapie prion propagation. *Prion*, *10*(2), 153–164. https://doi.org/10.1080/19336896.2016.1166324

Dassati, S., Waldner, A., & Schweigreiter, R. (2014). Apolipoprotein D takes center stage in the stress response of the aging and degenerative brain. *Neurobiology of Aging*, *35*(7), 1632–1642. https://doi.org/10.1016/j.neurobiolaging.2014.01.148

de Magalhães, J. P., Curado, J., & Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, *25*(7), 875–881. https://doi.org/10.1093/bioinformatics/btp073

de Melo, A. S. L. F., Lima, J. L. D., Malta, M. C. S., Marroquim, N. F., Moreira, Á. R., de Almeida Ladeia, I., Dos Santos Cardoso, F., Gonçalves, D. B., Dutra, B. G., & Dos Santos, J. C. C. (2021). The role of microglia in prion diseases and possible therapeutic targets: a literature review. *Prion*, *15*(1), 191–206. https://doi.org/10.1080/19336896.2021.1991771

Del-Aguila, J. L., Li, Z., Dube, U., Mihindukulasuriya, K. A., Budde, J. P., Fernandez, M. V., Ibanez, L., Bradley, J., Wang, F., Bergmann, K., Davenport, R., Morris, J. C., Holtzman, D. M., Perrin, R. J., Benitez, B. A., Dougherty, J., Cruchaga, C., & Harari, O. (2019). A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain. *Alzheimer's Research & Therapy*, *11*(1), 71. https://doi.org/10.1186/s13195-019-0524-x

Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clement, O., Simmons, R. K., Lister, R., & Forrest, A. R. R. (2019). Systematic bias assessment in solid tissue 10x scRNA-seq workflows. *BioRxiv*. https://doi.org/10.1101/832444

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Domingues, H. S., Portugal, C. C., Socodato, R., & Relvas, J. B. (2016). Oligodendrocyte, astrocyte, and microglia crosstalk in myelin development, damage, and repair. *Frontiers in Cell and Developmental Biology*, *4*, 71. https://doi.org/10.3389/fcell.2016.00071

Dong, A., Liu, S., & Li, Y. (2018). Gap junctions in the nervous system: probing functional connections using new imaging approaches. *Frontiers in Cellular Neuroscience*, *12*, 320. https://doi.org/10.3389/fncel.2018.00320

Duò, A., Robinson, M. D., & Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, *7*, 1141. https://doi.org/10.12688/f1000research.15666.3

Edgeworth, J. A., Gros, N., Alden, J., Joiner, S., Wadsworth, J. D. F., Linehan, J., Brandner, S., Jackson, G. S., Weissmann, C., & Collinge, J. (2010). Spontaneous generation of mammalian prions. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14402–14406. https://doi.org/10.1073/pnas.1004036107

Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics & Bioinformatics / Beijing Genomics Institute*, *13*(1), 4–16. https://doi.org/10.1016/j.gpb.2015.01.009

Ferreira, P. G., Muñoz-Aguirre, M., Reverter, F., Sá Godinho, C. P., Sousa, A., Amadoz, A., Sodaei, R., Hidalgo, M. R., Pervouchine, D., Carbonell-Caballero, J., Nurtdinov, R., Breschi, A., Amador, R., Oliveira, P., Çubuk, C., Curado, J., Aguet, F., Oliveira, C., Dopazo, J., … Guigó, R. (2018). The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature Communications*, *9*(1), 490. https://doi.org/10.1038/s41467-017-02772-x

Fok-Seang, J., Mathews, G. A., ffrench-Constant, C., Trotter, J., & Fawcett, J. W. (1995). Migration of oligodendrocyte precursors on astrocytes and meningeal cells. *Developmental Biology*, *171*(1), 1–15. https://doi.org/10.1006/dbio.1995.1255

Fraser, H., & Dickinson, A. G. (1973). Scrapie in mice. *Journal of Comparative Pathology*, *83*(1), 29–40. https://doi.org/10.1016/0021-9975(73)90024-8

Freytag, S., Tian, L., Lönnstedt, I., Ng, M., & Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. [version 2; peer review: 3 approved]. *F1000Research*, *7*, 1297. https://doi.org/10.12688/f1000research.15809.2

Frost, B., & Diamond, M. I. (2010). Prion-like mechanisms in neurodegenerative diseases. *Nature Reviews. Neuroscience*, *11*(3), 155–159. https://doi.org/10.1038/nrn2786

Frost, G. R., & Li, Y.-M. (2017). The role of astrocytes in amyloid production and Alzheimer's disease. *Open Biology*, *7*(12). https://doi.org/10.1098/rsob.170228

Gao, C., Wei, J., Zhang, B.-Y., Shi, Q., Chen, C., Wang, J., Shi, Q., & Dong, X.-P. (2016). MiRNA expression profiles in the brains of mice infected with scrapie agents 139A, ME7 and S15. *Emerging Microbes & Infections*, *5*(11), e115. https://doi.org/10.1038/emi.2016.120

Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C., & Deplancke, B. (2017). ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, *33*(19), 3123–3125. https://doi.org/10.1093/bioinformatics/btx337

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Giese, A., Brown, D. R., Groschup, M. H., Feldmann, C., Haist, I., & Kretzschmar, H. A. (1998). Role of microglia in neuronal cell death in prion disease. *Brain Pathology*, *8*(3), 449–457. https://doi.org/10.1111/j.1750-3639.1998.tb00167.x

Giles, K., Berry, D. B., Condello, C., Dugger, B. N., Li, Z., Oehler, A., Bhardwaj, S., Elepano, M., Guan, S., Silber, B. M., Olson, S. H., & Prusiner, S. B. (2016). Optimization of Aryl Amides that Extend Survival in Prion-Infected Mice. *The Journal of Pharmacology and Experimental Therapeutics*, *358*(3), 537–547. https://doi.org/10.1124/jpet.116.235556

Giles, K., Berry, D. B., Condello, C., Hawley, R. C., Gallardo-Godoy, A., Bryant, C., Oehler, A., Elepano, M., Bhardwaj, S., Patel, S., Silber, B. M., Guan, S., DeArmond, S. J., Renslo, A. R., & Prusiner, S. B. (2015). Different 2-Aminothiazole Therapeutics Produce Distinct Patterns of

Scrapie Prion Neuropathology in Mouse Brains. *The Journal of Pharmacology and Experimental Therapeutics*, *355*(1), 2–12. https://doi.org/10.1124/jpet.115.224659

Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., & Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*, *518*(7539), 365–369. https://doi.org/10.1038/nature14252

Glöckner, F., & Ohm, T. G. (2003). Hippocampal apolipoprotein D level depends on Braak stage and APOE genotype. *Neuroscience*, *122*(1), 103–110. https://doi.org/10.1016/s0306-4522(03)00529-3

Gómez-Rubio, V. (2017). ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). *Journal of Statistical Software*, *77*(Book Review 2). https://doi.org/10.18637/jss.v077.b02

Goniotaki, D., Lakkaraju, A. K. K., Shrivastava, A. N., Bakirci, P., Sorce, S., Senatore, A., Marpakwar, R., Hornemann, S., Gasparini, F., Triller, A., & Aguzzi, A. (2017). Inhibition of group-I metabotropic glutamate receptors protects against prion toxicity. *PLoS Pathogens*, *13*(11), e1006733. https://doi.org/10.1371/journal.ppat.1006733

Goold, R., McKinnon, C., & Tabrizi, S. J. (2015). Prion degradation pathways: Potential for therapeutic intervention. *Molecular and Cellular Neurosciences*, *66*(Pt A), 12–20. https://doi.org/10.1016/j.mcn.2014.12.009

Greenwood, A. D., Horsch, M., Stengel, A., Vorberg, I., Lutzny, G., Maas, E., Schädler, S., Erfle, V., Beckers, J., Schätzl, H., & Leib-Mösch, C. (2005). Cell line dependent RNA expression profiles of prion-infected mouse neuronal cells. *Journal of Molecular Biology*, *349*(3), 487–500. https://doi.org/10.1016/j.jmb.2005.03.076

Griffiths, J. A., Scialdone, A., & Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, *14*(4), e8046. https://doi.org/10.15252/msb.20178046

Grindberg, R. V., Yee-Greenbaum, J. L., McConnell, M. J., Novotny, M., O'Shaughnessy, A. L., Lambert, G. M., Araúzo-Bravo, M. J., Lee, J., Fishman, M., Robbins, G. E., Lin, X., Venepally, P., Badger, J. H., Galbraith, D. W., Gage, F. H., & Lasken, R. S. (2013). RNA-sequencing from single nuclei. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(49), 19802–19807. https://doi.org/10.1073/pnas.1319700110

Groveman, B. R., Foliaki, S. T., Orru, C. D., Zanusso, G., Carroll, J. A., Race, B., & Haigh, C. L. (2019). Sporadic Creutzfeldt-Jakob disease prion infection of human cerebral organoids. *Acta Neuropathologica Communications*, *7*(1), 90. https://doi.org/10.1186/s40478-019-0742-2

Groveman, B. R., Smith, A., Williams, K., & Haigh, C. L. (2021). Cerebral organoids as a new model for prion disease. *PLoS Pathogens*, *17*(7), e1009747. https://doi.org/10.1371/journal.ppat.1009747

Grubman, A., Chew, G., Ouyang, J. F., Sun, G., Choo, X. Y., McLean, C., Simmons, R. K., Buckberry, S., Vargas-Landin, D. B., Poppe, D., Pflueger, J., Lister, R., Rackham, O. J. L., Petretto, E., & Polo, J. M. (2019). A single-cell atlas of entorhinal cortex from individuals with

Alzheimer’s disease reveals cell-type-specific gene expression regulation. *Nature Neuroscience*, *22*(12), 2087–2097. https://doi.org/10.1038/s41593-019-0539-4

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., & van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, *525*(7568), 251–255. https://doi.org/10.1038/nature14966

Grunstein, M., & Hogness, D. S. (1975). Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(10), 3961–3965. https://doi.org/10.1073/pnas.72.10.3961

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., & Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*, *18*(4), 675–685. https://doi.org/10.1016/j.devcel.2010.02.012

Gu, X.-L., Long, C.-X., Sun, L., Xie, C., Lin, X., & Cai, H. (2010). Astrocytic expression of Parkinson’s disease-related A53T alpha-synuclein causes neurodegeneration in mice. *Molecular Brain*, *3*, 12. https://doi.org/10.1186/1756-6606-3-12

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D. A., Rozenblatt-Rosen, O., Zhang, F., & Regev, A. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, *14*(10), 955–958. https://doi.org/10.1038/nmeth.4407

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J. J., Hession, C., Zhang, F., & Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*, *353*(6302), 925–928. https://doi.org/10.1126/science.aad7038

Halliday, G. M., & Stevens, C. H. (2011). Glia: initiators and progressors of pathology in Parkinson’s disease. *Movement Disorders*, *26*(1), 6–17. https://doi.org/10.1002/mds.23455

Hamaguchi, T., Noguchi-Shinohara, M., Nozaki, I., Nakamura, Y., Sato, T., Kitamoto, T., Mizusawa, H., & Yamada, M. (2009). Medical procedures and risk for sporadic Creutzfeldt-Jakob disease, Japan, 1999-2008. *Emerging Infectious Diseases*, *15*(2), 265–271. https://doi.org/10.3201/eid1502.080749

Hannaoui, S., Gougerot, A., Privat, N., Levavasseur, E., Bizat, N., Hauw, J.-J., Brandel, J.-P., & Haïk, S. (2014). Cycline efficacy on the propagation of human prions in primary cultured neurons is strain-specific. *The Journal of Infectious Diseases*, *209*(7), 1144–1148. https://doi.org/10.1093/infdis/jit623

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Harries-Jones, R., Knight, R., Will, R. G., Cousens, S., Smith, P. G., & Matthews, W. B. (1988). Creutzfeldt-Jakob disease in England and Wales, 1980-1984: a case-control study of potential risk factors. *Journal of Neurology, Neurosurgery, and Psychiatry*, *51*(9), 1113–1119. https://doi.org/10.1136/jnnp.51.9.1113

Hartmann, K., Sepulveda-Falla, D., Rose, I. V. L., Madore, C., Muth, C., Matschke, J., Butovsky, O., Liddelow, S., Glatzel, M., & Krasemann, S. (2019). Complement 3+-astrocytes are highly abundant in prion diseases, but their abolishment led to an accelerated disease course and early dysregulation of microglia. *Acta Neuropathologica Communications*, *7*(1), 83. https://doi.org/10.1186/s40478-019-0735-1

Hayer, K. E., Pizarro, A., Lahens, N. F., Hogenesch, J. B., & Grant, G. R. (2015). Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, *31*(24), 3938–3945. https://doi.org/10.1093/bioinformatics/btv488

Heimberg, G., Bhatnagar, R., El-Samad, H., & Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems*, *2*(4), 239–250. https://doi.org/10.1016/j.cels.2016.04.001

Heneka, M. T., Kummer, M. P., & Latz, E. (2014). Innate immune activation in neurodegenerative disease. *Nature Reviews. Immunology*, *14*(7), 463–477. https://doi.org/10.1038/nri3705

Heng, Y., Dubbelaar, M. L., Marie, S. K. N., Boddeke, E. W. G. M., & Eggen, B. J. L. (2021). The effects of postmortem delay on mouse and human microglia gene expression. *Glia*, *69*(4), 1053–1060. https://doi.org/10.1002/glia.23948

He, X., Jittiwat, J., Kim, J.-H., Jenner, A. M., Farooqui, A. A., Patel, S. C., & Ong, W.-Y. (2009). Apolipoprotein D modulates F2-isoprostane and 7-ketocholesterol formation and has a neuroprotective effect on organotypic hippocampal cultures after kainate-induced excitotoxic injury. *Neuroscience Letters*, *455*(3), 183–186. https://doi.org/10.1016/j.neulet.2009.03.038

Highet, B., Parker, R., Faull, R. L. M., Curtis, M. A., & Ryan, B. (2021). RNA Quality in Post-mortem Human Brain Tissue Is Affected by Alzheimer's Disease. *Frontiers in Molecular Neuroscience*, *14*, 780352. https://doi.org/10.3389/fnmol.2021.780352

Hill, Andrew F, & Collinge, J. (2003). Subclinical prion infection. *Trends in Microbiology*, *11*(12), 578–584. https://doi.org/10.1016/j.tim.2003.10.007

Hill, Andrew F, Joiner, S., Wadsworth, J. D. F., Sidle, K. C. L., Bell, J. E., Budka, H., Ironside, J. W., & Collinge, J. (2003). Molecular classification of sporadic Creutzfeldt-Jakob disease. *Brain: A Journal of Neurology*, *126*(Pt 6), 1333–1346. https://doi.org/10.1093/brain/awg125

Hill, A F, Desbruslais, M., Joiner, S., Sidle, K. C., Gowland, I., Collinge, J., Doey, L. J., & Lantos, P. (1997). The same prion strain causes vCJD and BSE. *Nature*, *389*(6650), 448–450, 526. https://doi.org/10.1038/38925

Hill, A F, Joiner, S., Linehan, J., Desbruslais, M., Lantos, P. L., & Collinge, J. (2000). Species-barrier-independent prion replication in apparently resistant species. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(18), 10248–10253. https://doi.org/10.1073/pnas.97.18.10248

Hirsch, E. C., & Hunot, S. (2009). Neuroinflammation in Parkinson's disease: a target for neuroprotection? *Lancet Neurology*, *8*(4), 382–397. https://doi.org/10.1016/S1474-4422(09)70062-6

Hol, E. M., Roelofs, R. F., Moraal, E., Sonnemans, M. A. F., Sluijs, J. A., Proper, E. A., de Graan, P. N. E., Fischer, D. F., & van Leeuwen, F. W. (2003). Neuronal expression of GFAP in patients with Alzheimer pathology and identification of novel GFAP splice forms. *Molecular Psychiatry*, *8*(9), 786–796. https://doi.org/10.1038/sj.mp.4001379

Hosokawa, M., Klegeris, A., Maguire, J., & McGeer, P. L. (2003). Expression of complement messenger RNAs and proteins by human oligodendroglial cells. *Glia*, *42*(4), 417–423. https://doi.org/10.1002/glia.10234

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325

Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, *21*(1), 218. https://doi.org/10.1186/s13059-020-02132-x

Ho, D. M., Artavanis-Tsakonas, S., & Louvi, A. (2020). The Notch pathway in CNS homeostasis and neurodegeneration. *Wiley Interdisciplinary Reviews. Developmental Biology*, *9*(1), e358. https://doi.org/10.1002/wdev.358

Hsiao, K. K., Scott, M., Foster, D., Groth, D. F., DeArmond, S. J., & Prusiner, S. B. (1990). Spontaneous neurodegeneration in transgenic mice with mutant prion protein. *Science*, *250*(4987), 1587–1590. https://doi.org/10.1126/science.1980379

Hwang, D., Lee, I. Y., Yoo, H., Gehlenborg, N., Cho, J.-H., Petritis, B., Baxter, D., Pitstick, R., Young, R., Spicer, D., Price, N. D., Hohmann, J. G., Dearmond, S. J., Carlson, G. A., & Hood, L. E. (2009). A systems approach to prion disease. *Molecular Systems Biology*, *5*, 252. https://doi.org/10.1038/msb.2009.10

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, *18 Suppl 1*, S233-40. https://doi.org/10.1093/bioinformatics/18.suppl_1.s233

Ikegami, A., Haruwaka, K., & Wake, H. (2019). Microglia: Lifelong modulator of neural circuits. *Neuropathology*, *39*(3), 173–180. https://doi.org/10.1111/neup.12560

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., & Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, *17*, 29. https://doi.org/10.1186/s13059-016-0888-1

Imran, M., & Mahmood, S. (2011). An overview of animal prion diseases. *Virology Journal*, *8*, 493. https://doi.org/10.1186/1743-422X-8-493

Ishikura, N., Clever, J. L., Bouzamondo-Bernstein, E., Samayoa, E., Prusiner, S. B., Huang, E. J., & DeArmond, S. J. (2005). Notch-1 activation and dendritic atrophy in prion disease.

*Proceedings of the National Academy of Sciences of the United States of America*, *102*(3), 886–891. https://doi.org/10.1073/pnas.0408612101

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., & Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/science.1247651

Jaunmuktane, Z., Mead, S., Ellis, M., Wadsworth, J. D. F., Nicoll, A. J., Kenny, J., Launchbury, F., Linehan, J., Richard-Loendt, A., Walker, A. S., Rudge, P., Collinge, J., & Brandner, S. (2015). Evidence for human transmission of amyloid-β pathology and cerebral amyloid angiopathy. *Nature*, *525*(7568), 247–250. https://doi.org/10.1038/nature15369

Jeffrey, M., Goodsir, C. M., Race, R. E., & Chesebro, B. (2004). Scrapie-specific neuronal lesions are independent of neuronal PrP expression. *Annals of Neurology*, *55*(6), 781–792. https://doi.org/10.1002/ana.20093

Jones, E., Hummerich, H., Viré, E., Uphill, J., Dimitriadis, A., Speedy, H., Campbell, T., Norsworthy, P., Quinn, L., Whitfield, J., Linehan, J., Jaunmuktane, Z., Brandner, S., Jat, P., Nihat, A., How Mok, T., Ahmed, P., Collins, S., Stehmann, C., … Mead, S. (2020). Identification of novel risk loci and causal insights for sporadic Creutzfeldt-Jakob disease: a genome-wide association study. *Lancet Neurology*, *19*(10), 840–848. https://doi.org/10.1016/S1474-4422(20)30273-8

Julius, C., Hutter, G., Wagner, U., Seeger, H., Kana, V., Kranich, J., Klöhn, P.-C., Weissmann, C., Miele, G., & Aguzzi, A. (2008). Transcriptional stability of cultured cells upon prion infection. *Journal of Molecular Biology*, *375*(5), 1222–1233. https://doi.org/10.1016/j.jmb.2007.11.003

Kanata, E., Llorens, F., Dafou, D., Dimitriadis, A., Thüne, K., Xanthopoulos, K., Bekas, N., Espinosa, J. C., Schmitz, M., Marín-Moreno, A., Capece, V., Shormoni, O., Andréoletti, O., Bonn, S., Torres, J. M., Ferrer, I., Zerr, I., & Sklaviadis, T. (2019). RNA editing alterations define manifestation of prion diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(39), 19727–19735. https://doi.org/10.1073/pnas.1803521116

Kawasaki, Y., Kawagoe, K., Chen, C., Teruya, K., Sakasegawa, Y., & Doh-ura, K. (2007). Orally administered amyloidophilic compound is effective in prolonging the incubation periods of animals cerebrally infected with prion diseases in a prion strain-dependent manner. *Journal of Virology*, *81*(23), 12889–12898. https://doi.org/10.1128/JVI.01563-07

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T. K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., Itzkovitz, S., Colonna, M., Schwartz, M., & Amit, I. (2017). A Unique Microglia Type Associated with Restricting Development of Alzheimer’s Disease. *Cell*, *169*(7), 1276-1290.e17. https://doi.org/10.1016/j.cell.2017.05.018

Ke, R., Mignardi, M., Hauling, T., & Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Human Mutation*, *37*(12), 1363–1367. https://doi.org/10.1002/humu.23051

Khalifé, M., Young, R., Passet, B., Halliez, S., Vilotte, M., Jaffrezic, F., Marthey, S., Béringue, V., Vaiman, D., Le Provost, F., Laude, H., & Vilotte, J.-L. (2011). Transcriptomic analysis brings new insight into the biological role of the prion protein during mouse embryogenesis. *Plos One*, *6*(8), e23253. https://doi.org/10.1371/journal.pone.0023253

Khan, S., & Kaihara, K. A. (2019). Single-Cell RNA-Sequencing of Peripheral Blood Mononuclear Cells with ddSEQ. *Methods in Molecular Biology*, *1979*, 155–176. https://doi.org/10.1007/978-1-4939-9240-9_10

Khosravani, H., Zhang, Y., Tsutsui, S., Hameed, S., Altier, C., Hamid, J., Chen, L., Villemaire, M., Ali, Z., Jirik, F. R., & Zamponi, G. W. (2008). Prion protein attenuates excitotoxicity by inhibiting NMDA receptors. *The Journal of Cell Biology*, *181*(3), 551–565. https://doi.org/10.1083/jcb.200711002

Kimberlin, R. H., & Walker, C. (1977). Characteristics of a short incubation model of scrapie in the golden hamster. *The Journal of General Virology*, *34*(2), 295–304. https://doi.org/10.1099/0022-1317-34-2-295

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, H., Park, H. J., Choi, H., Chang, Y., Park, H., Shin, J., Kim, J., Lengner, C. J., Lee, Y. K., & Kim, J. (2019). Modeling G2019S-LRRK2 Sporadic Parkinson’s Disease in 3D Midbrain Organoids. *Stem Cell Reports*, *12*(3), 518–531. https://doi.org/10.1016/j.stemcr.2019.01.020

Kim, H.-J., Tark, D.-S., Lee, Y.-H., Kim, M.-J., Lee, W.-Y., Cho, I.-S., Sohn, H.-J., & Yokoyama, T. (2012). Establishment of a cell line persistently infected with chronic wasting disease prions. *The Journal of Veterinary Medical Science*, *74*(10), 1377–1380. https://doi.org/10.1292/jvms.12-0061

Kim, H. O., Snyder, G. P., Blazey, T. M., Race, R. E., Chesebro, B., & Skinner, P. J. (2008). Prion disease induced alterations in gene expression in spleen and brain prior to clinical symptoms. *Advances and Applications in Bioinformatics and Chemistry : AABC*, *1*, 29–50.

Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Teichmann, S. A., & Marioni, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, *6*, 8687. https://doi.org/10.1038/ncomms9687

King, A. (2018). The search for better animal models of Alzheimer’s disease. *Nature*, *559*(7715), S13–S15. https://doi.org/10.1038/d41586-018-05722-9

Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews. Genetics*, *20*(5), 273–282. https://doi.org/10.1038/s41576-018-0088-9

Kitamoto, T., Tateishi, J., Sawa, H., & Doh-Ura, K. (1989). Positive transmission of Creutzfeldt-Jakob disease verified by murine kuru plaques. *Laboratory Investigation*, *60*(4), 507–512.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

Klöhn, P. C., Stoltze, L., Flechsig, E., Enari, M., & Weissmann, C. (2003). A quantitative, highly sensitive cell-based infectivity assay for mouse scrapie prions. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11666–11671. https://doi.org/10.1073/pnas.1834432100

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell*, *58*(4), 610–620. https://doi.org/10.1016/j.molcel.2015.04.005

Konnova, E. A., & Swanberg, M. (2018). Animal models of parkinson’s disease. In T. B. Stoker & J. C. Greenland (Eds.), *Parkinson’s disease: pathogenesis and clinical aspects*. Codon Publications. https://doi.org/10.15586/codonpublications.parkinsonsdisease.2018.ch5

Krance, S. H., Luke, R., Shenouda, M., Israwi, A. R., Colpitts, S. J., Darwish, L., Strauss, M., & Watts, J. C. (2020). Cellular models for discovering prion disease therapeutics: Progress and challenges. *Journal of Neurochemistry*, *153*(2), 150–172. https://doi.org/10.1111/jnc.14956

Krejciova, Z., Alibhai, J., Zhao, C., Krencik, R., Rzechorzek, N. M., Ullian, E. M., Manson, J., Ironside, J. W., Head, M. W., & Chandran, S. (2017). Human stem cell-derived astrocytes replicate human prions in a PRNP genotype-dependent manner. *The Journal of Experimental Medicine*, *214*(12), 3481–3495. https://doi.org/10.1084/jem.20161547

Kristiansen, M., Deriziotis, P., Dimcheff, D. E., Jackson, G. S., Ovaa, H., Naumann, H., Clarke, A. R., van Leeuwen, F. W. B., Menéndez-Benito, V., Dantuma, N. P., Portis, J. L., Collinge, J., & Tabrizi, S. J. (2007). Disease-associated prion protein oligomers inhibit the 26S proteasome. *Molecular Cell*, *26*(2), 175–188. https://doi.org/10.1016/j.molcel.2007.04.001

Kumar, P., Kumar, D., Jha, S. K., Jha, N. K., & Ambasta, R. K. (2016). Ion channels in neurological disorders. *Advances in Protein Chemistry and Structural Biology*, *103*, 97–136. https://doi.org/10.1016/bs.apcsb.2015.10.006

Lacar, B., Linker, S. B., Jaeger, B. N., Krishnaswami, S. R., Barron, J. J., Kelder, M. J. E., Parylak, S. L., Paquola, A. C. M., Venepally, P., Novotny, M., O’Connor, C., Fitzpatrick, C., Erwin, J. A., Hsu, J. Y., Husband, D., McConnell, M. J., Lasken, R., & Gage, F. H. (2016). Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nature Communications*, *7*, 11022. https://doi.org/10.1038/ncomms11022

Ladogana, A., Puopolo, M., Croes, E. A., Budka, H., Jarius, C., Collins, S., Klug, G. M., Sutcliffe, T., Giulivi, A., Alperovitch, A., Delasnerie-Laupretre, N., Brandel, J. P., Poser, S., Kretzschmar, H., Rietveld, I., Mitrova, E., Cuesta, J. de P., Martinez-Martin, P., Glatzel, M., … Zerr, I. (2005). Mortality from Creutzfeldt-Jakob disease and related disorders in Europe, Australia, and Canada. *Neurology*, *64*(9), 1586–1591. https://doi.org/10.1212/01.WNL.0000160117.56690.B2

Lahens, N. F., Kavakli, I. H., Zhang, R., Hayer, K., Black, M. B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R. S., Grant, G. R., & Hogenesch, J. B. (2014). IVT-seq reveals extreme bias in RNA sequencing. *Genome Biology*, *15*(6), R86. https://doi.org/10.1186/gb-2014-15-6-r86

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., Wildberg, A., Gao, D., Fung, H.-L., Chen, S., Vijayaraghavan, R., Wong, J., Chen, A., Sheng, X., Kaper, F., Shen, R., Ronaghi, M., Fan, J.-B., Wang, W., … Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, *352*(6293), 1586–1590. https://doi.org/10.1126/science.aaf1204

Lakkaraju, A. K., Sorce, S., Senatore, A., Nuvolone, M., Guo, J., Schwarz, P., Moos, R., Pelczar, P., & Aguzzi, A. (2021). Glial activation in prion diseases is strictly nonautonomous and requires neuronal PrP$^{Sc}$. *BioRxiv*. https://doi.org/10.1101/2021.01.03.425136

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Lee, H.-G., Wheeler, M. A., & Quintana, F. J. (2022). Function and therapeutic value of astrocytes in neurological diseases. *Nature Reviews. Drug Discovery*. https://doi.org/10.1038/s41573-022-00390-x

Legname, G., Nguyen, H.-O. B., Baskakov, I. V., Cohen, F. E., Dearmond, S. J., & Prusiner, S. B. (2005). Strain-specified characteristics of mouse synthetic prions. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(6), 2168–2173. https://doi.org/10.1073/pnas.0409079102

Liddelow, S. A., & Barres, B. A. (2017). Reactive astrocytes: production, function, and therapeutic potential. *Immunity*, *46*(6), 957–967. https://doi.org/10.1016/j.immuni.2017.06.006

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, K., Li, J., Zheng, J., & Qin, S. (2019). Reactive astrocytes in neurodegenerative diseases. *Aging and Disease*, *10*(3), 664–675. https://doi.org/10.14336/AD.2018.0720

Li, Y., Wang, R., Qiao, N., Peng, G., Zhang, K., Tang, K., Han, J.-D. J., & Jing, N. (2017). Transcriptome analysis reveals determinant stages controlling human embryonic stem cell commitment to neuronal cells. *The Journal of Biological Chemistry*, *292*(48), 19590–19604. https://doi.org/10.1074/jbc.M117.796383

Li, Z., Del-Aguila, J. L., Dube, U., Budde, J., Martinez, R., Black, K., Xiao, Q., Cairns, N. J., Dominantly Inherited Alzheimer Network (DIAN), Dougherty, J. D., Lee, J.-M., Morris, J. C., Bateman, R. J., Karch, C. M., Cruchaga, C., & Harari, O. (2018). Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure. *Genome Medicine*, *10*(1), 43. https://doi.org/10.1186/s13073-018-0551-4

Liddelow, S. A., Guttenplan, K. A., Clarke, L. E., Bennett, F. C., Bohlen, C. J., Schirmer, L., Bennett, M. L., Münch, A. E., Chung, W.-S., Peterson, T. C., Wilton, D. K., Frouin, A., Napier, B. A., Panicker, N., Kumar, M., Buckwalter, M. S., Rowitch, D. H., Dawson, V. L., Dawson, T. M., … Barres, B. A. (2017). Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*, *541*(7638), 481–487. https://doi.org/10.1038/nature21029

Loerch, P. M., Lu, T., Dakin, K. A., Vann, J. M., Isaacs, A., Geula, C., Wang, J., Pan, Y., Gabuzda, D. H., Li, C., Prolla, T. A., & Yankner, B. A. (2008). Evolution of the aging brain transcriptome and synaptic regulation. *Plos One*, *3*(10), e3329. https://doi.org/10.1371/journal.pone.0003329

López-Pérez, Ó., Badiola, J. J., Bolea, R., Ferrer, I., Llorens, F., & Martín-Burriel, I. (2020). An update on autophagy in prion diseases. *Frontiers in Bioengineering and Biotechnology*, *8*, 975. https://doi.org/10.3389/fbioe.2020.00975

Love, M. I., Anders, S., Kim, V., & Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. [version 1; peer review: 2 approved]. *F1000Research*, *4*, 1070. https://doi.org/10.12688/f1000research.7035.1

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6), e8746. https://doi.org/10.15252/msb.20188746

Lun, A. T. L., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, *17*, 75. https://doi.org/10.1186/s13059-016-0947-7

Lun, A. T. L., McCarthy, D. J., & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. [version 2; peer review: 3 approved, 2 approved with reservations]. *F1000Research*, *5*, 2122. https://doi.org/10.12688/f1000research.9501.2

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Mahadik, S. P., Khan, M. M., Evans, D. R., & Parikh, V. V. (2002). Elevated plasma level of apolipoprotein D in schizophrenia and its treatment and outcome. *Schizophrenia Research*, *58*(1), 55–62. https://doi.org/10.1016/s0920-9964(01)00378-4

Mahal, S. P., Baker, C. A., Demczyk, C. A., Smith, E. W., Julius, C., & Weissmann, C. (2007). Prion strain discrimination in cell culture: the cell panel assay. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(52), 20908–20913. https://doi.org/10.1073/pnas.0710054104

Mahal, S. P., Demczyk, C. A., Smith, E. W., Klohn, P.-C., & Weissmann, C. (2008). Assaying prions in cell culture: the standard scrapie cell assay (SSCA) and the scrapie cell assay in end point format (SCEPA). *Methods in Molecular Biology*, *459*, 49–68. https://doi.org/10.1007/978-1-59745-234-2_4

Mahdieh, N., Mikaeeli, S., Tavasoli, A. R., Rezaei, Z., Maleki, M., & Rabbani, B. (2018). Genotype, phenotype and in silico pathogenicity analysis of HEXB mutations: Panel based sequencing for differential diagnosis of gangliosidosis. *Clinical Neurology and Neurosurgery*, *167*, 43–53. https://doi.org/10.1016/j.clineuro.2018.02.011

Mahfoud, R., Garmy, N., Maresca, M., Yahi, N., Puigserver, A., & Fantini, J. (2002). Identification of a common sphingolipid-binding domain in Alzheimer, prion, and HIV-1 proteins. *The Journal of Biological Chemistry*, *277*(13), 11292–11296. https://doi.org/10.1074/jbc.M111679200

Mahmoud, S., Gharagozloo, M., Simard, C., & Gris, D. (2019). Astrocytes Maintain Glutamate Homeostasis in the CNS by Controlling the Balance between Glutamate Uptake and Release. *Cells*, *8*(2). https://doi.org/10.3390/cells8020184

Maier, T., Strater, N., Schuette, C. G., Klingenstein, R., Sandhoff, K., & Saenger, W. (2003). The X-ray crystal structure of human beta-hexosaminidase B provides new insights into Sandhoff disease. *Journal of Molecular Biology*, *328*(3), 669–681. https://doi.org/10.1016/s0022-2836(03)00311-5

Majer, A., Medina, S. J., Niu, Y., Abrenica, B., Manguiat, K. J., Frost, K. L., Philipson, C. S., Sorensen, D. L., & Booth, S. A. (2012). Early mechanisms of pathobiology are revealed by transcriptional temporal dynamics in hippocampal CA1 neurons of prion infected mice. *PLoS Pathogens*, *8*(11), e1003002. https://doi.org/10.1371/journal.ppat.1003002

Mallucci, G. R. (2009). Prion neurodegeneration: starts and stops at the synapse. *Prion*, *3*(4), 195–201.

Manuelidis, L., Tesin, D. M., Sklaviadis, T., & Manuelidis, E. E. (1987). Astrocyte gene expression in Creutzfeldt-Jakob disease. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(16), 5937–5941. https://doi.org/10.1073/pnas.84.16.5937

Marbiah, M. M., Harvey, A., West, B. T., Louzolo, A., Banerjee, P., Alden, J., Grigoriadis, A., Hummerich, H., Kan, H.-M., Cai, Y., Bloom, G. S., Jat, P., Collinge, J., & Klöhn, P.-C. (2014). Identification of a gene regulatory network associated with prion replication. *The EMBO Journal*, *33*(14), 1527–1547. https://doi.org/10.15252/embj.201387150

Marino, S., Vooijs, M., van Der Gulden, H., Jonkers, J., & Berns, A. (2000). Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. *Genes & Development*, *14*(8), 994–1004.

Martins, V. R., Beraldo, F. H., Hajj, G. N., Lopes, M. H., Lee, K. S., Prado, M. A., & Linden, R. (2010). Prion protein: orchestrating neurotrophic activities. *Current Issues in Molecular Biology*, *12*(2), 63–86.

Masters, C. L., Harris, J. O., Gajdusek, D. C., Gibbs, C. J., Bernoulli, C., & Asher, D. M. (1979). Creutzfeldt-Jakob disease: patterns of worldwide occurrence and the significance of familial and sporadic clustering. *Annals of Neurology*, *5*(2), 177–188. https://doi.org/10.1002/ana.410050212

Masuda, T., Sankowski, R., Staszewski, O., Böttcher, C., Amann, L., Sagar, Scheiwe, C., Nessler, S., Kunz, P., van Loo, G., Coenen, V. A., Reinacher, P. C., Michel, A., Sure, U., Gold, R., Grün, D., Priller, J., Stadelmann, C., & Prinz, M. (2019). Spatial and temporal heterogeneity of mouse and human microglia at single-cell resolution. *Nature*, *566*(7744), 388–392. https://doi.org/10.1038/s41586-019-0924-x

Mathys, H., Adaikkan, C., Gao, F., Young, J. Z., Manet, E., Hemberg, M., De Jager, P. L., Ransohoff, R. M., Regev, A., & Tsai, L.-H. (2017). Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Reports*, *21*(2), 366–380. https://doi.org/10.1016/j.celrep.2017.09.039

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett, D. A., Kellis, M., & Tsai, L.-H. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, *570*(7761), 332–337. https://doi.org/10.1038/s41586-019-1195-2

Mays, C. E., & Soto, C. (2016). The stress of prion disease. *Brain Research*, *1648*(Pt B), 553–560. https://doi.org/10.1016/j.brainres.2016.04.009

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., & Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, *33*(8), 1179–1186. https://doi.org/10.1093/bioinformatics/btw777

McKinnon, C., Goold, R., Andre, R., Devoy, A., Ortega, Z., Moonga, J., Linehan, J. M., Brandner, S., Lucas, J. J., Collinge, J., & Tabrizi, S. J. (2016). Prion-mediated neurodegeneration is associated with early impairment of the ubiquitin-proteasome system. *Acta Neuropathologica*, *131*(3), 411–425. https://doi.org/10.1007/s00401-015-1508-y

Mead, S, Beck, J., Dickinson, A., Fisher, E. M., & Collinge, J. (2000). Examination of the human prion protein-like gene doppel for genetic susceptibility to sporadic and variant Creutzfeldt-Jakob disease. *Neuroscience Letters*, *290*(2), 117–120. https://doi.org/10.1016/s0304-3940(00)01319-7

Mead, Simon, Uphill, J., Beck, J., Poulter, M., Campbell, T., Lowe, J., Adamson, G., Hummerich, H., Klopp, N., Rückert, I.-M., Wichmann, H.-E., Azazi, D., Plagnol, V., Pako, W. H., Whitfield, J., Alpers, M. P., Whittaker, J., Balding, D. J., Zerr, I., … Collinge, J. (2012). Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to PRNP. *Human Molecular Genetics*, *21*(8), 1897–1906. https://doi.org/10.1093/hmg/ddr607

Mendiola, A. S., Ryu, J. K., Bardehle, S., Meyer-Franke, A., Ang, K. K.-H., Wilson, C., Baeten, K. M., Hanspers, K., Merlini, M., Thomas, S., Petersen, M. A., Williams, A., Thomas, R., Rafalski, V. A., Meza-Acevedo, R., Tognatta, R., Yan, Z., Pfaff, S. J., Machado, M. R., … Akassoglou, K. (2020). Transcriptional profiling and therapeutic targeting of oxidative stress in neuroinflammation. *Nature Immunology*, *21*(5), 513–524. https://doi.org/10.1038/s41590-020-0654-0

Method of the Year 2020: spatially resolved transcriptomics. (2021). *Nature Methods*, *18*(1), 1. https://doi.org/10.1038/s41592-020-01042-x

Meyer-Luehmann, M., Spires-Jones, T. L., Prada, C., Garcia-Alloza, M., de Calignon, A., Rozkalne, A., Koenigsknecht-Talboo, J., Holtzman, D. M., Bacskai, B. J., & Hyman, B. T. (2008). Rapid appearance and local toxicity of amyloid-beta plaques in a mouse model of Alzheimer’s disease. *Nature*, *451*(7179), 720–724. https://doi.org/10.1038/nature06616

Miller, S. A., Policastro, R. A., Savant, S. S., Sriramkumar, S., Ding, N., Lu, X., Mohammad, H. P., Cao, S., Kalin, J. H., Cole, P. A., Zentner, G. E., & O’Hagan, H. M. (2020). Lysine-Specific Demethylase 1 Mediates AKT Activity and Promotes Epithelial-to-Mesenchymal Transition in PIK3CA-Mutant Colorectal Cancer. *Molecular Cancer Research*, *18*(2), 264–277. https://doi.org/10.1158/1541-7786.MCR-19-0748

Moody, L. R., Herbst, A. J., Yoo, H. S., Vanderloo, J. P., & Aiken, J. M. (2009). Comparative prion disease gene expression profiling using the prion disease mimetic, cuprizone. *Prion*, *3*(2), 99–109. https://doi.org/10.4161/pri.3.2.9059

Moore, R. A., Sturdevant, D. E., Chesebro, B., & Priola, S. A. (2014). Proteomics analysis of amyloid and nonamyloid prion disease phenotypes reveals both common and divergent mechanisms of neuropathogenesis. *Journal of Proteome Research*, *13*(11), 4620–4634. https://doi.org/10.1021/pr500329w

Moore, R. C., Lee, I. Y., Silverman, G. L., Harrison, P. M., Strome, R., Heinrich, C., Karunaratne, A., Pasternak, S. H., Chishti, M. A., Liang, Y., Mastrangelo, P., Wang, K., Smit, A. F., Katamine, S., Carlson, G. A., Cohen, F. E., Prusiner, S. B., Melton, D. W., Tremblay, P., … Westaway, D. (1999). Ataxia in prion protein (PrP)-deficient mice is associated with upregulation of the novel PrP-like protein doppel. *Journal of Molecular Biology*, *292*(4), 797–817. https://doi.org/10.1006/jmbi.1999.3108

Moreno, J. A., Radford, H., Peretti, D., Steinert, J. R., Verity, N., Martin, M. G., Halliday, M., Morgan, J., Dinsdale, D., Ortori, C. A., Barrett, D. A., Tsaytler, P., Bertolotti, A., Willis, A. E., Bushell, M., & Mallucci, G. R. (2012). Sustained translational repression by eIF2α-P mediates prion neurodegeneration. *Nature*, *485*(7399), 507–511. https://doi.org/10.1038/nature11058

Mosher, K. I., & Wyss-Coray, T. (2014). Microglial dysfunction in brain aging and Alzheimer’s disease. *Biochemical Pharmacology*, *88*(4), 594–604. https://doi.org/10.1016/j.bcp.2014.01.008

Mou, T., Deng, W., Gu, F., Pawitan, Y., & Vu, T. N. (2019). Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. *Frontiers in Genetics*, *10*, 1331. https://doi.org/10.3389/fgene.2019.01331

Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., Sinnamon, J. R., Shendure, J., Trapnell, C., O'Roak, B. J., Xia, Z., Steemers, F. J., & Adey, A. C. (2018). Highly scalable generation of DNA methylation profiles in single cells. *Nature Biotechnology*, *36*(5), 428–431. https://doi.org/10.1038/nbt.4112

Muñoz-Gutiérrez, J. F., Pierlé, S. A., Schneider, D. A., Baszler, T. V., & Stanton, J. B. (2016). Transcriptomic determinants of scrapie prion propagation in cultured ovine microglia. *Plos One*, *11*(1), e0147727. https://doi.org/10.1371/journal.pone.0147727

Myerowitz, R., Lawson, D., Mizukami, H., Mi, Y., Tifft, C. J., & Proia, R. L. (2002). Molecular pathophysiology in Tay-Sachs and Sandhoff diseases as revealed by gene expression profiling. *Human Molecular Genetics*, *11*(11), 1343–1350. https://doi.org/10.1093/hmg/11.11.1343

Nagy, C., Maitra, M., Tanti, A., Suderman, M., Théroux, J.-F., Davoli, M. A., Perlman, K., Yerko, V., Wang, Y. C., Tripathy, S. J., Pavlidis, P., Mechawar, N., Ragoussis, J., & Turecki, G. (2020). Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nature Neuroscience*, *23*(6), 771–781. https://doi.org/10.1038/s41593-020-0621-y

Natarajan, K. N., Miao, Z., Jiang, M., Huang, X., Zhou, H., Xie, J., Wang, C., Qin, S., Zhao, Z., Wu, L., Yang, N., Li, B., Hou, Y., Liu, S., & Teichmann, S. A. (2019). Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biology*, *20*(1), 70. https://doi.org/10.1186/s13059-019-1676-5

Neale, M. H., Mountjoy, S. J., Edwards, J. C., Vilette, D., Laude, H., Windl, O., & Saunders, G. C. (2010). Infection of cell lines with experimental and natural ovine scrapie agents. *Journal of Virology*, *84*(5), 2444–2452. https://doi.org/10.1128/JVI.01855-09

Neves, A. C. (2019). Prion paradigm: understanding neurodegenerativedisorders. *Biomedical Journal of Scientific & Technical Research*, *13*(5). https://doi.org/10.26717/BJSTR.2019.13.002464

Nishida, N., Harris, D. A., Vilette, D., Laude, H., Frobert, Y., Grassi, J., Casanova, D., Milhavet, O., & Lehmann, S. (2000). Successful transmission of three mouse-adapted scrapie strains to murine neuroblastoma cell lines overexpressing wild-type mouse prion protein. *Journal of Virology*, *74*(1), 320–325. https://doi.org/10.1128/jvi.74.1.320-325.2000

Nishida, N., Tremblay, P., Sugimoto, T., Shigematsu, K., Shirabe, S., Petromilli, C., Erpel, S. P., Nakaoke, R., Atarashi, R., Houtani, T., Torchia, M., Sakaguchi, S., DeArmond, S. J., Prusiner, S. B., & Katamine, S. (1999). A mouse prion protein transgene rescues mice deficient for the prion protein gene from purkinje cell degeneration and demyelination. *Laboratory Investigation*, *79*(6), 689–697.

Norsworthy, P. J., Thompson, A. G. B., Mok, T. H., Guntoro, F., Dabin, L. C., Nihat, A., Paterson, R. W., Schott, J. M., Collinge, J., Mead, S., & Viré, E. A. (2020). A blood miRNA signature associates with sporadic Creutzfeldt-Jakob disease diagnosis. *Nature Communications*, *11*(1), 3960. https://doi.org/10.1038/s41467-020-17655-x

Nuvolone, M., Hermann, M., Sorce, S., Russo, G., Tiberi, C., Schwarz, P., Minikel, E., Sanoudou, D., Pelczar, P., & Aguzzi, A. (2016). Strictly co-isogenic C57BL/6J-Prnp-/- mice: A rigorous resource for prion science. *The Journal of Experimental Medicine*, *213*(3), 313–327. https://doi.org/10.1084/jem.20151610

O'Flanagan, C. H., Campbell, K. R., Zhang, A. W., Kabeer, F., Lim, J. L. P., Biele, J., Eirew, P., Lai, D., McPherson, A., Kong, E., Bates, C., Borkowski, K., Wiens, M., Hewitson, B., Hopkins, J., Pham, J., Ceglia, N., Moore, R., Mungall, A. J., … Aparicio, S. (2019). Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biology*, *20*(1), 210. https://doi.org/10.1186/s13059-019-1830-0

O'Shea, M., Maytham, E. G., Linehan, J. M., Brandner, S., Collinge, J., & Lloyd, S. E. (2008). Investigation of mcp1 as a quantitative trait gene for prion disease incubation time in mouse. *Genetics*, *180*(1), 559–566. https://doi.org/10.1534/genetics.108.090894

Ordoñez, C., Navarro, A., Perez, C., Astudillo, A., Martínez, E., & Tolivia, J. (2006). Apolipoprotein D expression in substantia nigra of Parkinson disease. *Histology and Histopathology*, *21*(4), 361–366. https://doi.org/10.14670/HH-21.361

Osaki, T., Uzel, S. G. M., & Kamm, R. D. (2018). Microphysiological 3D model of amyotrophic lateral sclerosis (ALS) from human iPS-derived muscle cells and optogenetic motor neurons. *Science Advances*, *4*(10), eaat5847. https://doi.org/10.1126/sciadv.aat5847

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, *11*(12), 220. https://doi.org/10.1186/gb-2010-11-12-220

Palmer, M. S., Dryden, A. J., Hughes, J. T., & Collinge, J. (1991). Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease. *Nature*, *352*(6333), 340–342. https://doi.org/10.1038/352340a0

Paolicelli, R. C., Bolasco, G., Pagani, F., Maggi, L., Scianni, M., Panzanelli, P., Giustetto, M., Ferreira, T. A., Guiducci, E., Dumas, L., Ragozzino, D., & Gross, C. T. (2011). Synaptic pruning by microglia is necessary for normal brain development. *Science*, *333*(6048), 1456–1458. https://doi.org/10.1126/science.1202529

Parchi, P., Castellani, R., Capellari, S., Ghetti, B., Young, K., Chen, S. G., Farlow, M., Dickson, D. W., Sima, A. A., Trojanowski, J. Q., Petersen, R. B., & Gambetti, P. (1996). Molecular basis of phenotypic variability in sporadic Creutzfeldt-Jakob disease. *Annals of Neurology*, *39*(6), 767–778. https://doi.org/10.1002/ana.410390613

Parchi, P., Giese, A., Capellari, S., Brown, P., Schulz-Schaeffer, W., Windl, O., Zerr, I., Budka, H., Kopp, N., Piccardo, P., Poser, S., Rojiani, A., Streichemberger, N., Julien, J., Vital, C., Ghetti, B., Gambetti, P., & Kretzschmar, H. (1999). Classification of sporadic Creutzfeldt-Jakob disease based on molecular and phenotypic analysis of 300 subjects. *Annals of Neurology*, *46*(2), 224–233. https://doi.org/10.1002/1531-8249(199908)46:2<224::AID-ANA12>3.0.CO;2-W

Parkhurst, C. N., Yang, G., Ninan, I., Savas, J. N., Yates, J. R., Lafaille, J. J., Hempstead, B. L., Littman, D. R., & Gan, W.-B. (2013). Microglia promote learning-dependent synapse formation through brain-derived neurotrophic factor. *Cell*, *155*(7), 1596–1609. https://doi.org/10.1016/j.cell.2013.11.030

Patel, M. V. (2018). iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. *Bioinformatics*, *34*(24), 4305–4306. https://doi.org/10.1093/bioinformatics/bty517

Pehar, M., Harlan, B. A., Killoy, K. M., & Vargas, M. R. (2017). Role and therapeutic potential of astrocytes in amyotrophic lateral sclerosis. *Current Pharmaceutical Design*, *23*(33), 5010–5021. https://doi.org/10.2174/1381612823666170622095802

Perez-Nievas, B. G., & Serrano-Pozo, A. (2018). Deciphering the astrocyte reaction in alzheimer’s disease. *Frontiers in Aging Neuroscience*, *10*, 114. https://doi.org/10.3389/fnagi.2018.00114

Perkins, J. R., Antunes-Martins, A., Calvo, M., Grist, J., Rust, W., Schmid, R., Hildebrandt, T., Kohl, M., Orengo, C., McMahon, S. B., & Bennett, D. L. H. (2014). A comparison of RNA-seq and exon arrays for whole genome transcription profiling of the L5 spinal nerve transection model of neuropathic pain in the rat. *Molecular Pain*, *10*, 7. https://doi.org/10.1186/1744-8069-10-7

Petit, C. S. V., Besnier, L., Morel, E., Rousset, M., & Thenet, S. (2013). Roles of the cellular prion protein in the regulation of cell-cell junctions and barrier function. *Tissue Barriers*, *1*(2), e24377. https://doi.org/10.4161/tisb.24377

Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Płoski, R. (2016). Next Generation Sequencing—General Information about the Technology, Possibilities, and Limitations. In *Clinical Applications for Next-Generation Sequencing* (pp. 1–18). Elsevier. https://doi.org/10.1016/B978-0-12-801739-5.00001-5

Ponath, G., Park, C., & Pitt, D. (2018). The role of astrocytes in multiple sclerosis. *Frontiers in Immunology*, *9*, 217. https://doi.org/10.3389/fimmu.2018.00217

Prinz, M., Montrasio, F., Furukawa, H., van der Haar, M. E., Schwarz, P., Rülicke, T., Giger, O. T., Häusler, K.-G., Perez, D., Glatzel, M., & Aguzzi, A. (2004). Intrinsic resistance of oligodendrocytes to prion infection. *The Journal of Neuroscience*, *24*(26), 5974–5981. https://doi.org/10.1523/JNEUROSCI.0122-04.2004

Prusiner, S. B., Cochran, S. P., Groth, D. F., Downey, D. E., Bowman, K. A., & Martinez, H. M. (1982). Measurement of the scrapie agent using an incubation time interval assay. *Annals of Neurology*, *11*(4), 353–358. https://doi.org/10.1002/ana.410110406

Prusiner, S. B., & Hsiao, K. K. (1994). Human prion diseases. *Annals of Neurology*, *35*(4), 385–395. https://doi.org/10.1002/ana.410350404

Prusiner, S. B. (1982). Novel proteinaceous infectious particles cause scrapie. *Science*, *216*(4542), 136–144. https://doi.org/10.1126/science.6801762

Prusiner, S. B. (1998). Prions. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(23), 13363–13383. https://doi.org/10.1073/pnas.95.23.13363

Purro, S. A., Farrow, M. A., Linehan, J., Nazari, T., Thomas, D. X., Chen, Z., Mengel, D., Saito, T., Saido, T., Rudge, P., Brandner, S., Walsh, D. M., & Collinge, J. (2018). Transmission of amyloid-β protein pathology from cadaveric pituitary growth hormone. *Nature*, *564*(7736), 415–419. https://doi.org/10.1038/s41586-018-0790-y

Race, B., Williams, K., Orrú, C. D., Hughson, A. G., Lubke, L., & Chesebro, B. (2018). Lack of transmission of chronic wasting disease to cynomolgus macaques. *Journal of Virology*, *92*(14). https://doi.org/10.1128/JVI.00550-18

Race, R. E., Caughey, B., Graham, K., Ernst, D., & Chesebro, B. (1988). Analyses of frequency of infection, specific infectivity, and prion protein biosynthesis in scrapie-infected neuroblastoma cell clones. *Journal of Virology*, *62*(8), 2845–2849. https://doi.org/10.1128/JVI.62.8.2845-2849.1988

Race, R. E., Fadness, L. H., & Chesebro, B. (1987). Characterization of scrapie infection in mouse neuroblastoma cells. *The Journal of General Virology*, *68 ( Pt 5)*, 1391–1399. https://doi.org/10.1099/0022-1317-68-5-1391

Raeber, A. J., Race, R. E., Brandner, S., Priola, S. A., Sailer, A., Bessen, R. A., Mucke, L., Manson, J., Aguzzi, A., Oldstone, M. B., Weissmann, C., & Chesebro, B. (1997). Astrocyte-specific expression of hamster prion protein (PrP) renders PrP knockout mice susceptible to hamster scrapie. *The EMBO Journal*, *16*(20), 6057–6065. https://doi.org/10.1093/emboj/16.20.6057

Raj, A., & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, *135*(2), 216–226. https://doi.org/10.1016/j.cell.2008.09.050

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., Noble, W. S., Duan, Z., & Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nature Methods*, *14*(3), 263–266. https://doi.org/10.1038/nmeth.4155

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., & Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, *30*(8), 777–782. https://doi.org/10.1038/nbt.2282

Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A. G., & Liguori, M. J. (2018). Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Frontiers in Genetics*, *9*, 636. https://doi.org/10.3389/fgene.2018.00636

Raymond, G. J., Olsen, E. A., Lee, K. S., Raymond, L. D., Bryant, P. K., Baron, G. S., Caughey, W. S., Kocisko, D. A., McHolland, L. E., Favara, C., Langeveld, J. P. M., van Zijderveld, F. G., Mayer, R. T., Miller, M. W., Williams, E. S., & Caughey, B. (2006). Inhibition of protease-resistant prion protein formation in a transformed deer cell line infected with chronic wasting disease. *Journal of Virology*, *80*(2), 596–604. https://doi.org/10.1128/JVI.80.2.596-604.2006

Richt, J. A., Kasinathan, P., Hamir, A. N., Castilla, J., Sathiyaseelan, T., Vargas, F., Sathiyaseelan, J., Wu, H., Matsushita, H., Koster, J., Kato, S., Ishida, I., Soto, C., Robl, J. M., & Kuroiwa, Y. (2007). Production of cattle lacking prion protein. *Nature Biotechnology*, *25*(1), 132–138. https://doi.org/10.1038/nbt1271

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., & Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176–182. https://doi.org/10.1126/science.aam8999

Rossi, D., & Volterra, A. (2009). Astrocytic dysfunction: insights on the role in neurodegeneration. *Brain Research Bulletin*, *80*(4–5), 224–232. https://doi.org/10.1016/j.brainresbull.2009.07.012

Rostom, R., Svensson, V., Teichmann, S. A., & Kar, G. (2017). Computational approaches for interpreting scRNA-seq data. *FEBS Letters*, *591*(15), 2213–2225. https://doi.org/10.1002/1873-3468.12684https://sciwheel.com/work/bibliography/8745978

Rubenstein, R., Deng, H., Race, R. E., Ju, W., Scalici, C. L., Papini, M. C., Kascsak, R. J., & Carp, R. I. (1992). Demonstration of scrapie strain diversity in infected PC12 cells. *The Journal of General Virology*, *73 ( Pt 11)*, 3027–3031. https://doi.org/10.1099/0022-1317-73-11-3027

Rue-Albrecht, K., Marini, F., Soneson, C., & Lun, A. T. L. (2018). iSEE: Interactive SummarizedExperiment Explorer. [version 1; peer review: 3 approved]. *F1000Research*, *7*, 741. https://doi.org/10.12688/f1000research.14966.1

Rus, H., & Niculescu, F. (2001). The complement system in central nervous system diseases. *Immunologic Research*, *24*(1), 79–86. https://doi.org/10.1385/IR:24:1:79

Ruzicka, W. B., Mohammadi, S., Davila-Velderrain, J., Subburaju, S., Tso, D. R., Hourihan, M., & Kellis, M. (2020). Single-cell dissection of schizophrenia reveals neurodevelopmental-synaptic axis and transcriptional resilience. *MedRxiv*. https://doi.org/10.1101/2020.11.06.20225342

Saba, R., Goodman, C. D., Huzarewich, R. L. C. H., Robertson, C., & Booth, S. A. (2008). A miRNA signature of prion induced neurodegeneration. *Plos One*, *3*(11), e3652. https://doi.org/10.1371/journal.pone.0003652

Saborio, G. P., Permanne, B., & Soto, C. (2001). Sensitive detection of pathological prion protein by cyclic amplification of protein misfolding. *Nature*, *411*(6839), 810–813. https://doi.org/10.1038/35081095

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Safar, J. G. (2016). *Prion paradigm of human neurodegenerative diseases caused by protein misfolding* (Vol. 1). Oxford University Press. https://doi.org/10.1093/med/9780190233563.003.0005

Sailer, A., Büeler, H., Fischer, M., Aguzzi, A., & Weissmann, C. (1994). No propagation of prions in mice devoid of PrP. *Cell*, *77*(7), 967–968. https://doi.org/10.1016/0092-8674(94)90436-7

Sakaguchi, S., Katamine, S., Nishida, N., Moriuchi, R., Shigematsu, K., Sugimoto, T., Nakatani, A., Kataoka, Y., Houtani, T., Shirabe, S., Okada, H., Hasegawa, S., Miyamoto, T., & Noda, T. (1996). Loss of cerebellar Purkinje cells in aged mice homozygous for a disrupted PrP gene. *Nature*, *380*(6574), 528–531. https://doi.org/10.1038/380528a0

Salminen, A., Ojala, J., Suuronen, T., Kaarniranta, K., & Kauppinen, A. (2008). Amyloid-beta oligomers set fire to inflammasomes and induce Alzheimer’s pathology. *Journal of Cellular and Molecular Medicine*, *12*(6A), 2255–2262. https://doi.org/10.1111/j.1582-4934.2008.00496.x

Sandberg, M. K., Al-Doujaily, H., Sharps, B., Clarke, A. R., & Collinge, J. (2011). Prion propagation and toxicity in vivo occur in two distinct mechanistic phases. *Nature*, *470*(7335), 540–542. https://doi.org/10.1038/nature09768

Sandberg, M. K., Al-Doujaily, H., Sharps, B., De Oliveira, M. W., Schmidt, C., Richard-Londt, A., Lyall, S., Linehan, J. M., Brandner, S., Wadsworth, J. D. F., Clarke, A. R., & Collinge, J. (2014). Prion neuropathology follows the accumulation of alternate prion protein isoforms after infective titre has peaked. *Nature Communications*, *5*, 4347. https://doi.org/10.1038/ncomms5347

Santello, M., Toni, N., & Volterra, A. (2019). Astrocyte function from information processing to cognition and cognitive impairment. *Nature Neuroscience*, *22*(2), 154–166. https://doi.org/10.1038/s41593-018-0325-8

Sasaki, A., Hirato, J., & Nakazato, Y. (1993). Immunohistochemical study of microglia in the Creutzfeldt-Jakob diseased brain. *Acta Neuropathologica*, *86*(4), 337–344. https://doi.org/10.1007/BF00369445

Schafflick, D., Xu, C. A., Hartlehnert, M., Cole, M., Schulte-Mecklenbeck, A., Lautwein, T., Wolbert, J., Heming, M., Meuth, S. G., Kuhlmann, T., Gross, C. C., Wiendl, H., Yosef, N., & Meyer Zu Horste, G. (2020). Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nature Communications*, *11*(1), 247. https://doi.org/10.1038/s41467-019-14118-w

Scheckel, C., Imeri, M., Schwarz, P., & Aguzzi, A. (2020). Ribosomal profiling during prion disease uncovers progressive translational derangement in glia but not in neurons. *ELife*, *9*. https://doi.org/10.7554/eLife.62911

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Schirmer, L., Velmeshev, D., Holmqvist, S., Kaufmann, M., Werneburg, S., Jung, D., Vistnes, S., Stockley, J. H., Young, A., Steindel, M., Tung, B., Goyal, N., Bhaduri, A., Mayer, S., Engler, J. B., Bayraktar, O. A., Franklin, R. J. M., Haeussler, M., Reynolds, R., … Rowitch, D. H. (2019). Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature*, *573*(7772), 75–82. https://doi.org/10.1038/s41586-019-1404-z

Schmitt-Ulms, G., Legname, G., Baldwin, M. A., Ball, H. L., Bradon, N., Bosque, P. J., Crossin, K. L., Edelman, G. M., DeArmond, S. J., Cohen, F. E., & Prusiner, S. B. (2001). Binding of neural cell adhesion molecules (N-CAMs) to the cellular prion protein. *Journal of Molecular Biology*, *314*(5), 1209–1225. https://doi.org/10.1006/jmbi.2000.5183

Scholtens, D., & von Heydebreck, A. (2005). Analysis of Differential Gene Expression Studies. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 229–248). Springer New York. https://doi.org/10.1007/0-387-29362-0_14

Schwab, C., & McGeer, P. L. (2002). Complement activated C4d immunoreactive oligodendrocytes delineate small cortical plaques in multiple sclerosis. *Experimental Neurology*, *174*(1), 81–88. https://doi.org/10.1006/exnr.2001.7851

Scott, M., Foster, D., Mirenda, C., Serban, D., Coufal, F., Wälchli, M., Torchia, M., Groth, D., Carlson, G., DeArmond, S. J., Westaway, D., & Prusiner, S. B. (1989). Transgenic mice expressing hamster prion protein produce species-specific scrapie infectivity and amyloid plaques. *Cell*, *59*(5), 847–857. https://doi.org/10.1016/0092-8674(89)90608-9

Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., & Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *IScience*, *23*(3), 100882. https://doi.org/10.1016/j.isci.2020.100882

Skinner, P. J., Abbassi, H., Chesebro, B., Race, R. E., Reilly, C., & Haase, A. T. (2006). Gene expression alterations in brains of mice infected with three strains of scrapie. *BMC Genomics*, *7*, 114. https://doi.org/10.1186/1471-2164-7-114

Smith, H. L., Freeman, O. J., Butcher, A. J., Holmqvist, S., Humoud, I., Schätzl, T., Hughes, D. T., Verity, N. C., Swinden, D. P., Hayes, J., de Weerd, L., Rowitch, D. H., Franklin, R. J. M., & Mallucci, G. R. (2020). Astrocyte Unfolded Protein Response Induces a Specific Reactivity State that Causes Non-Cell-Autonomous Neuronal Degeneration. *Neuron*, *105*(5), 855-866.e5. https://doi.org/10.1016/j.neuron.2019.12.014

Solassol, J., Crozet, C., & Lehmann, S. (2003). Prion propagation in cultured cells. *British Medical Bulletin*, *66*, 87–97. https://doi.org/10.1093/bmb/66.1.87

Solforosi, L., Milani, M., Mancini, N., Clementi, M., & Burioni, R. (2013). A closer look at prion strains: characterization and important implications. *Prion*, *7*(2), 99–108. https://doi.org/10.4161/pri.23490

Solomon, I. H., Huettner, J. E., & Harris, D. A. (2010). Neurotoxic mutants of the prion protein induce spontaneous ionic currents in cultured cells. *The Journal of Biological Chemistry*, *285*(34), 26719–26726. https://doi.org/10.1074/jbc.M110.134619

Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, *15*(4), 255–261. https://doi.org/10.1038/nmeth.4612

Sorce, S., Nuvolone, M., Russo, G., Chincisan, A., Heinzer, D., Avar, M., Pfammatter, M., Schwarz, P., Delic, M., Müller, M., Hornemann, S., Sanoudou, D., Scheckel, C., & Aguzzi, A. (2020). Genome-wide transcriptomics identifies an early preclinical signature of prion infection. *PLoS Pathogens*, *16*(6), e1008653. https://doi.org/10.1371/journal.ppat.1008653

Soto, C., & Satani, N. (2011). The intricate mechanisms of neurodegeneration in prion diseases. *Trends in Molecular Medicine*, *17*(1), 14–24. https://doi.org/10.1016/j.molmed.2010.09.001

Spilman, P., Lessard, P., Sattavat, M., Bush, C., Tousseyn, T., Huang, E. J., Giles, K., Golde, T., Das, P., Fauq, A., Prusiner, S. B., & Dearmond, S. J. (2008). A gamma-secretase inhibitor and quinacrine reduce prions and prevent dendritic degeneration in murine brains. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(30), 10595–10600. https://doi.org/10.1073/pnas.0803671105

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, *12*(1), 5692. https://doi.org/10.1038/s41467-021-25960-2

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews. Genetics*, *16*(3), 133–145. https://doi.org/10.1038/nrg3833

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sun, C., Zhang, J., Zheng, D., Wang, J., Yang, H., & Zhang, X. (2018). Transcriptome variations among human embryonic stem cell lines are associated with their differentiation propensity. *Plos One*, *13*(2), e0192625. https://doi.org/10.1371/journal.pone.0192625

Sun, N., Meng, X., Liu, Y., Song, D., Jiang, C., & Cai, J. (2021). Applications of brain organoids in neurodevelopment and neurological diseases. *Journal of Biomedical Science*, *28*(1), 30. https://doi.org/10.1186/s12929-021-00728-4

Suzuki, Y. (2020). Advent of a new sequencing era: long-read and on-site sequencing. *Journal of Human Genetics*, *65*(1), 1. https://doi.org/10.1038/s10038-019-0683-4

Swindell, W. R., Xing, X., Voorhees, J. J., Elder, J. T., Johnston, A., & Gudjonsson, J. E. (2014). Integrative RNA-seq and microarray data analysis reveals GC content and gene length biases in the psoriasis transcriptome. *Physiological Genomics*, *46*(15), 533–546. https://doi.org/10.1152/physiolgenomics.00022.2014

Szabo, P. A., Levitin, H. M., Miron, M., Snyder, M. E., Senda, T., Yuan, J., Cheng, Y. L., Bush, E. C., Dogra, P., Thapa, P., Farber, D. L., & Sims, P. A. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications*, *10*(1), 4706. https://doi.org/10.1038/s41467-019-12464-3

Taketo, M., Schroeder, A. C., Mobraaten, L. E., Gunning, K. B., Hanten, G., Fox, R. R., Roderick, T. H., Stewart, C. L., Lilly, F., & Hansen, C. T. (1991). FVB/N: an inbred mouse strain preferable for transgenic analyses. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(6), 2065–2069. https://doi.org/10.1073/pnas.88.6.2065

Tanaka, M., Chien, P., Naber, N., Cooke, R., & Weissman, J. S. (2004). Conformational variations in an infectious protein determine prion strain differences. *Nature*, *428*(6980), 323–328. https://doi.org/10.1038/nature02392

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., & Surani, M. A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell*, *6*(5), 468–478. https://doi.org/10.1016/j.stem.2010.03.015

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Tang, Y., & Le, W. (2016). Differential roles of M1 and M2 microglia in neurodegenerative diseases. *Molecular Neurobiology*, *53*(2), 1181–1194. https://doi.org/10.1007/s12035-014-9070-5

Taraboulos, A., Serban, D., & Prusiner, S. B. (1990). Scrapie prion proteins accumulate in the cytoplasm of persistently infected cultured cells. *The Journal of Cell Biology*, *110*(6), 2117–2132. https://doi.org/10.1083/jcb.110.6.2117

Taylor, D. R., & Hooper, N. M. (2006). The prion protein and lipid rafts. *Molecular Membrane Biology*, *23*(1), 89–99. https://doi.org/10.1080/09687860500449994

Taylor, S. C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., & Fenrich, J. (2019). The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time. *Trends in Biotechnology*, *37*(7), 761–774. https://doi.org/10.1016/j.tibtech.2018.12.002

Telling, G. C., Parchi, P., DeArmond, S. J., Cortelli, P., Montagna, P., Gabizon, R., Mastrianni, J., Lugaresi, E., Gambetti, P., & Prusiner, S. B. (1996). Evidence for the conformation of the pathologic isoform of the prion protein enciphering and propagating prion diversity. *Science*, *274*(5295), 2079–2082. https://doi.org/10.1126/science.274.5295.2079

Thomas, E. A., Dean, B., Pavey, G., & Sutcliffe, J. G. (2001). Increased CNS levels of apolipoprotein D in schizophrenic and bipolar subjects: implications for the pathophysiology of psychiatric disorders. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(7), 4066–4071. https://doi.org/10.1073/pnas.071056198

Thrupp, N., Sala Frigerio, C., Wolfs, L., Skene, N. G., Fattorelli, N., Poovathingal, S., Fourne, Y., Matthews, P. M., Theys, T., Mancuso, R., de Strooper, B., & Fiers, M. (2020). Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. *Cell Reports*, *32*(13), 108189. https://doi.org/10.1016/j.celrep.2020.108189

Tian, C., Liu, D., Sun, Q.-L., Chen, C., Xu, Y., Wang, H., Xiang, W., Kretzschmar, H. A., Li, W., Chen, C., Shi, Q., Gao, C., Zhang, J., Zhang, B.-Y., Han, J., & Dong, X.-P. (2013). Comparative analysis of gene expression profiles between cortex and thalamus in Chinese fatal familial insomnia patients. *Molecular Neurobiology*, *48*(1), 36–48. https://doi.org/10.1007/s12035-013-8426-6

Tian, C., Liu, D., Xiang, W., Kretzschmar, H. A., Sun, Q.-L., Gao, C., Xu, Y., Wang, H., Fan, X.-Y., Meng, G., Li, W., & Dong, X.-P. (2014). Analyses of the similarity and difference of global gene expression profiles in cortex regions of three neurodegenerative diseases: sporadic Creutzfeldt-Jakob disease (sCJD), fatal familial insomnia (FFI), and Alzheimer's disease (AD). *Molecular Neurobiology*, *50*(2), 473–481. https://doi.org/10.1007/s12035-014-8758-x

Ugalde, C. L., Lewis, V., Stehmann, C., McLean, C. A., Lawson, V. A., Collins, S. J., & Hill, A. F. (2020). Markers of A1 astrocytes stratify to molecular sub-types in sporadic Creutzfeldt-Jakob disease brain. *Brain Communications*, *2*(2), fcaa029. https://doi.org/10.1093/braincomms/fcaa029

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., & Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, *14*(6), 565–571. https://doi.org/10.1038/nmeth.4292

Van Der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-SNE. *J Mach Learn Res*, *9*(26).

Vilette, D., Andreoletti, O., Archer, F., Madelaine, M. F., Vilotte, J. L., Lehmann, S., & Laude, H. (2001). Ex vivo propagation of infectious sheep scrapie agent in heterologous epithelial cells expressing ovine prion protein. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(7), 4055–4059. https://doi.org/10.1073/pnas.061337998

Vincenti, J. E., Murphy, L., Grabert, K., McColl, B. W., Cancellotti, E., Freeman, T. C., & Manson, J. C. (2015). Defining the Microglia Response during the Time Course of Chronic Neurodegeneration. *Journal of Virology*, *90*(6), 3003–3017. https://doi.org/10.1128/JVI.02613-15

Vincent, A. J., Gasperini, R., Foa, L., & Small, D. H. (2010). Astrocytes in Alzheimer's disease: emerging roles in calcium dysregulation and synaptic plasticity. *Journal of Alzheimer's Disease*, *22*(3), 699–714. https://doi.org/10.3233/JAD-2010-101089

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., Carbone, L., Steemers, F. J., & Adey, A. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, *14*(3), 302–308. https://doi.org/10.1038/nmeth.4154

Wadsworth, J. D. F., Adamson, G., Joiner, S., Brock, L., Powell, C., Linehan, J. M., Beck, J. A., Brandner, S., Mead, S., & Collinge, J. (2017). Methods for molecular diagnosis of human prion disease. *Methods in Molecular Biology*, *1658*, 311–346. https://doi.org/10.1007/978-1-4939-7244-9_22

Wadsworth, J. D. F., Joiner, S., Linehan, J. M., Jack, K., Al-Doujaily, H., Costa, H., Ingold, T., Taema, M., Zhang, F., Sandberg, M. K., Brandner, S., Tran, L., Vikøren, T., Våge, J., Madslien, K., Ytrehus, B., Benestad, S. L., Asante, E. A., & Collinge, J. (2021). Humanised transgenic mice are resistant to chronic wasting disease prions from Norwegian reindeer and moose. *The Journal of Infectious Diseases*. https://doi.org/10.1093/infdis/jiab033

Walker, D. G., Link, J., Lue, L.-F., Dalsing-Hernandez, J. E., & Boyes, B. E. (2006). Gene expression changes by amyloid beta peptide-stimulated human postmortem brain microglia identify activation of multiple inflammatory processes. *Journal of Leukocyte Biology*, *79*(3), 596–610. https://doi.org/10.1189/jlb.0705377

Walker, L. C., & Jucker, M. (2015). Neurodegenerative diseases: expanding the prion concept. *Annual Review of Neuroscience*, *38*, 87–103. https://doi.org/10.1146/annurev-neuro-071714-033828

Wallraff, A., Odermatt, B., Willecke, K., & Steinhäuser, C. (2004). Distinct types of astroglial cells in the hippocampus differ in gap junction coupling. *Glia*, *48*(1), 36–43. https://doi.org/10.1002/glia.20040

Wang, H. (2021). Microglia Heterogeneity in Alzheimer’s Disease: Insights From Single-Cell Technologies. *Frontiers in Synaptic Neuroscience*, *13*, 773590. https://doi.org/10.3389/fnsyn.2021.773590

Wang, S., Wang, B., Shang, D., Zhang, K., Yan, X., & Zhang, X. (2022). Ion channel dysfunction in astrocytes in neurodegenerative diseases. *Frontiers in Physiology*, *13*, 814285. https://doi.org/10.3389/fphys.2022.814285

Wang, T., Li, B., Nelson, C. E., & Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, *20*(1), 40. https://doi.org/10.1186/s12859-019-2599-6

Wang, Y., Cella, M., Mallinson, K., Ulrich, J. D., Young, K. L., Robinette, M. L., Gilfillan, S., Krishnan, G. M., Sudhakar, S., Zinselmeyer, B. H., Holtzman, D. M., Cirrito, J. R., & Colonna, M. (2015). TREM2 lipid sensing sustains the microglial response in an Alzheimer’s disease model. *Cell*, *160*(6), 1061–1071. https://doi.org/10.1016/j.cell.2015.01.049

Ward, H. J. T., Everington, D., Cousens, S. N., Smith-Bathgate, B., Gillies, M., Murray, K., Knight, R. S. G., Smith, P. G., & Will, R. G. (2008). Risk factors for sporadic Creutzfeldt-Jakob disease. *Annals of Neurology*, *63*(3), 347–354. https://doi.org/10.1002/ana.21294

Watson, J. D., & Crick, F. H. (1953). The structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, *18*, 123–131. https://doi.org/10.1101/SQB.1953.018.01.020

Watts, J. C., & Prusiner, S. B. (2014). Mouse models for studying the formation and propagation of prions. *The Journal of Biological Chemistry*, *289*(29), 19841–19849. https://doi.org/10.1074/jbc.R114.550707

Wemheuer, W. M., Wrede, A., & Schulz-Schaeffer, W. J. (2017). Types and strains: their essential role in understanding protein aggregation in neurodegenerative diseases. *Frontiers in Aging Neuroscience*, *9*, 187. https://doi.org/10.3389/fnagi.2017.00187

Williams, A. E., Lawson, L. J., Perry, V. H., & Fraser, H. (1994). Characterization of the microglial response in murine scrapie. *Neuropathology and Applied Neurobiology*, *20*(1), 47–55.

Woerman, A. L., Stöhr, J., Aoyagi, A., Rampersaud, R., Krejciova, Z., Watts, J. C., Ohyama, T., Patel, S., Widjaja, K., Oehler, A., Sanders, D. W., Diamond, M. I., Seeley, W. W., Middleton, L. T., Gentleman, S. M., Mordes, D. A., Südhof, T. C., Giles, K., & Prusiner, S. B. (2015). Propagation of prions causing synucleinopathies in cultured cells. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(35), E4949-58. https://doi.org/10.1073/pnas.1513426112

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Cambridge (Mass.))*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Xiang, W., Windl, O., Westner, I. M., Neumann, M., Zerr, I., Lederer, R. M., & Kretzschmar, H. A. (2005). Cerebral gene expression profiles in sporadic Creutzfeldt-Jakob disease. *Annals of Neurology*, *58*(2), 242–257. https://doi.org/10.1002/ana.20551

Yamada, T., Akiyama, H., & McGeer, P. L. (1990). Complement-activated oligodendroglia: a new pathogenic entity identified by immunostaining with antibodies to human complement proteins C3d and C4d. *Neuroscience Letters*, *112*(2–3), 161–166. https://doi.org/10.1016/0304-3940(90)90196-g

Yamada, T., McGeer, P. L., & McGeer, E. G. (1991). Relationship of Complement-Activated Oligodendrocytes to Reactive Microglia and Neuronal Pathology in Neurodegenerative Disease. *Dementia and Geriatric Cognitive Disorders*, *2*(2), 71–77. https://doi.org/10.1159/000107179

Yang, Z., & Wang, K. K. W. (2015). Glial fibrillary acidic protein: from intermediate filament assembly and gliosis to neurobiomarker. *Trends in Neurosciences*, *38*(6), 364–374. https://doi.org/10.1016/j.tins.2015.04.003

Yu, G., Chen, J., Xu, Y., Zhu, C., Yu, H., Liu, S., Sha, H., Chen, J., Xu, X., Wu, Y., Zhang, A., Ma, J., & Cheng, G. (2009). Generation of goats lacking prion protein. *Molecular Reproduction and Development*, *76*(1), 3. https://doi.org/10.1002/mrd.20960

Zappia, L., Phipson, B., & Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Computational Biology*, *14*(6), e1006245. https://doi.org/10.1371/journal.pcbi.1006245

Zeidler, M., Stewart, G., Cousens, S. N., Estibeiro, K., & Will, R. G. (1997). Codon 129 genotype and new variant CJD. *The Lancet*, *350*(9078), 668. https://doi.org/10.1016/s0140-6736(05)63366-1

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., & Linnarsson, S. (2018). Molecular architecture of the mouse nervous system. *Cell*, *174*(4), 999-1014.e22. https://doi.org/10.1016/j.cell.2018.06.021

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., Zhang, C., Xie, T., Tran, L., Dobrin, R., Fluder, E., Clurman, B., Melquist, S., Narayanan, M., Suver, C., Shah, H., Mahajan, M., Gillis, T., Mysore, J., … Emilsson, V. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer’s disease. *Cell*, *153*(3), 707–720. https://doi.org/10.1016/j.cell.2013.03.030

Zhang, Xiannian, Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., & Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, *73*(1), 130-142.e5. https://doi.org/10.1016/j.molcel.2018.10.020

Zhang, Xiao, Wan, J.-Q., & Tong, X.-P. (2018). Potassium channel dysfunction in neurons and astrocytes in Huntington’s disease. *CNS Neuroscience & Therapeutics*, *24*(4), 311–318. https://doi.org/10.1111/cns.12804

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *Plos One*, *9*(1), e78644. https://doi.org/10.1371/journal.pone.0078644

Zhao, Y., Li, X., Zhao, W., Wang, J., Yu, J., Wan, Z., Gao, K., Yi, G., Wang, X., Fan, B., Wu, Q., Chen, B., Xie, F., Wu, J., Zhang, W., Chen, F., Yang, H., Wang, J., Xu, X., … Liu, X. (2019). Single-cell transcriptomic landscape of nucleated cells in umbilical cord blood. *GigaScience*, *8*(5). https://doi.org/10.1093/gigascience/giz047

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049. https://doi.org/10.1038/ncomms14049

Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., Xu, X., Tang, F., Zhang, J., Qiao, J., & Wang, X. (2018). A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*, *555*(7697), 524–528. https://doi.org/10.1038/nature25980

Zhou, X., Liu, Z., Shen, K., Zhao, P., & Sun, M.-X. (2020). Cell lineage-specific transcriptome analysis for interpreting cell fate specification of proembryos. *Nature Communications*, *11*(1), 1366. https://doi.org/10.1038/s41467-020-15189-w

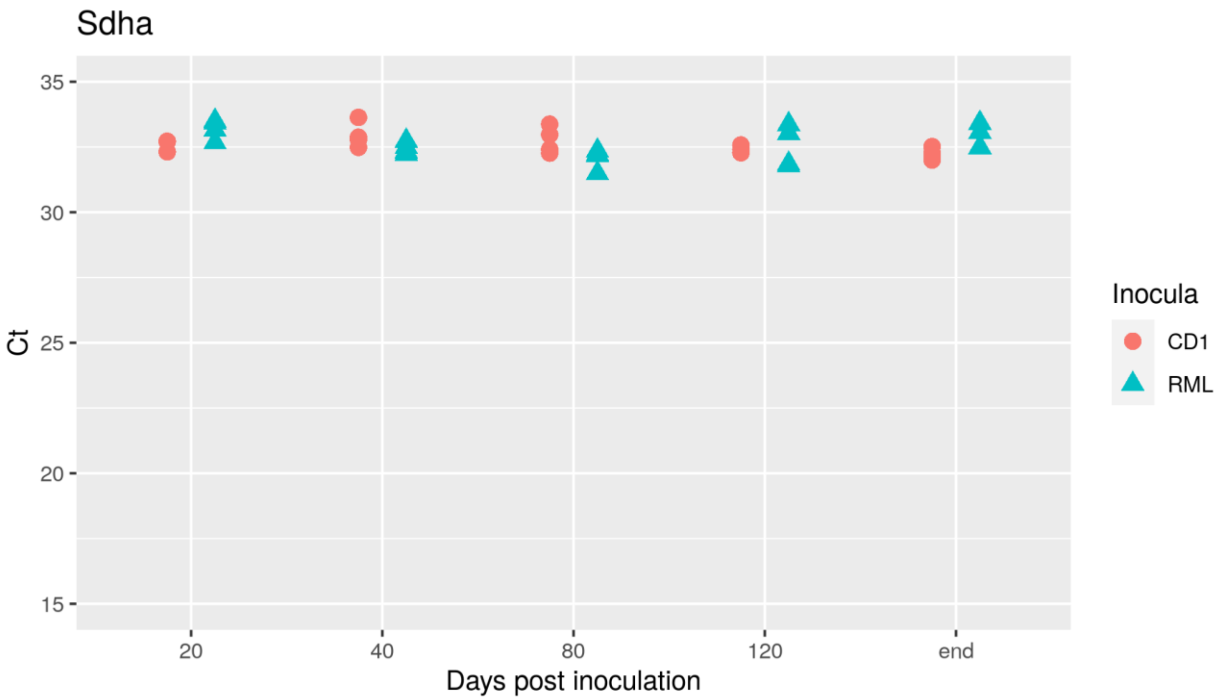
Zhu, A., Ibrahim, J. G., & Love, M. I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, *35*(12), 2084–2092. https://doi.org/10.1093/bioinformatics/bty895

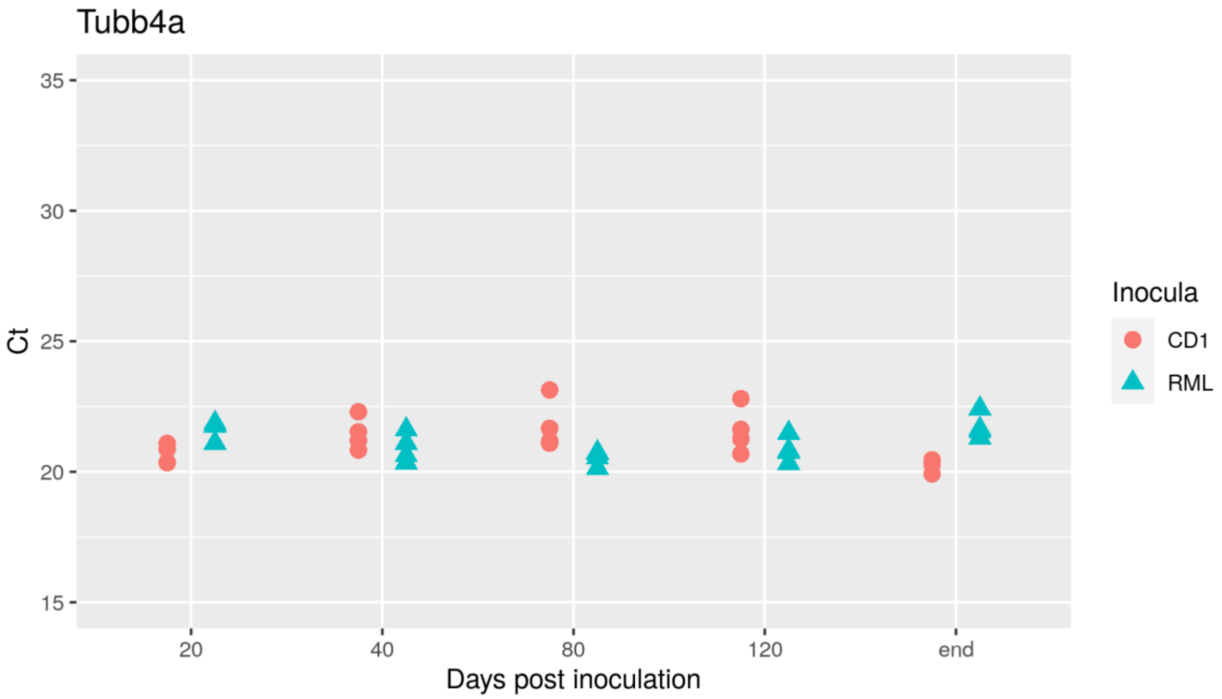
Zhu, S., Qing, T., Zheng, Y., Jin, L., & Shi, L. (2017). Advances in single-cell RNA sequencing and its applications in cancer research. *Oncotarget*, *8*(32), 53763–53779. https://doi.org/10.18632/oncotarget.17893

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., & Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, *65*(4), 631-643.e4. https://doi.org/10.1016/j.molcel.2017.01.023

# 7 Supplementary materials

## 7.1 Figures

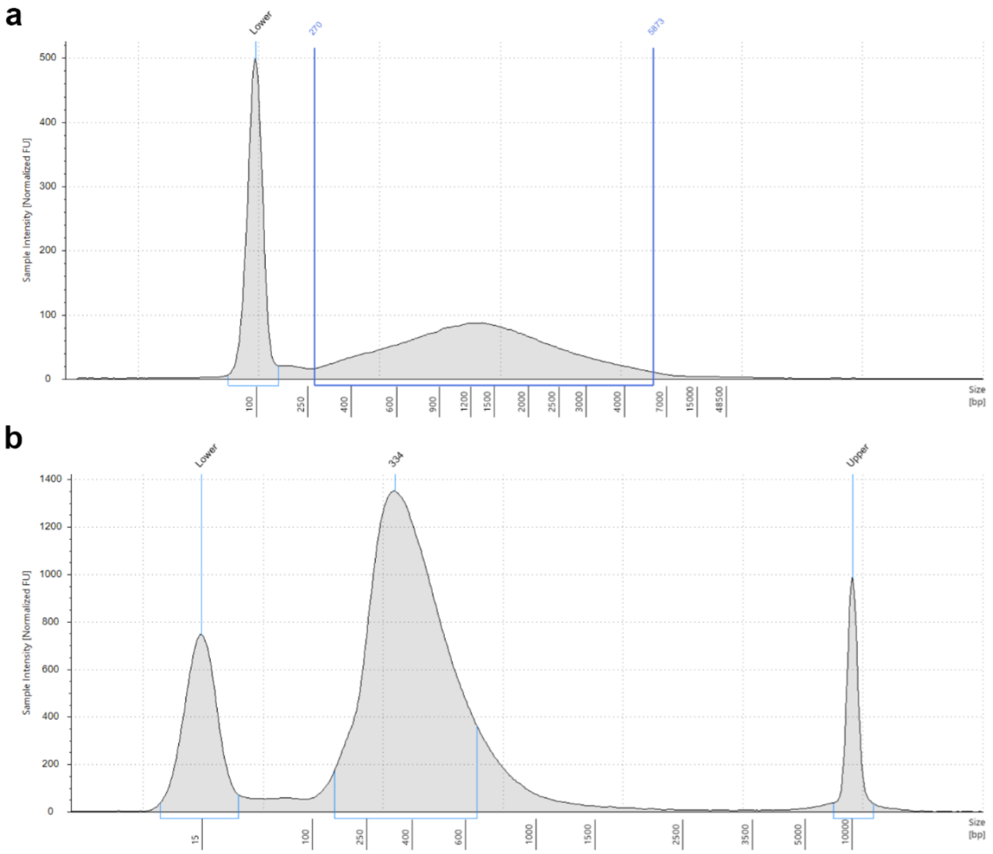




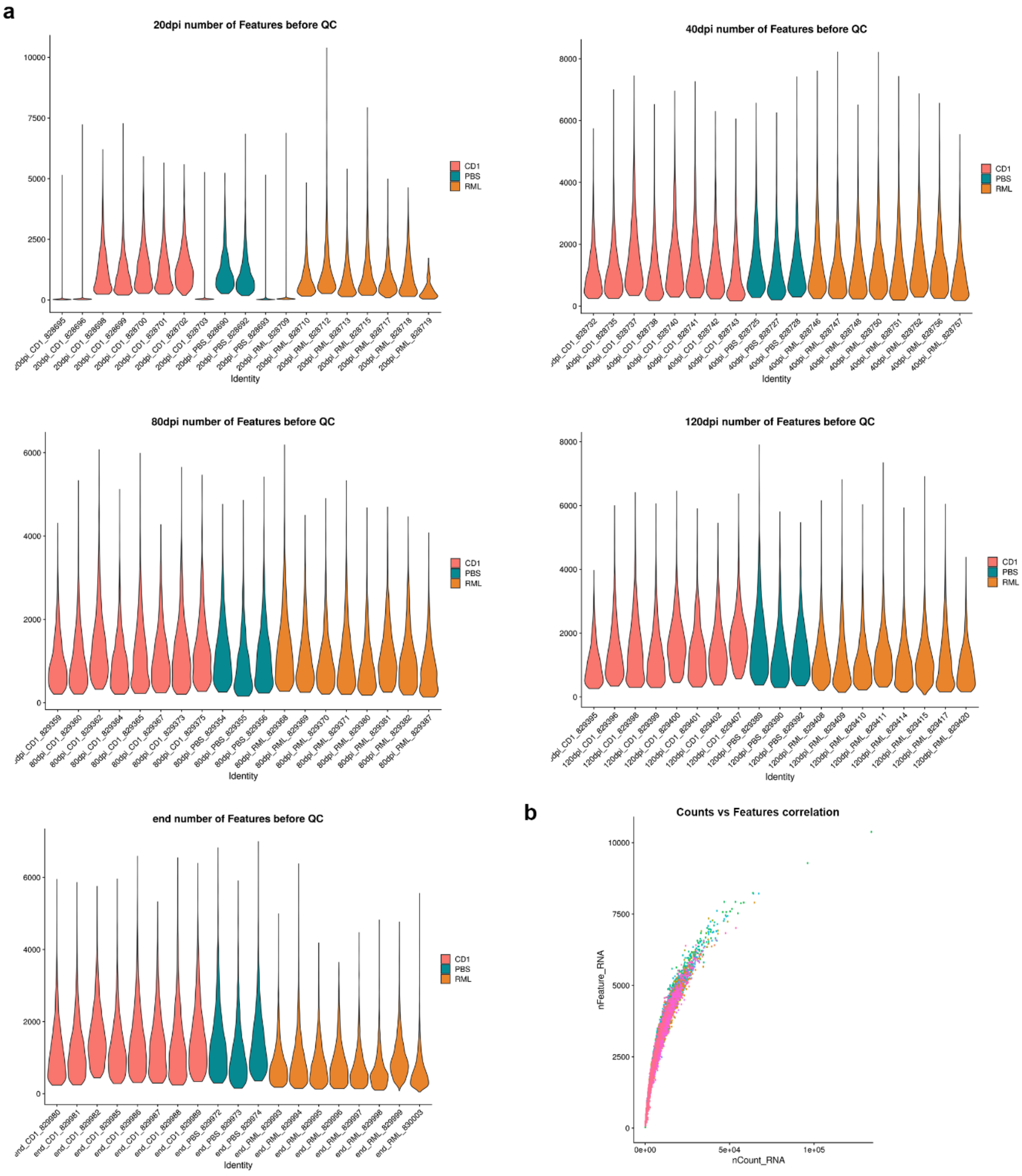
***Supplementary Figure 1: Sdha and Tubb4a were selected as internal reference genes for the normalisation of the real-time PCR data.*** *Raw Ct values were plotted for the two endogenous control genes. N = 4 biologically independent samples in each time point/inoculum combination.*

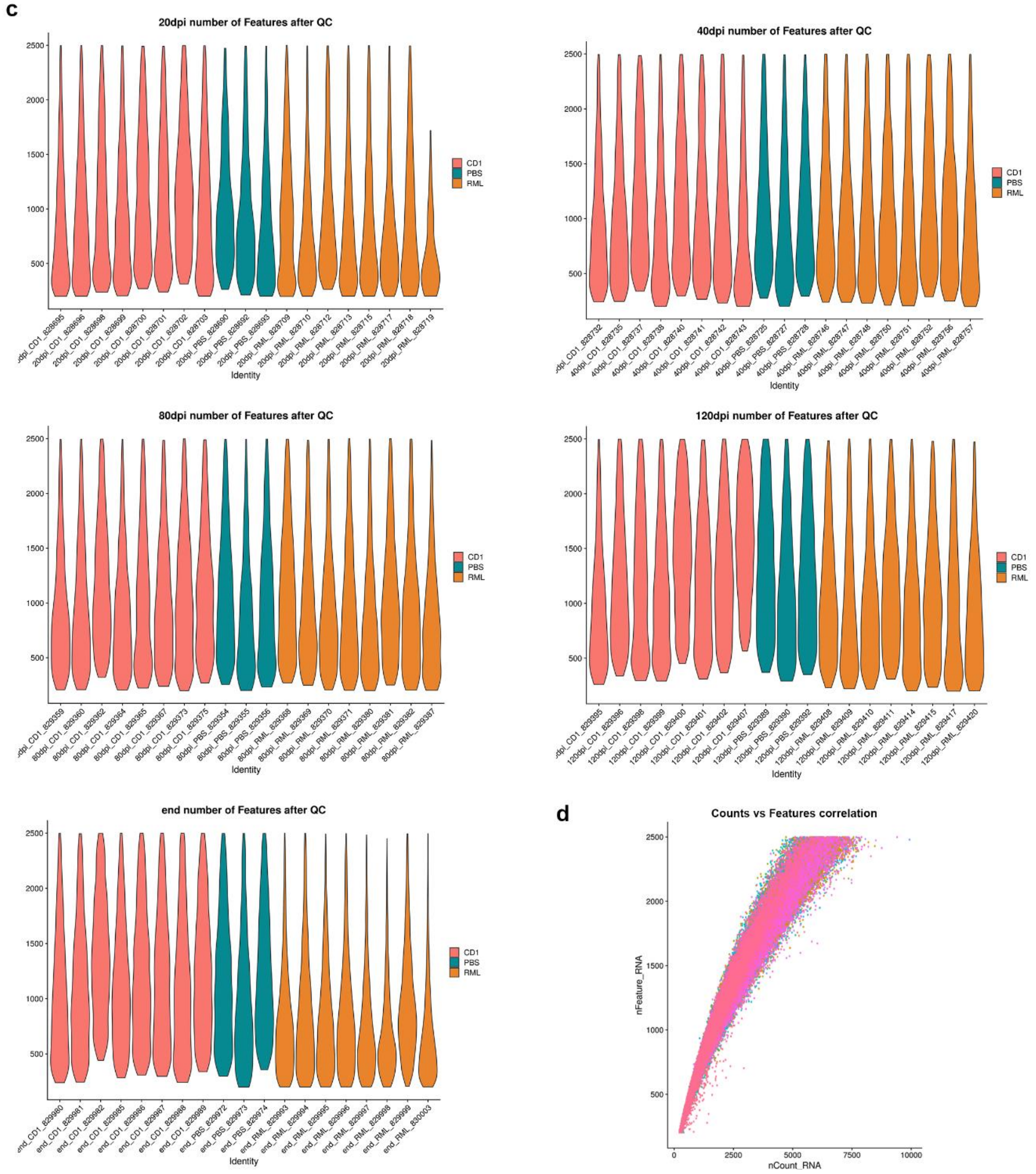
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 972 |  | 973 |  | 974 |  | 980 |  |  | 981 |  |  |
| B | 982 |  |  | 985 |  |  | 986 |  |  | 987 |  | 988 |
| C | 988 | 989 |  | 999 |  |  | 003 |  |  | 993 |  |  |
| D | 994 |  |  | 995 |  |  | 996 |  | 997 |  | 998 |  |
| E |  |  |  |  |  |  |  |  |  |  |  |  |
| F |  |  |  |  |  |  |  |  |  |  |  |  |
| G |  |  |  |  |  |  |  |  |  |  |  |  |
| H |  |  |  |  |  |  |  |  |  |  |  |  |

***Supplementary Figure 2: A representative image of the layout of a loaded SPLiT-seq 96-well plate used for the 1st round of barcoding.*** *The colours represent samples of different experimental groups: green for PBS, blue for CD1, and red for RML. The numbers correspond to sample IDs (from the end-stage mice in this case).*



***Supplementary Figure 3: Representative TapeStation traces of (a) a library after the first PCR amplification and (b) after tagmentation.** Samples were run on (a) a gDNA or (b) an HSD5000 tape. The x-axis represents the molecular weights of analysed nucleic acids and the y-axis the fluorescence intensity, which corresponds to the mass of nucleic acids assayed. The "Lower" and "Upper" peaks correspond to internal standards included in the loading buffer, essential for accurate molecular weight estimation and quantification.*

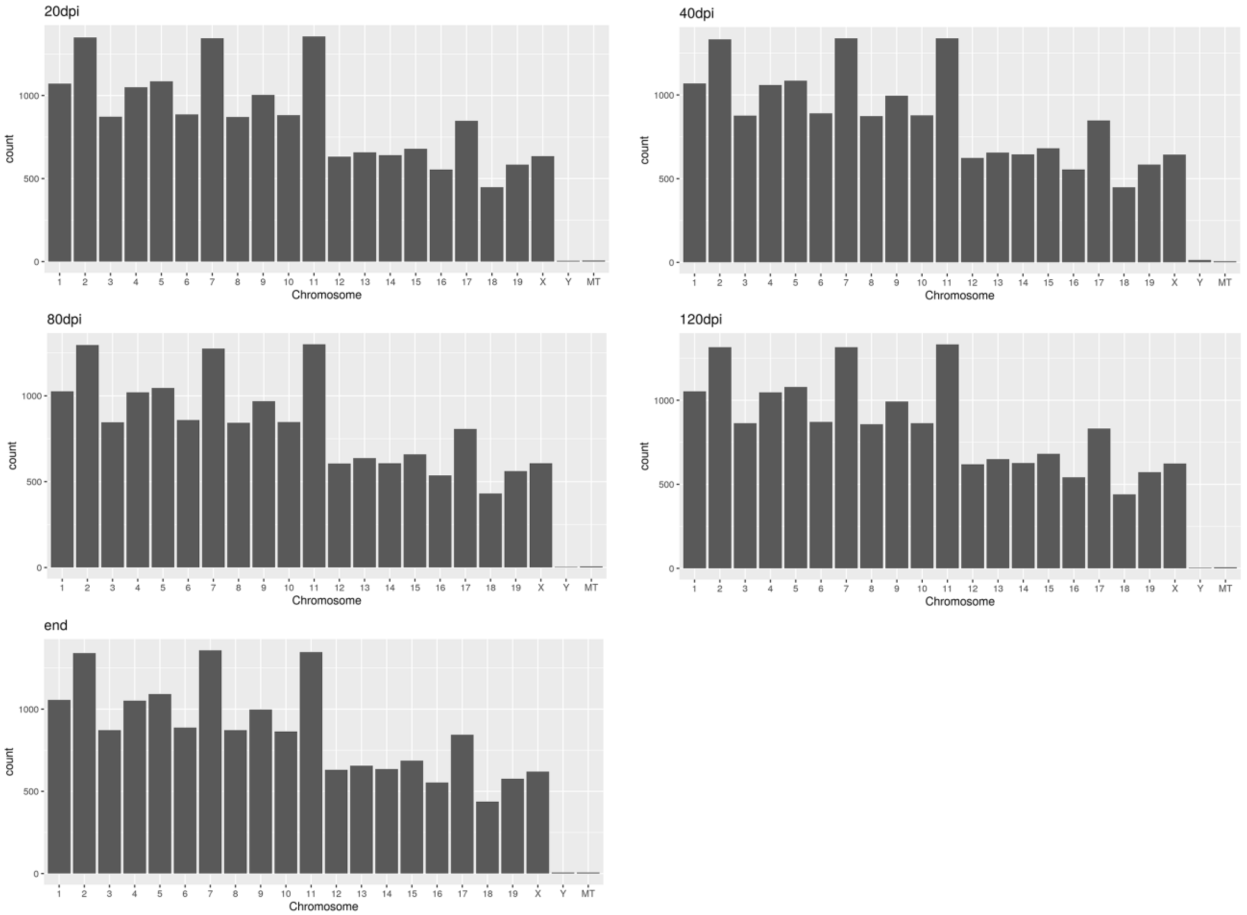

a
20dpi number of Features before QC
10000
7500
5000
2500
0
0dpi_CD1_828695
20dpi_CD1_828696
20dpi_CD1_828698
20dpi_CD1_828699
20dpi_CD1_828700
20dpi_CD1_828701
20dpi_CD1_828702
20dpi_CD1_828703
20dpi_PBS_828690
20dpi_PBS_828692
20dpi_PBS_828693
20dpi_RML_828709
20dpi_RML_828710
20dpi_RML_828712
20dpi_RML_828713
20dpi_RML_828715
20dpi_RML_828717
20dpi_RML_828718
20dpi_RML_828719
Identity
CD1
PBS
RML
40dpi number of Features before QC
8000
6000
4000
2000
0
dpi_CD1_828732
40dpi_CD1_828735
40dpi_CD1_828737
40dpi_CD1_828738
40dpi_CD1_828740
40dpi_CD1_828741
40dpi_CD1_828742
40dpi_CD1_828743
40dpi_PBS_828725
40dpi_PBS_828727
40dpi_PBS_828728
40dpi_RML_828746
40dpi_RML_828747
40dpi_RML_828748
40dpi_RML_828750
40dpi_RML_828751
40dpi_RML_828752
40dpi_RML_828756
40dpi_RML_828757
Identity
CD1
PBS
RML
80dpi number of Features before QC
6000
4000
2000
0
dpi_CD1_829359
80dpi_CD1_829360
80dpi_CD1_829362
80dpi_CD1_829364
80dpi_CD1_829365
80dpi_CD1_829367
80dpi_CD1_829373
80dpi_CD1_829375
80dpi_PBS_829354
80dpi_PBS_829355
80dpi_PBS_829356
80dpi_RML_829368
80dpi_RML_829369
80dpi_RML_829370
80dpi_RML_829371
80dpi_RML_829380
80dpi_RML_829381
80dpi_RML_829382
80dpi_RML_829387
Identity
CD1
PBS
RML
120dpi number of Features before QC
8000
6000
4000
2000
0
dpi_CD1_829395
120dpi_CD1_829396
120dpi_CD1_829398
120dpi_CD1_829399
120dpi_CD1_829400
120dpi_CD1_829401
120dpi_CD1_829402
120dpi_CD1_829407
120dpi_PBS_829389
120dpi_PBS_829390
120dpi_PBS_829392
120dpi_RML_829408
120dpi_RML_829409
120dpi_RML_829410
120dpi_RML_829411
120dpi_RML_829414
120dpi_RML_829415
120dpi_RML_829417
120dpi_RML_829420
Identity
CD1
PBS
RML
end number of Features before QC
6000
4000
2000
0
end_CD1_829980
end_CD1_829981
end_CD1_829982
end_CD1_829985
end_CD1_829986
end_CD1_829987
end_CD1_829988
end_CD1_829989
end_PBS_829972
end_PBS_829973
end_PBS_829974
end_RML_829993
end_RML_829994
end_RML_829995
end_RML_829996
end_RML_829997
end_RML_829998
end_RML_829999
end_RML_830003
Identity
CD1
PBS
RML
b
Counts vs Features correlation
10000
7500
5000
2500
0
nFeature_RNA
0e+00
5e+04
1e+05
nCount_RNA

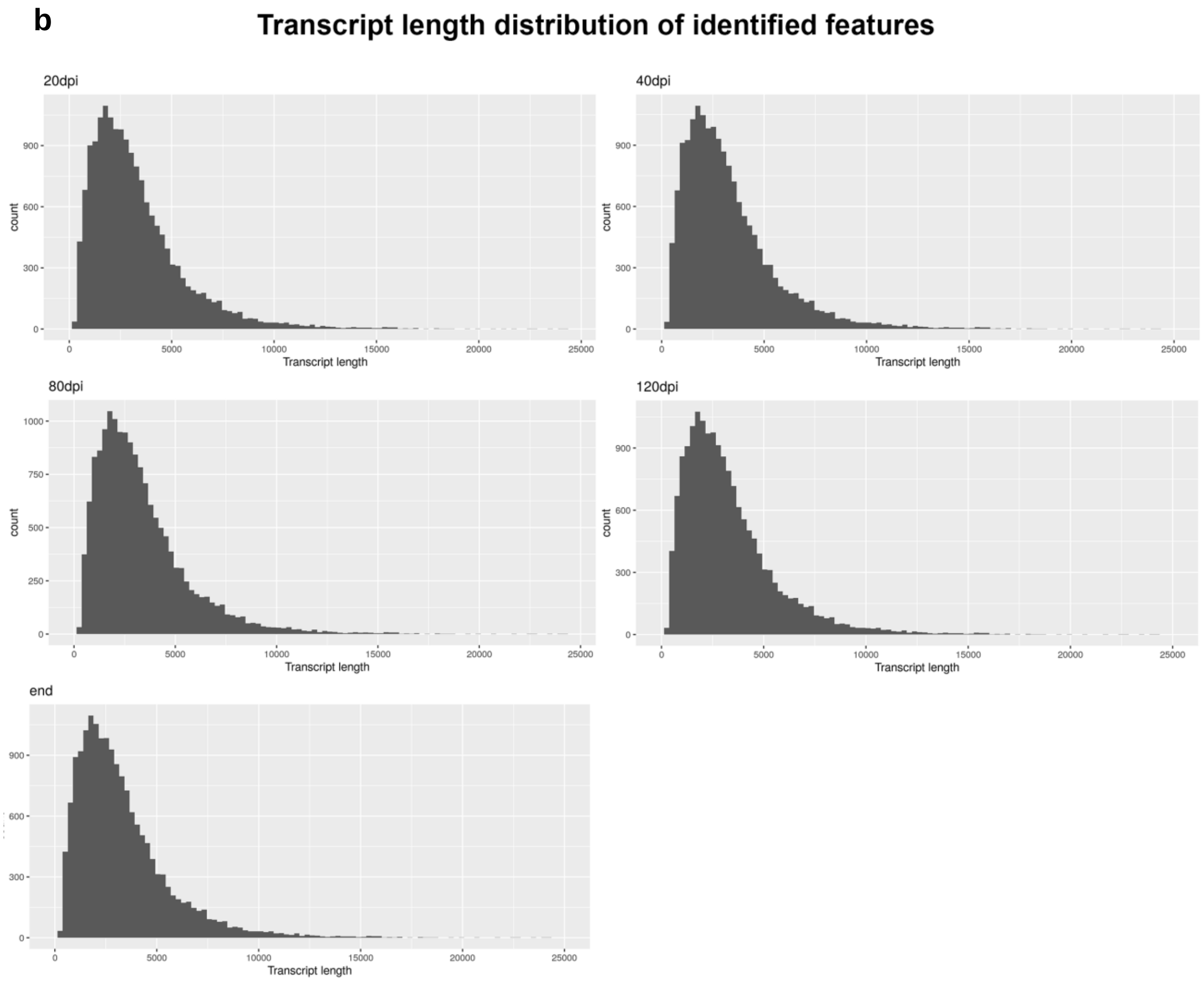


***Supplementary Figure 4: Filtering of the mouse transcriptomics data. (a)*** *The number of features (genes) identified before data filtering.* ***(b)*** *Correlation between the number of counts and features before data filtering.* ***(c)*** *The number of features identified after filtering out cells with fewer than 250 or more than 2500 features or a mitochondrial gene percentage of more than 1%.* ***(d)*** *Correlation between the number of counts and features after data filtering. Each violin represents a biological sample, colours represent the experimental groups.*

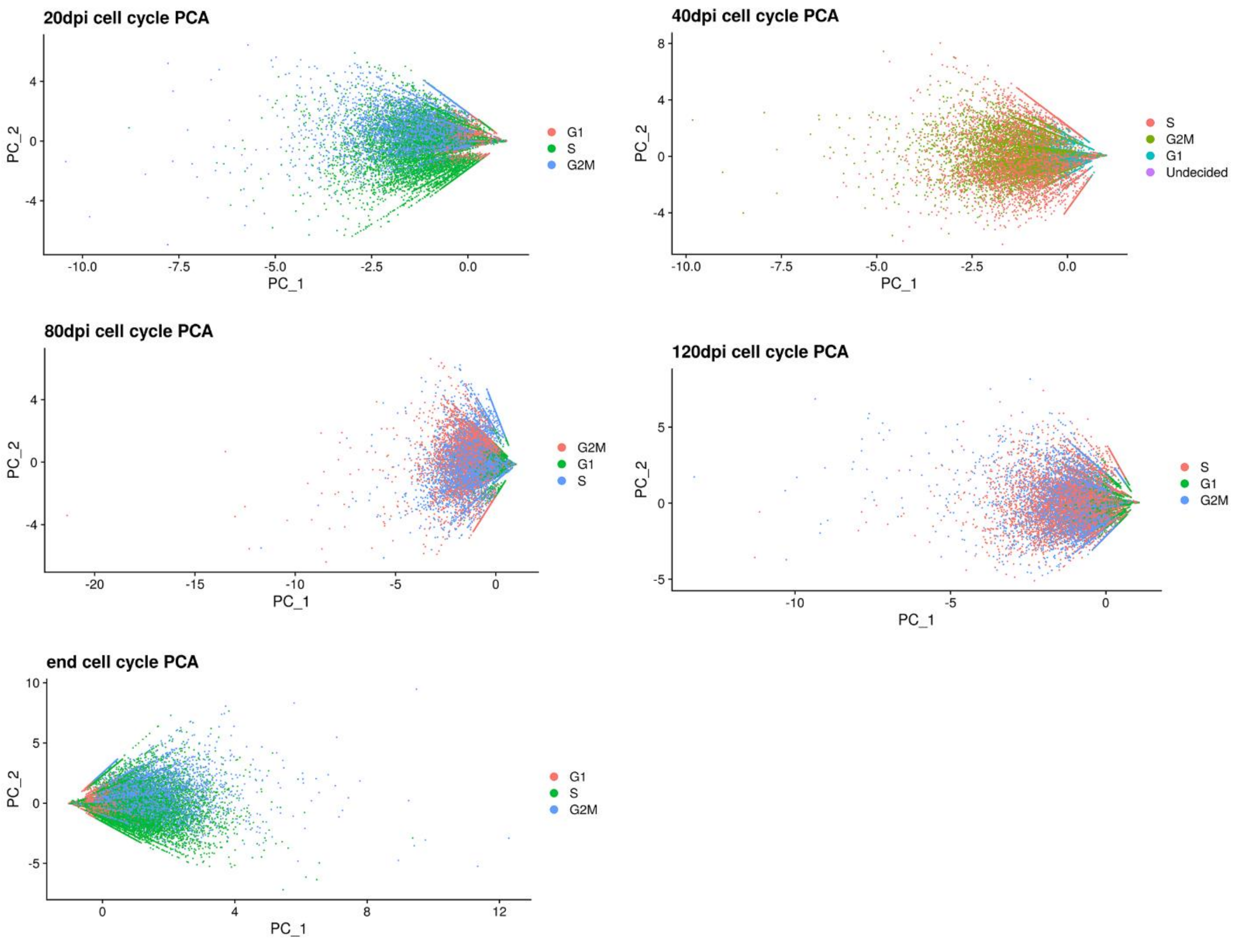
**a**

## Chromosome distribution of identified features

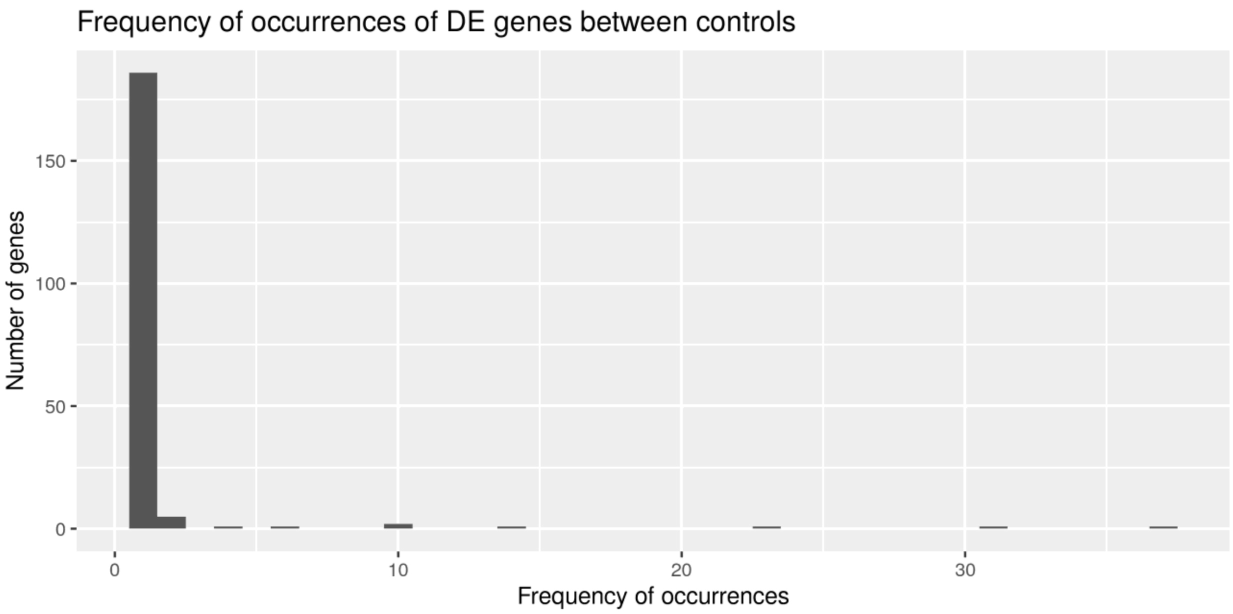




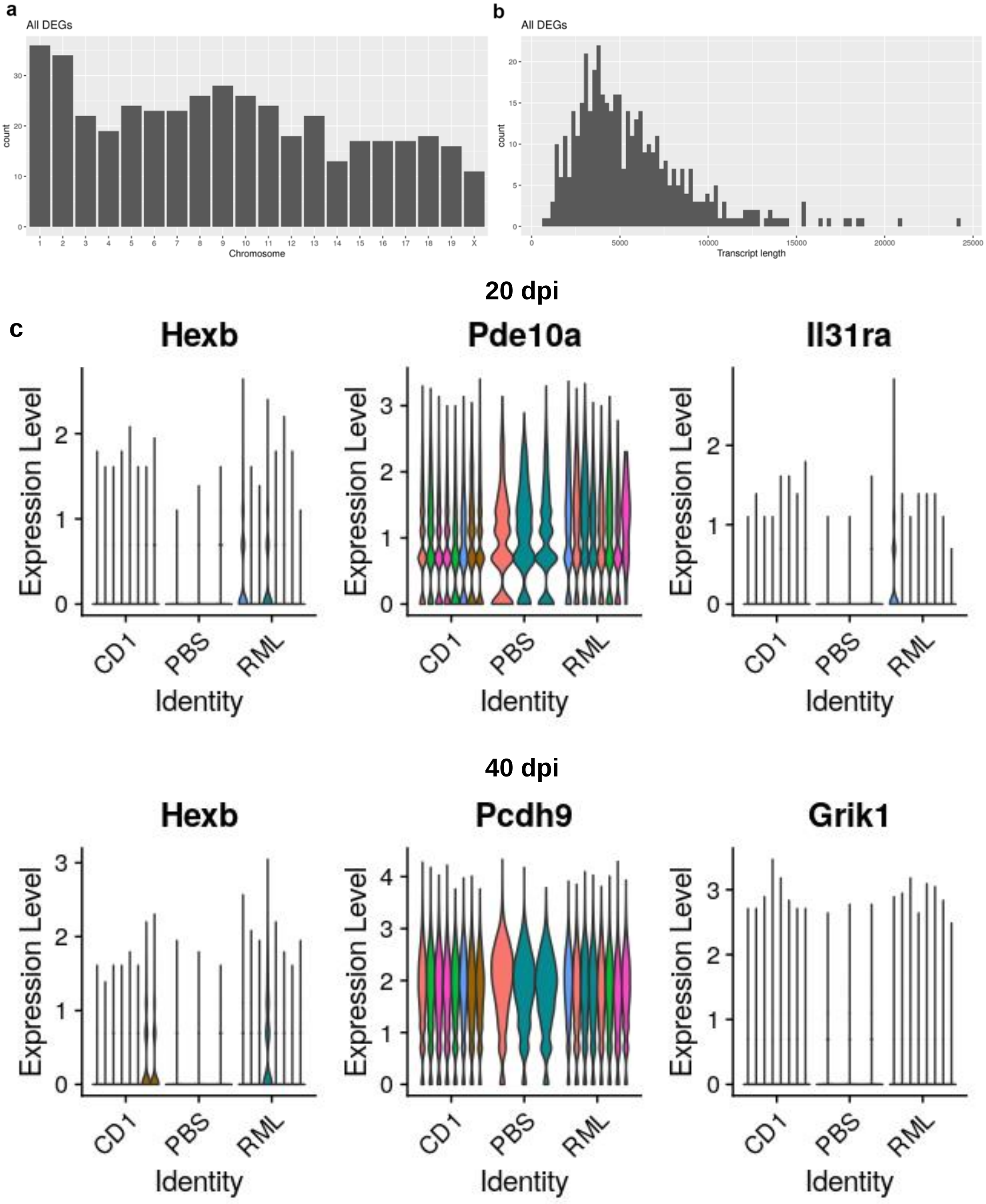
***Supplementary Figure 5: No bias was identified regarding the transcript lengths or chromosomes of the identified features. (a)*** *Histograms of the chromosomes where identified features reside.* ***(b)*** *Histograms of the transcript lengths of identified features. All identified features from each time point were used for these plots. Transcript length in nucleotides.*



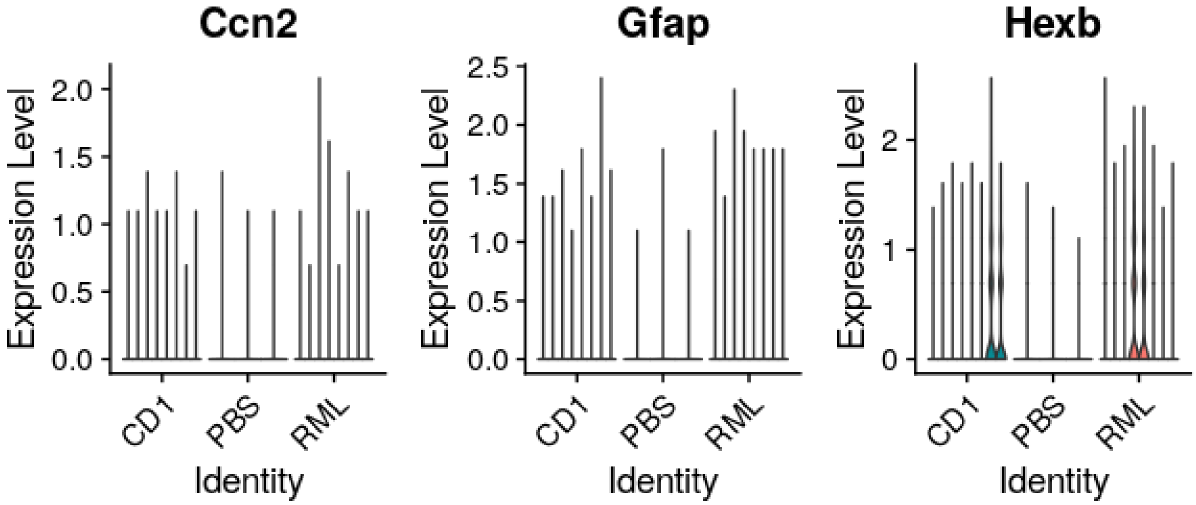
***Supplementary Figure 6: PCA plots of cell cycle genes suggest modest cell cycle effects.*** *Clusters of cells are not separated in the two-dimensional space of the first two principal components based on the cell cycle. The cell cycle genes used for the computation of the PCs are listed in the methods section.*



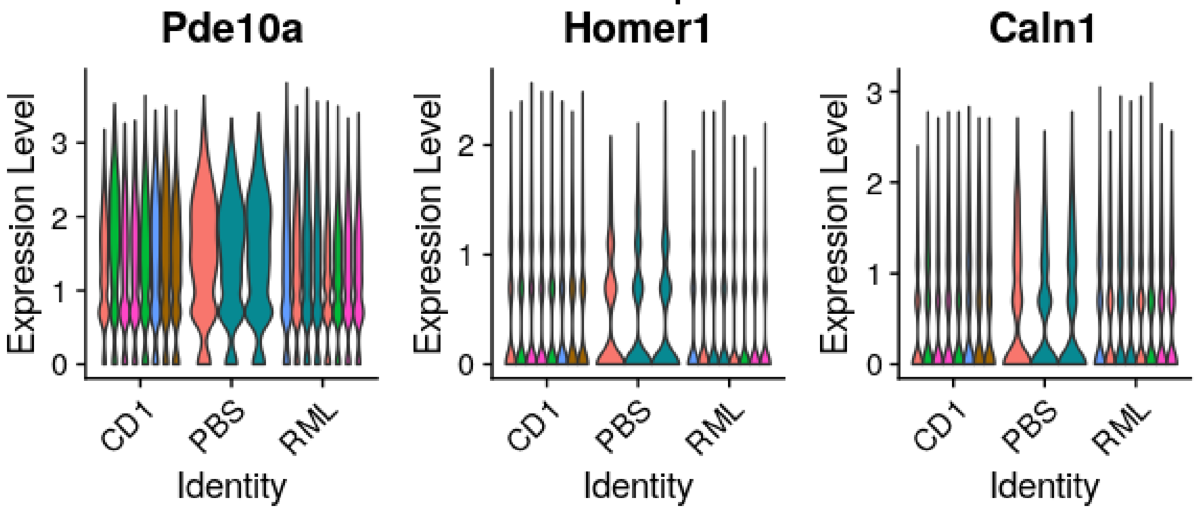
***Supplementary Figure 7: The majority of differentially expressed genes between the two control groups are identified only once across all cell clusters and time points.*** *The histogram shows the frequency that a gene was identified to be differentially expressed across all time points and cell clusters. After setting an occurrence threshold of 5, we identified a set of 7 genes that strongly deviated from the rest (Calm1, Cdk8, Cmss1, Malat1, mt-Rnr1, mt-Rnr2, and Rn18s) and were subsequently flagged for removal.*

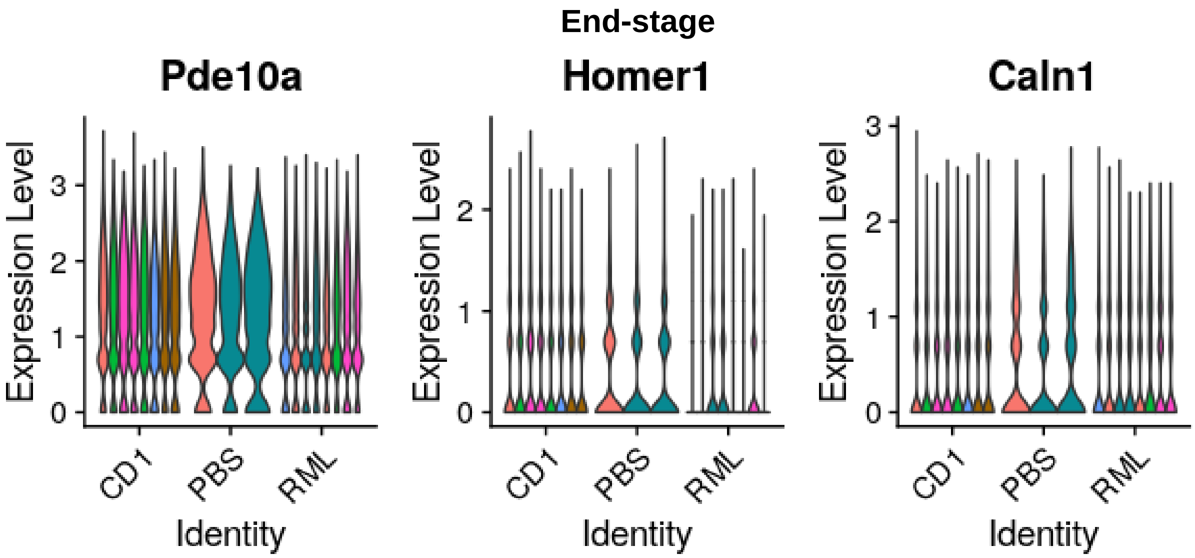

a
All DEGs
count
0
10
20
30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
b
All DEGs
count
0
5
10
15
20
0
5000
10000
15000
20000
25000
Transcript length
20 dpi
c
Hexb
Pde10a
Il31ra
Expression Level
Identity
CD1
PBS
RML
40 dpi
Hexb
Pcdh9
Grik1
Expression Level
Identity
CD1
PBS
RML

80 dpi

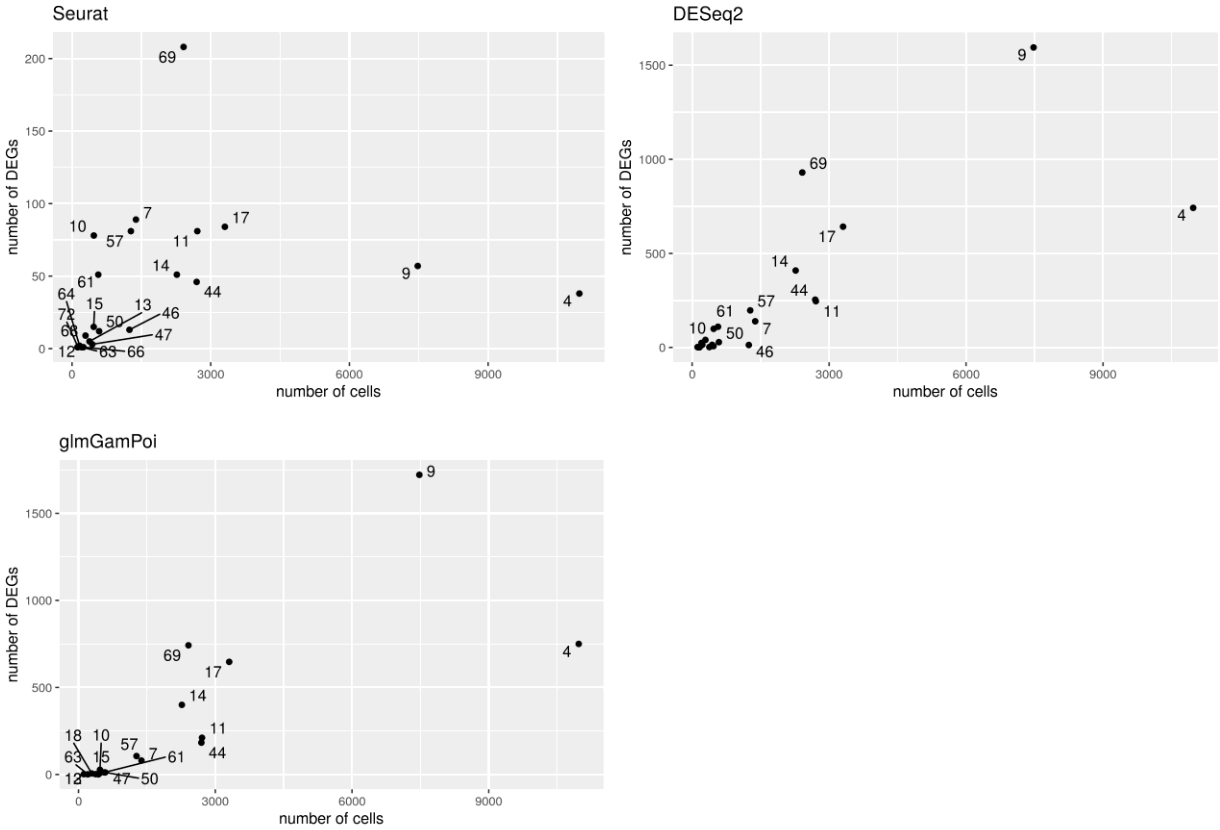


120 dpi

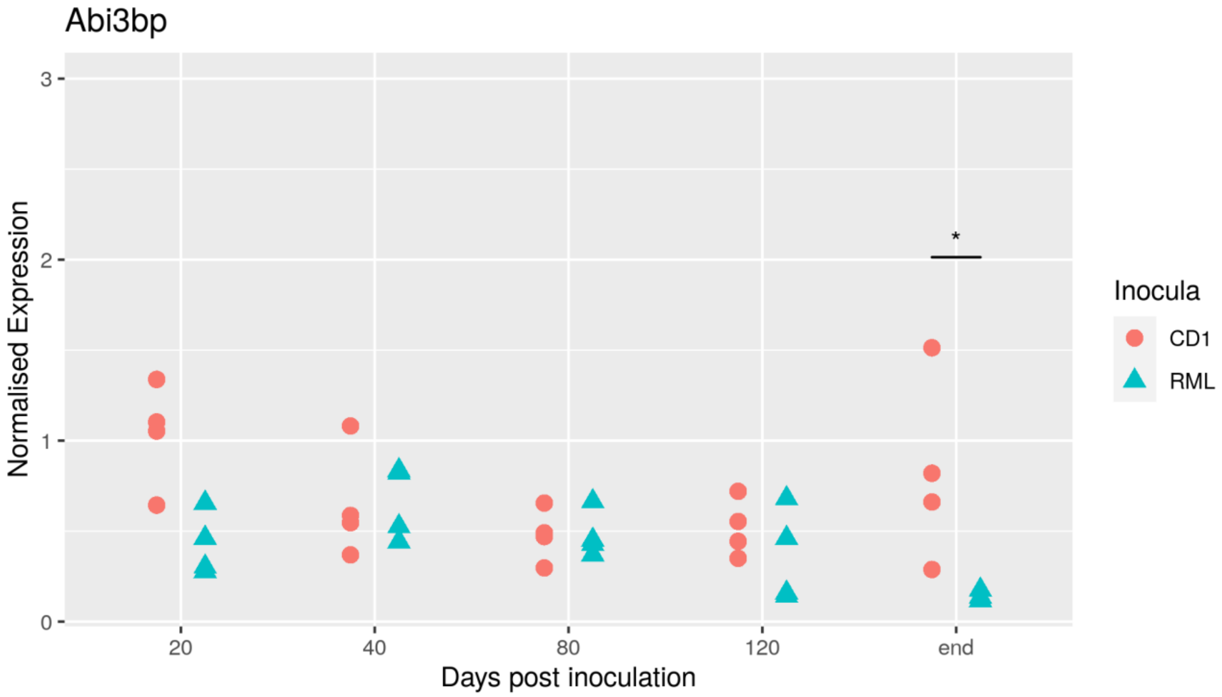


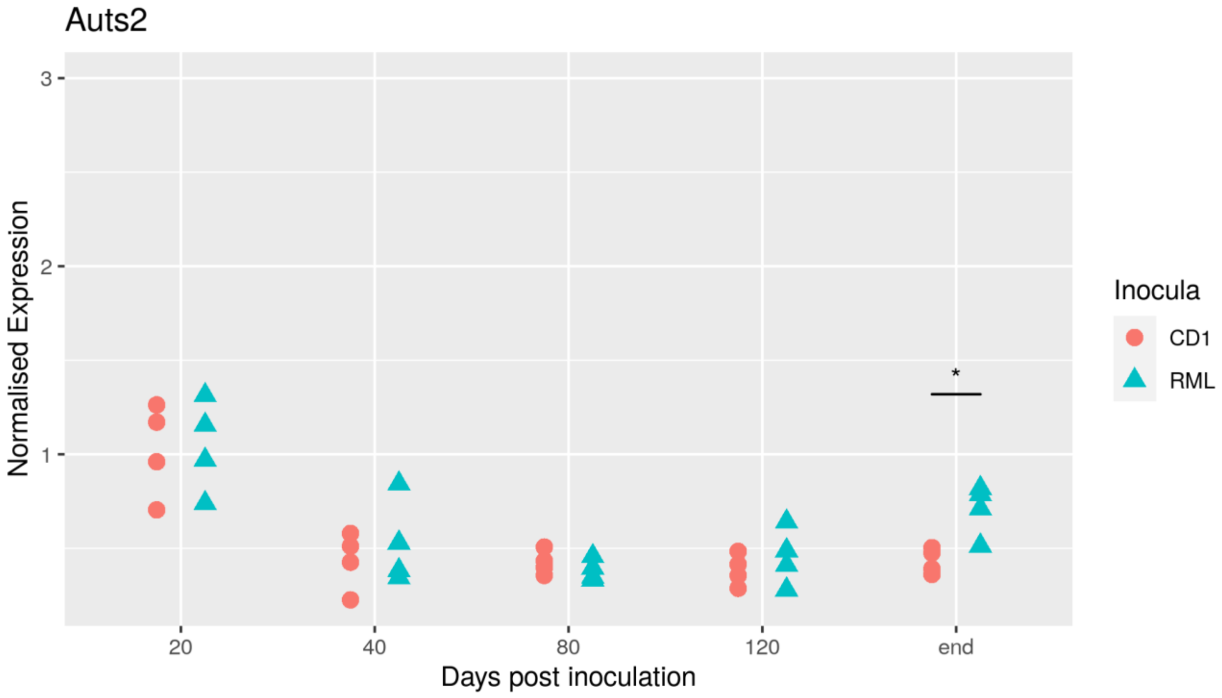


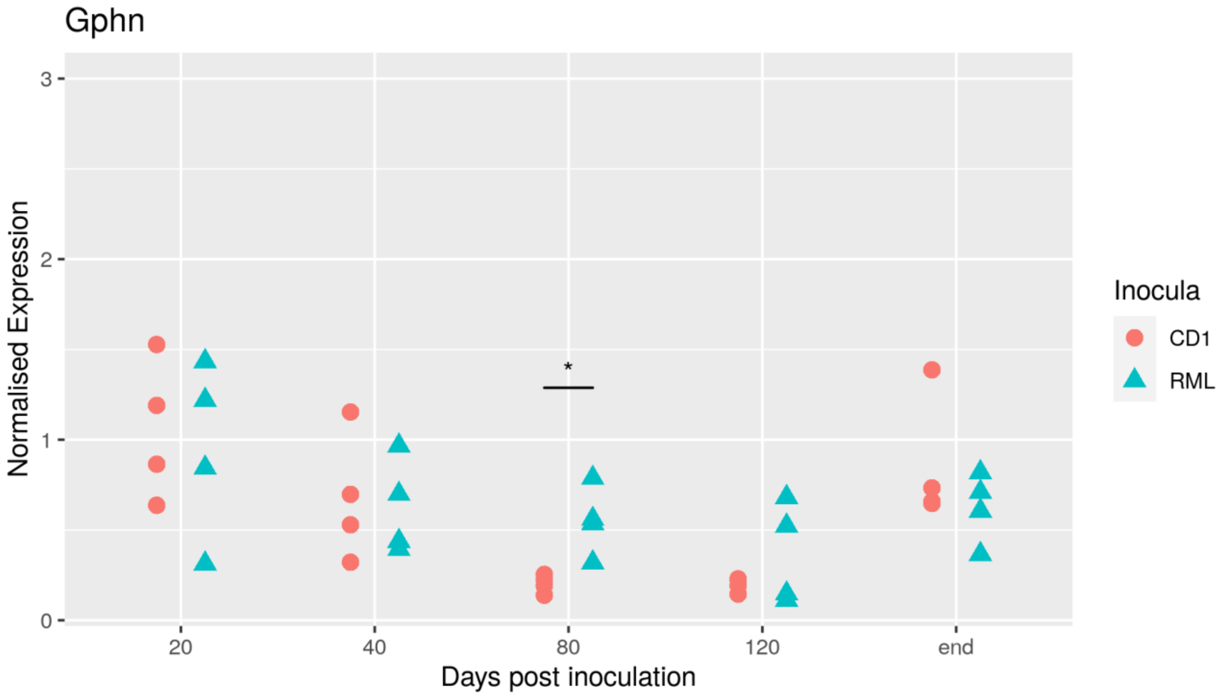
***Supplementary Figure 8: No bias was identified regarding the transcript length and chromosomes of differentially expressed genes and no outlier samples were found to drive differences in gene expression. (a)*** *The histogram shows the distribution of chromosomes that DEGs reside.* ***(b)*** *The histogram shows the transcript lengths of identified DEGs. Transcript length in nucleotides.* ***(c)*** *Representative plots of the expression of reported DEGs between different samples. While single-cell data is sparse and there is variability in the gene expression of some samples, there were no specific sample outliers that drove the results of the differential expression analysis.*

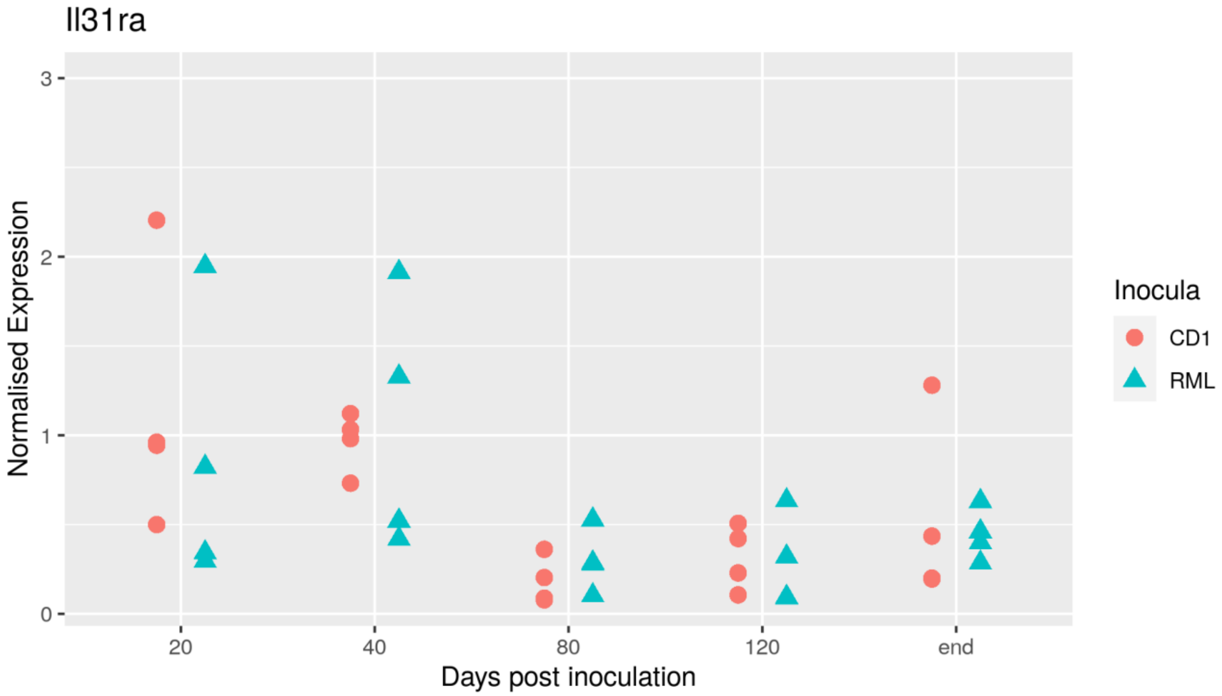


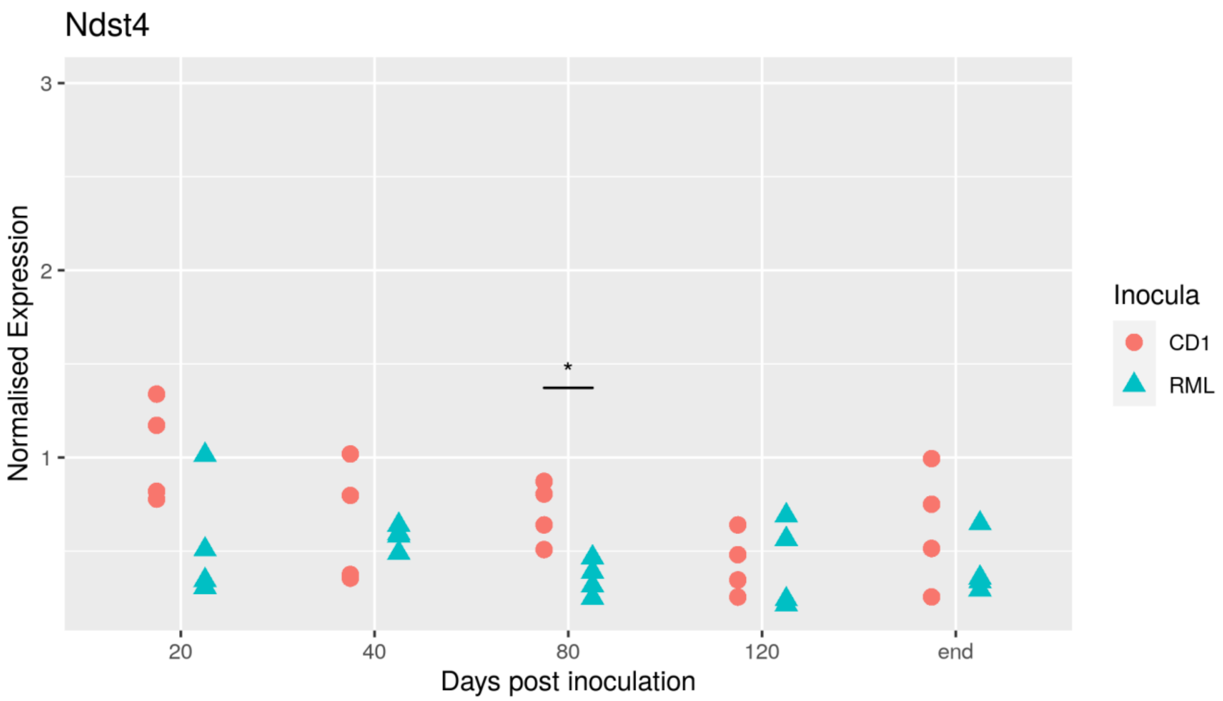
***Supplementary Figure 9: There is a positive correlation between the number of DEGs identified and the numbers of cells in each cluster when using any of the three DE approaches.*** *The three plots show the total number of DE genes identified in each cluster across all time points (y-axis) versus the total number of cells in the same cluster (x-axis) for DE analyses performed using Seurat, DESeq2 or glmGamPoi. The correlation coefficients were calculated to be 0.28 for Seurat, 0.78 for DESeq2 and 0.78 for glmGamPoi. Annotation labels correspond to cluster numbers.*

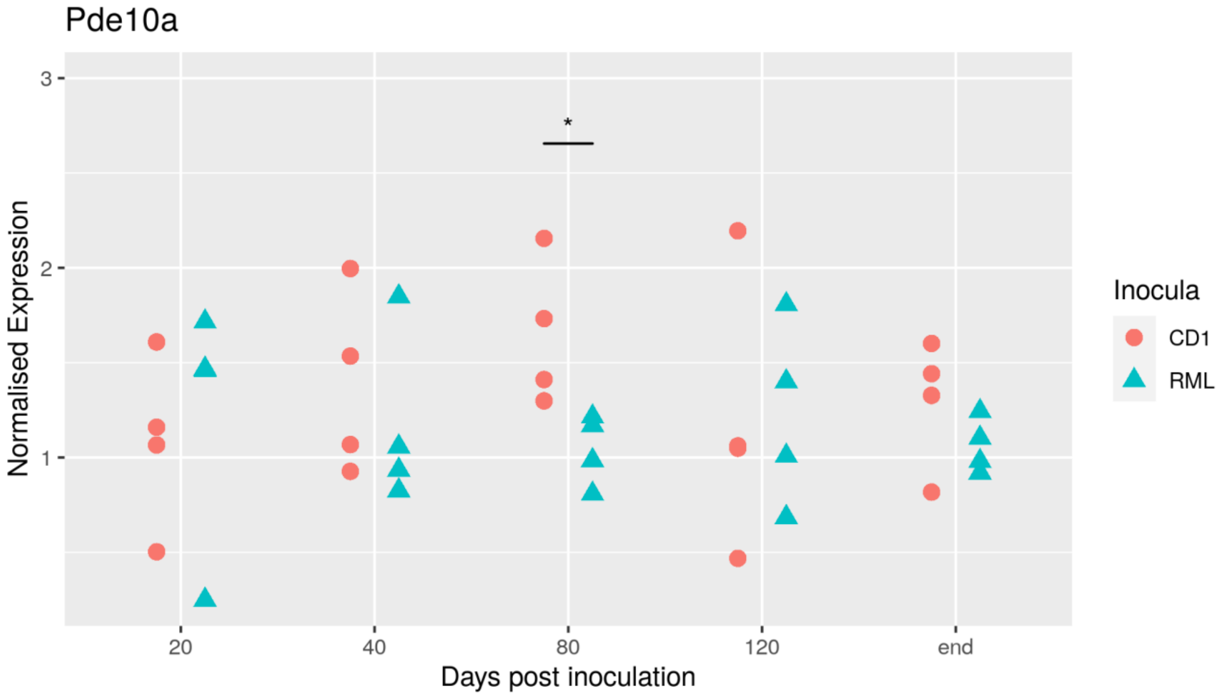

Abi3bp
Normalised Expression
3
2
1
0
20
40
80
120
end
Days post inoculation
*
Inocula
● CD1
▲ RML


Auts2
Normalised Expression
3
2
1
20
40
80
120
end
Days post inoculation
*
Inocula
● CD1
▲ RML

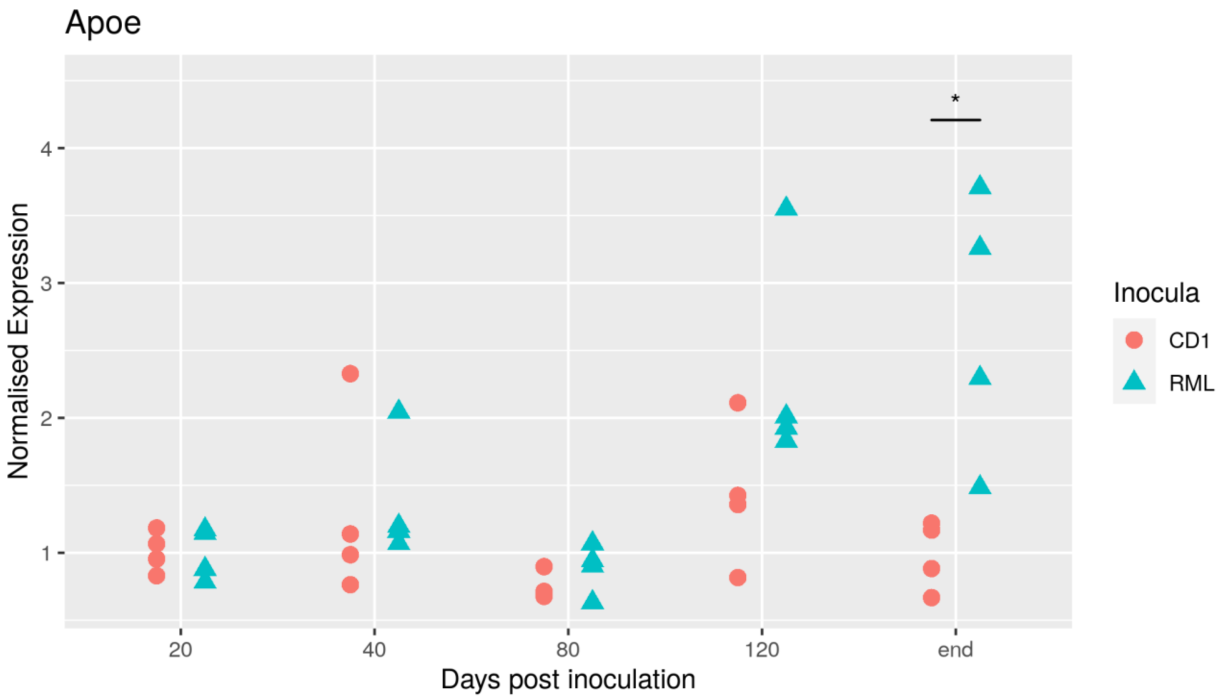

Gphn
Normalised Expression
3
2
1
0
*
20
40
80
120
end
Days post inoculation
Inocula
CD1
RML

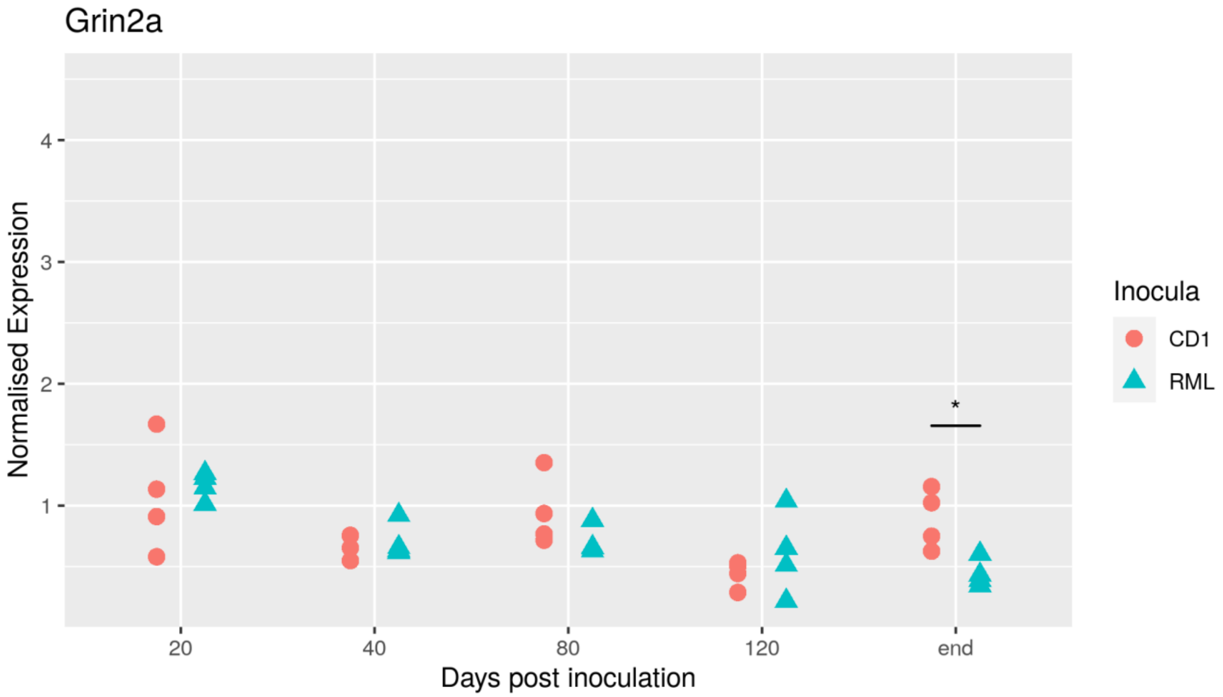

Il31ra
Normalised Expression
3
2
1
0
20
40
80
120
end
Days post inoculation
Inocula
CD1
RML

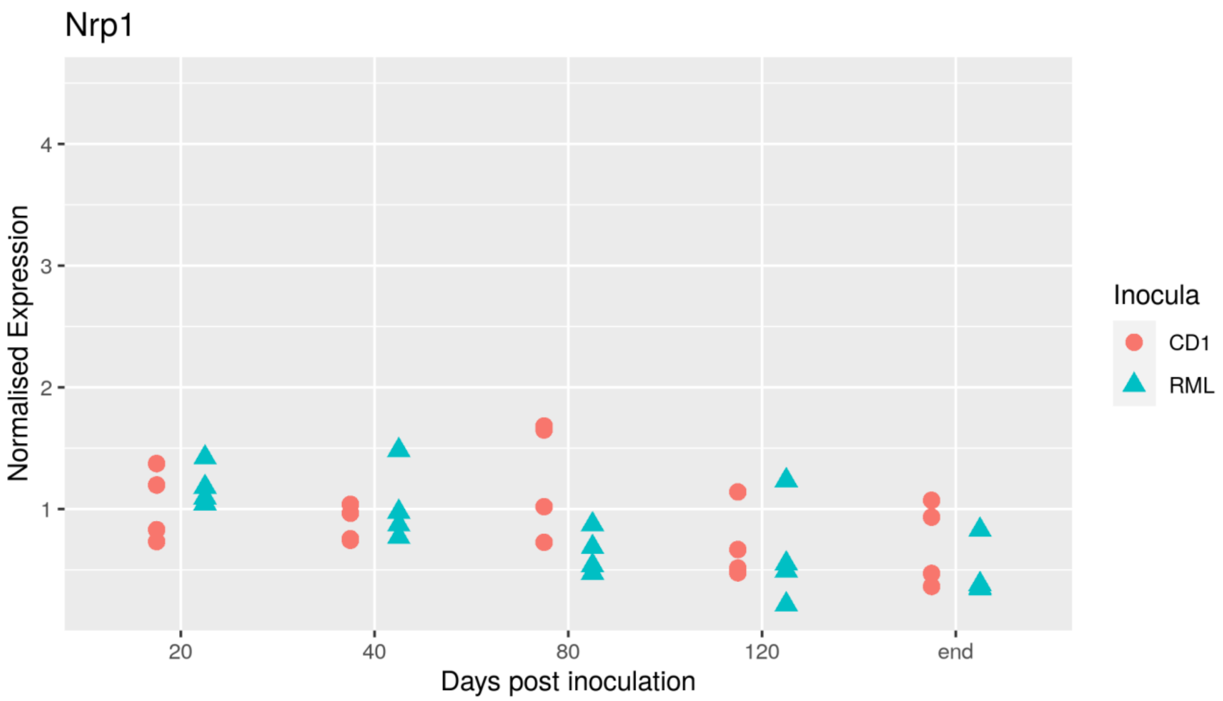


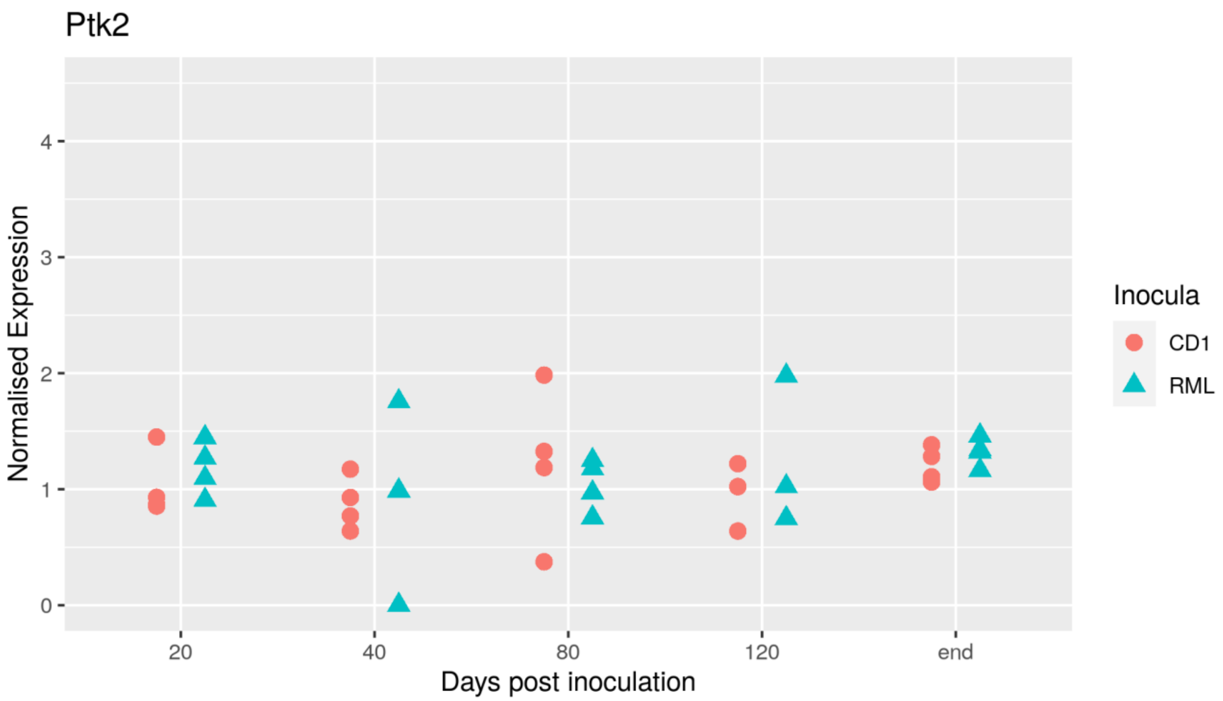


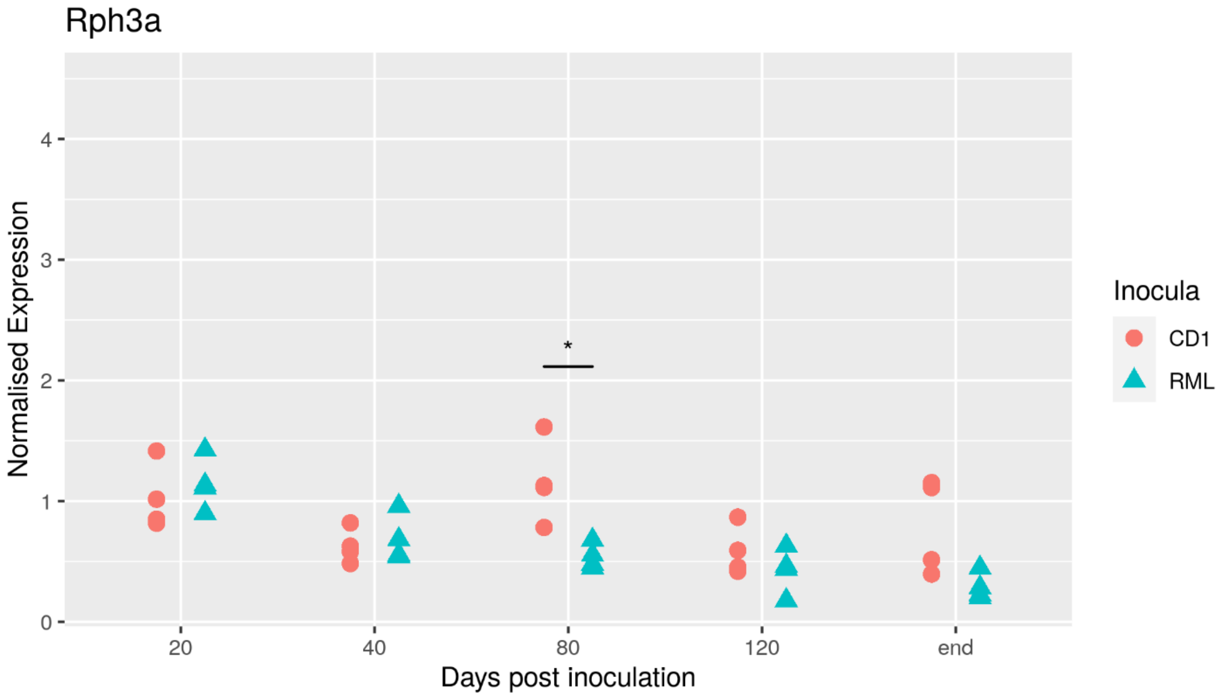
***Supplementary Figure 10: Real-time quantitative PCR expression data for 6 genes that were found to be differentially expressed in early and late time points.*** *Only 2 of the genes (Abi3bp and Auts2) were found to be significantly differentially expressed in the last time point. Statistical test: Wilcoxon rank-sum test. * represents a p-value < 0.05. Each point represents a biologically independent sample. The* *starting material was bulk brain nuclei suspension, so the effect of specific cell populations could have been diluted in the bulk material.*


Apoe

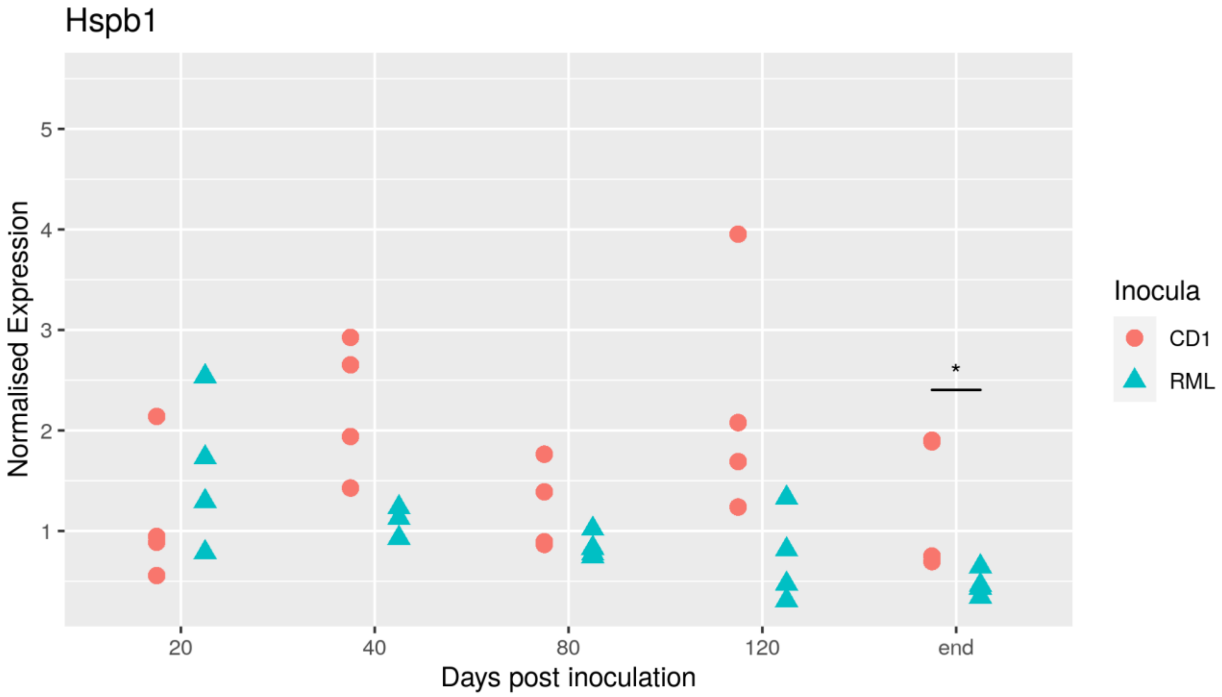

Grin2a

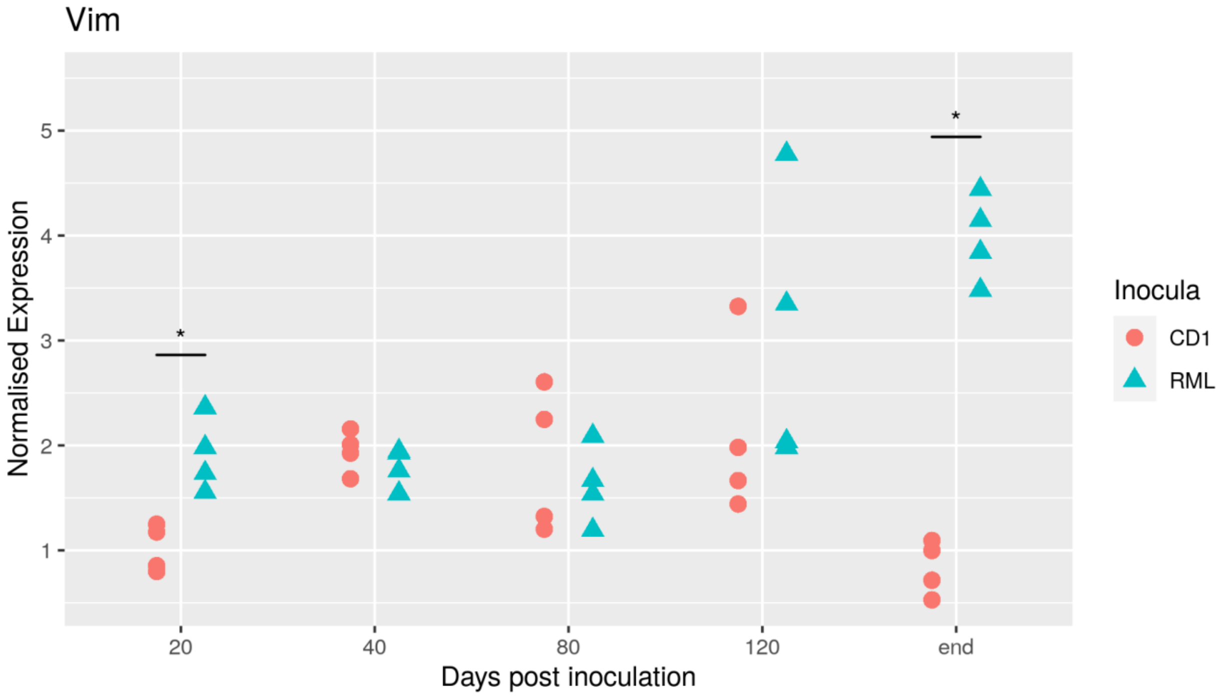

Nrp1
Normalised Expression
4
3
2
1
20
40
80
120
end
Days post inoculation
Inocula
● CD1
▲ RML

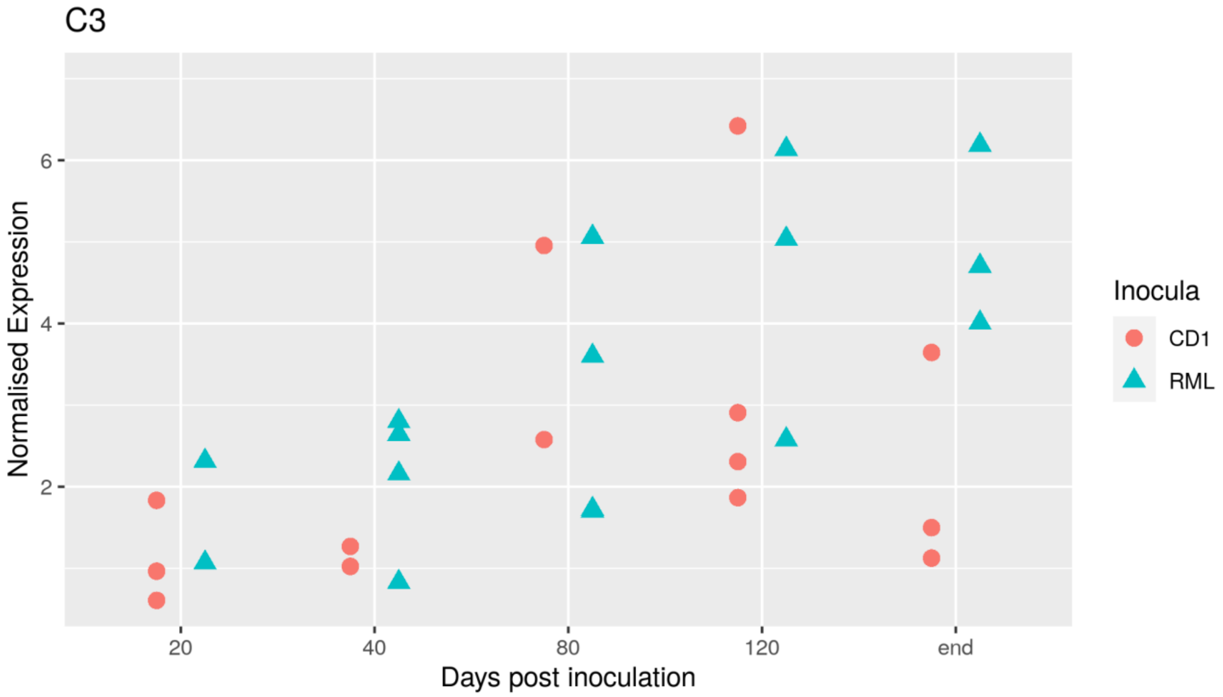

Ptk2
Normalised Expression
4
3
2
1
0
20
40
80
120
end
Days post inoculation
Inocula
● CD1
▲ RML

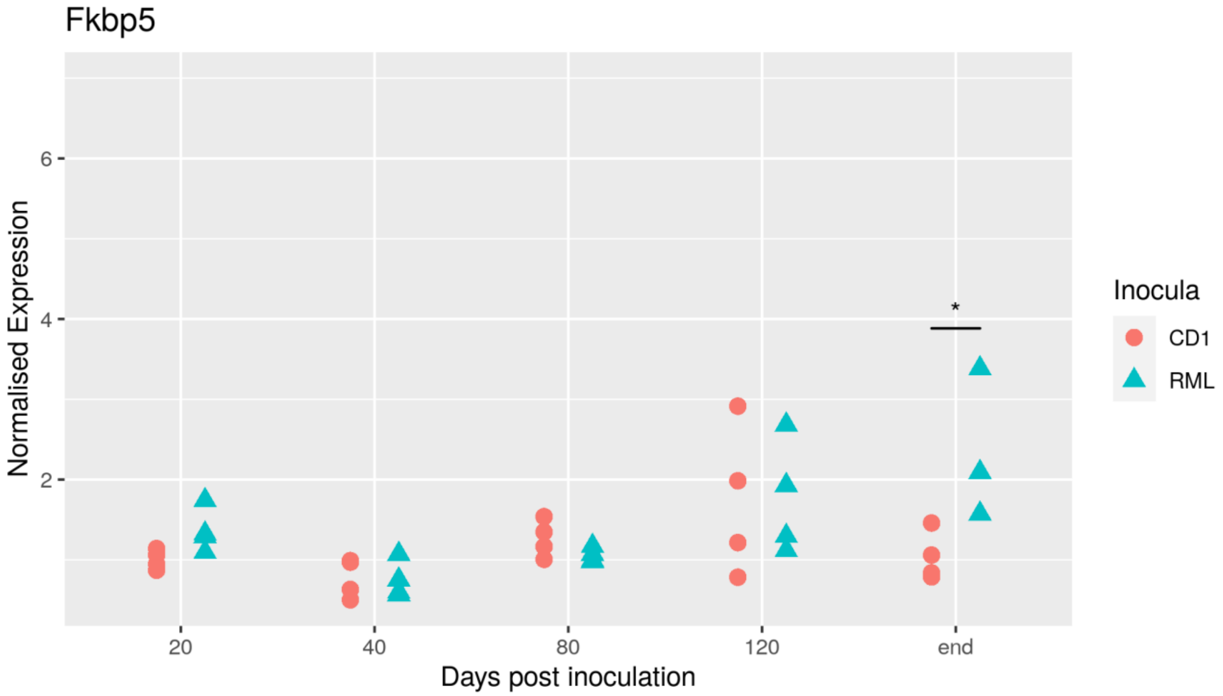


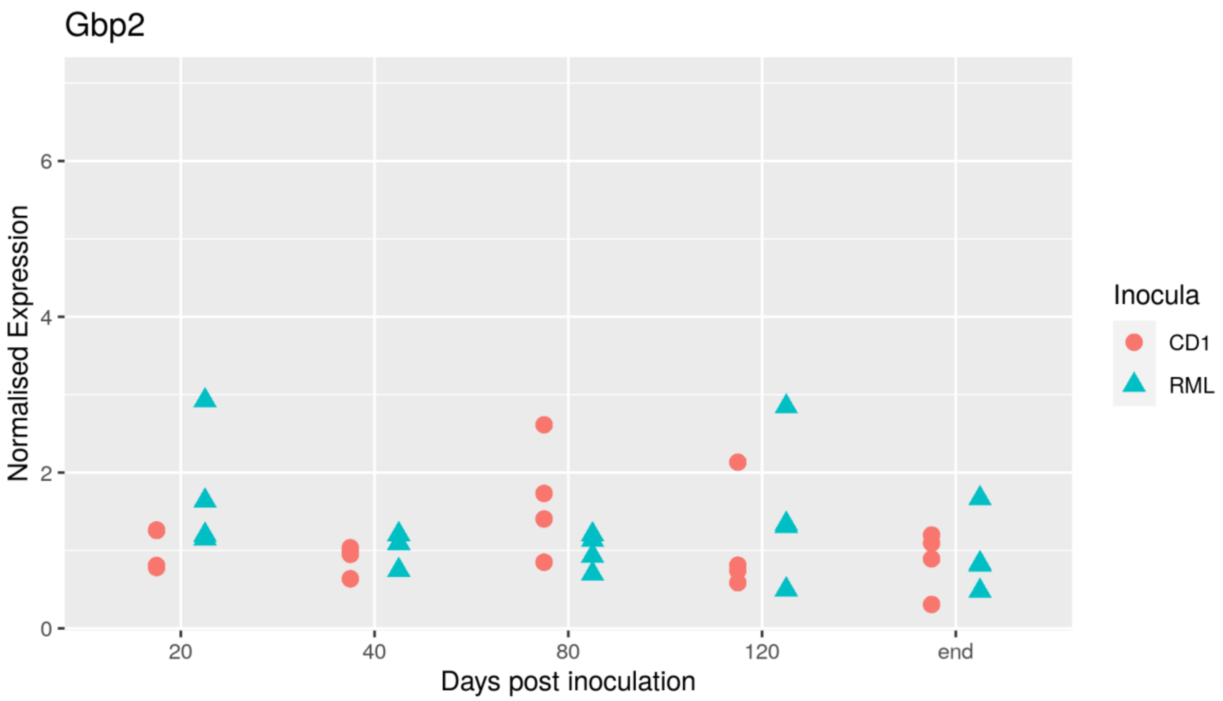
***Supplementary Figure 11: Real-time quantitative PCR expression data for 5 genes that were found to be differentially expressed in late time points and are also part of the synapse organisation Gene Ontology pathway.*** *Only 2 of the genes (Apoe and Grin2a) were found to be significantly differentially expressed in the last time point. Statistical test: Wilcoxon rank-sum test. * represents a p-value < 0.05. Each point represents a biologically independent sample. The starting material was bulk brain nuclei suspension, so the effect of specific cell populations could have been diluted in the bulk material.*

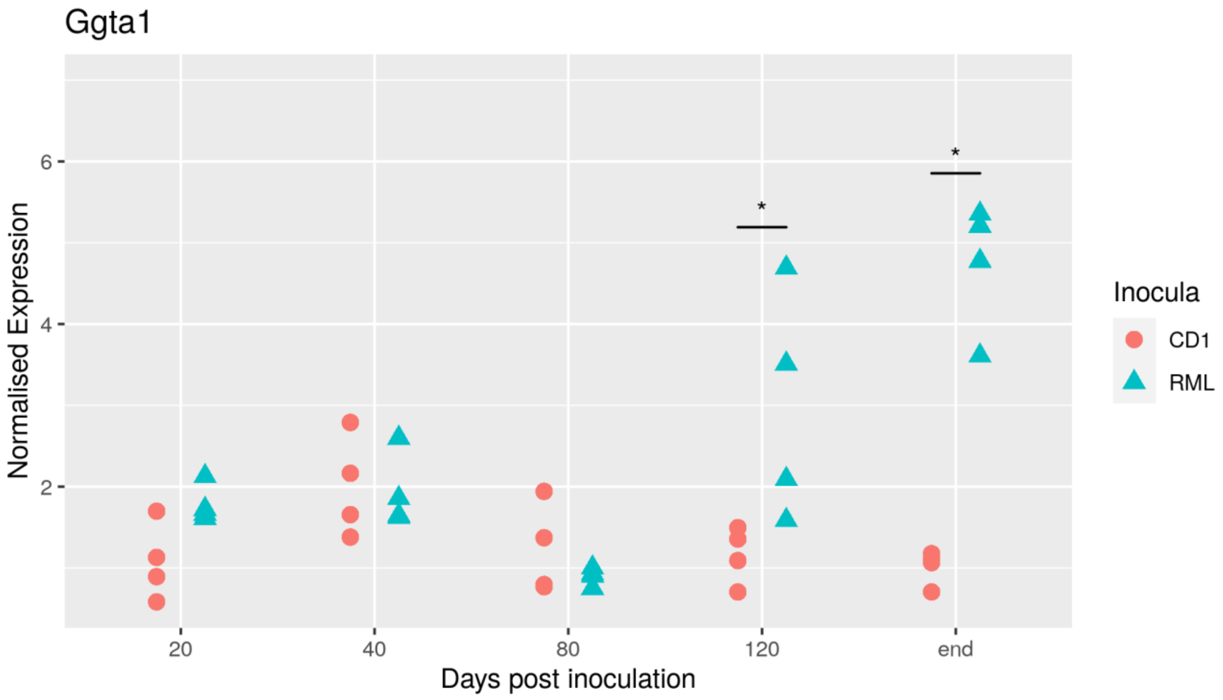

Hspb1
Normalised Expression
5
4
3
2
1
20
40
80
120
end
Days post inoculation
*
Inocula
● CD1
▲ RML

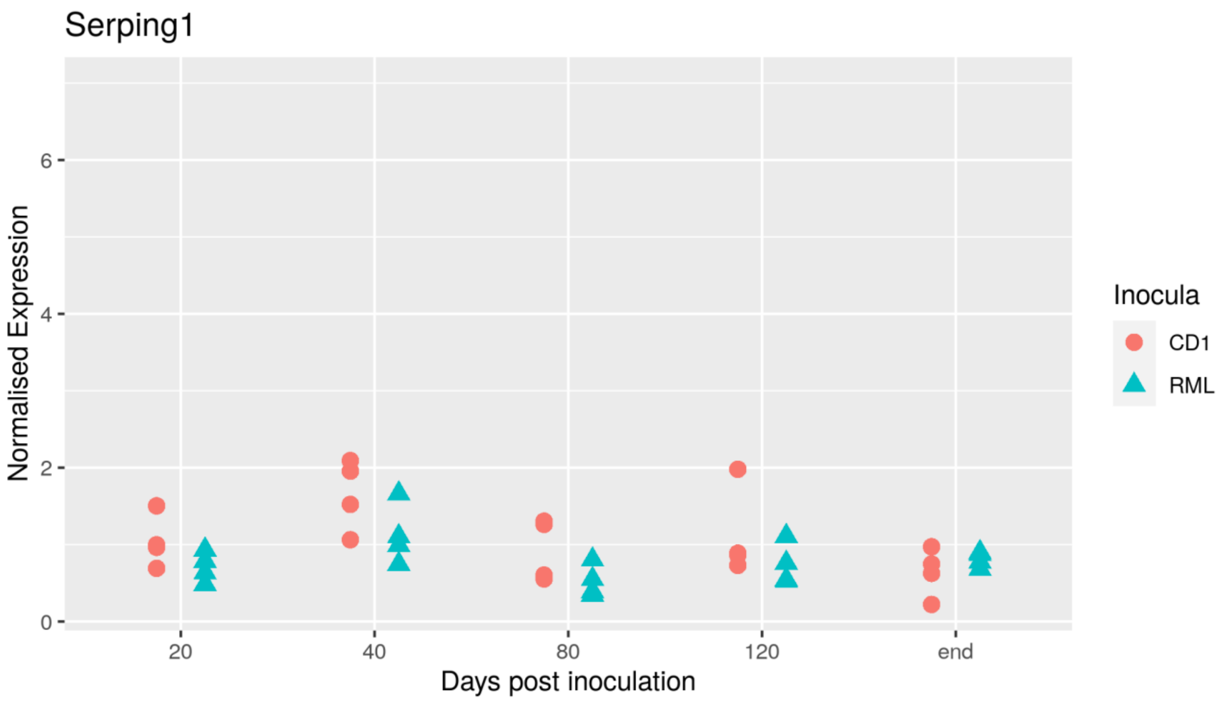

Vim
Normalised Expression
5
4
3
2
1
20
40
80
120
end
Days post inoculation
*
*
Inocula
● CD1
▲ RML

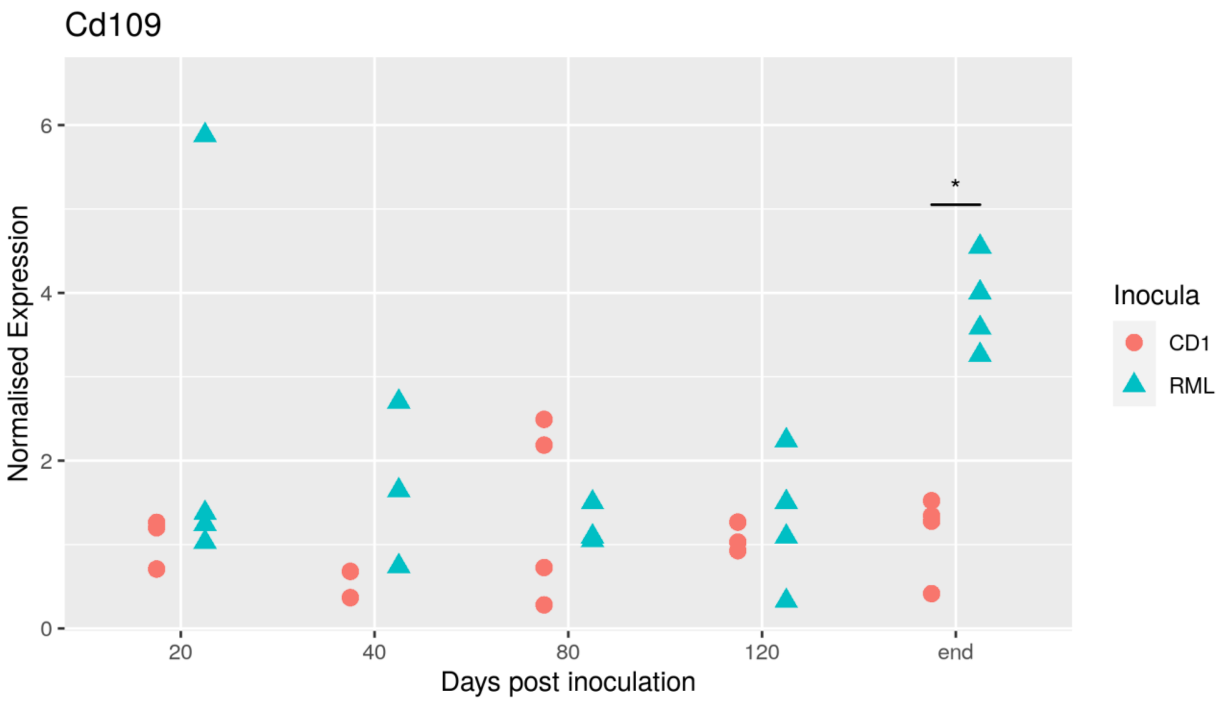

C3
Normalised Expression
6
4
2
20
40
80
120
end
Days post inoculation
Inocula
● CD1
▲ RML

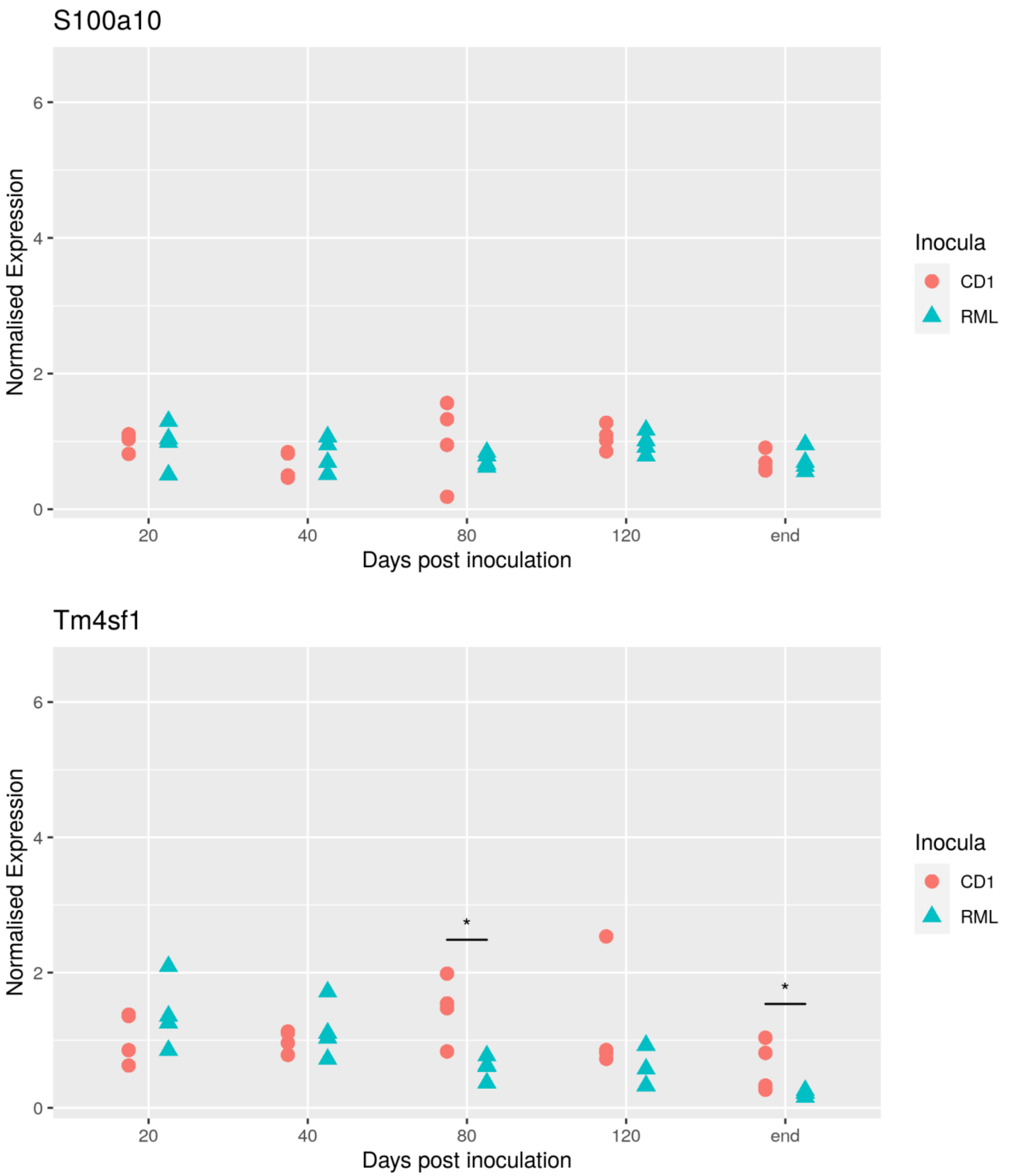

Fkbp5
Normalised Expression
6
4
2
*
20
40
80
120
end
Days post inoculation
Inocula
● CD1
▲ RML


Gbp2
Normalised Expression
0
2
4
6
20
40
80
120
end
Days post inoculation
Inocula
CD1
RML

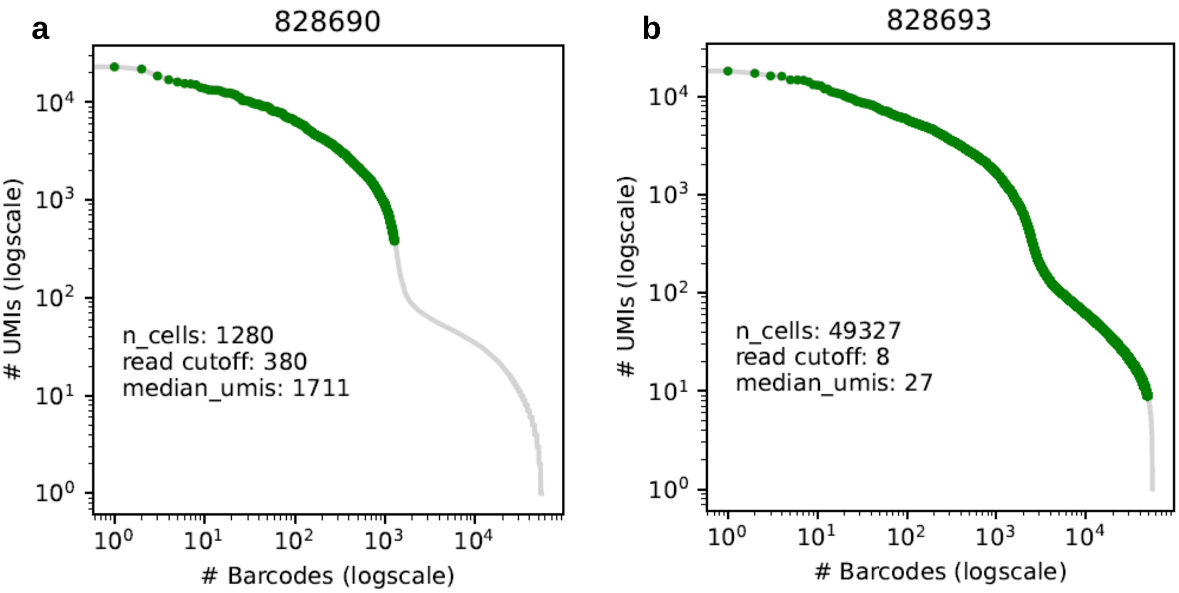

Ggta1
Normalised Expression
2
4
6
*
*
20
40
80
120
end
Days post inoculation
Inocula
CD1
RML


Serping1
Normalised Expression
0
2
4
6
20
40
80
120
end
Days post inoculation
Inocula
● CD1
▲ RML


Cd109
Normalised Expression
0
2
4
6
*
20
40
80
120
end
Days post inoculation
Inocula
● CD1
▲ RML



***Supplementary Figure 12: Real-time quantitative PCR expression data of Pan (Hspb1, Vim), A1 (C3, Fkbp5, Gbp2, Ggta1, Serping1), and A2 (Cd109, S100a10, Tm4sf1) astrocyte signature genes.*** *Both pan-astrocyte signature genes were found to be significantly upregulated in the last time point and Vim was also found to be significantly upregulated at 20 dpi, suggesting astrocyte activation. 2/5 A1 signature genes were significantly upregulated at the last time point, suggesting the existence of A1 astrocytes. Cd109 A2* *signature gene was significantly upregulated at the disease end-stage, while Tm4sf1 was significantly downregulated at 80 dpi and the end-stage. Statistical test: Wilcoxon rank-sum test. * represents a p-value < 0.05. Each point represents a biologically independent sample. The starting material was bulk brain nuclei suspension, so the effect of specific cell populations could have been diluted in the bulk material.*



***Supplementary Figure 13: Representative knee plots from two samples of the 20 dpi time point where the cell identification threshold has been set (a) correctly and (b) incorrectly.*** *The knee plots represent the sorted number of cell barcodes (x-axis) versus the number of UMI counts detected per cell barcode (y-axis). The plots are expected to contain two such knees, and the mid-point of the first knee is usually used as a cut-off to differentiate real cells from the background.* ***(a)*** *The splitseq-tools algorithm has successfully identified the knee of the plot and set a threshold of 380 reads per nucleus.* ***(b)*** *In some cases, when the slope is not pronounced enough, the algorithm can omit the first knee of the plot and identify the second one, substantially inflating the number of identified cells and decreasing the mean number of UMIs per cell. Here the algorithm has selected a threshold of only 8 reads per nucleus.*

## 7.2 Tables

| | PDG | Patient ID | PM No | Experimental group | Cause of death | PMI | PrP type | Codon 129 | Sex | Age | Clinical duration |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 12499 | 11836 | NH04-1692 | Biopsy sCJD | sporadic CJD | <0.5 h | T3 | MV | M | 60 | 4 m |
| 2 | 16602 | 12699 | NH06-0761 | Biopsy sCJD | sporadic CJD | <0.5 h | protease resistant PrP with an unusual fragment pattern | MV | F | 71 | 1 y |
| 3 | 21843 | 13672 | NP11-70 | Biopsy sCJD | sporadic CJD | <0.5 h | T3 | MV | M | 39 | 3 y 2 m |
| 4 | 67459 | NH11-348 | N/A | Biopsy control | N/A | <0.5 h | N/A | N/A | F | 57 | N/A |
| 5 | 67460 | NH15-1686 | N/A | Biopsy control | N/A | <0.5 h | N/A | N/A | M | 65 | N/A |
| 6 | 67461 | NH21-2794 | N/A | Biopsy control | N/A | <0.5 h | N/A | N/A | M | 60 | N/A |
| 7 | 51446 | 22496 | NP15-69 | Post-mortem sCJD | sporadic CJD | 6 d | N/A | MM | M | 64 | 4 y |
| 8 | 52470 | 24948 | NP15-78 | Post-mortem sCJD | sporadic CJD | 2 d | T2 | MM | F | 77 | 3 m |
| 9 | 52695 | 25040 | NP15-84 | Post-mortem sCJD | sporadic CJD | 8 d | T2 | MM | M | 71 | 2 m |
| 10 | 52935 | 25028 | NP15-90 | Post-mortem sCJD | sporadic CJD | 3 d | T3 | MM | F | 57 | 2 m |
| 11 | 53846 | 25824 | NP16-60 | Post-mortem sCJD | sporadic CJD | 10 d | T2 | MM | F | 74 | 3 w |
| 12 | 54580 | 25308 | NP16-19 | Post-mortem sCJD | sporadic CJD | 5 d | T2 | MM | M | 67 | 3 m |
| 13 | 56720 | 25936 | NP17-37 | Post-mortem sCJD | sporadic CJD | 5 d | T2 | MM | F | 68 | 2 m |
| 14 | 54466 | 25301 | NP16-14 | Post-mortem sCJD | sporadic CJD | 10 d | T3 | MM | F | 77 | 1 m |
| 15 | 56943 | 26879 | NP17-46 | Post-mortem sCJD | sporadic CJD | 12 d | T2 | MM | M | 77 | N/A |
| 16 | 56093 | 25835 | NP16-68 | Post-mortem sCJD | sporadic CJD | 4 d | T3 | MM | F | 72 | 1 m |
| 17 | 67094 | P83/10 | N/A | Post-mortem control | Myocardial infarction | 50 h | N/A | N/A | M | 81 | N/A |
| 18 | 67096 | P15/10 | N/A | Post-mortem control | Haemopericardium | 168 h | N/A | N/A | M | 69 | N/A |
| 19 | 67097 | P66/10 | N/A | Post-mortem control | N/A | 57 h | N/A | N/A | M | 87 | N/A |
| 20 | 67098 | P21/17 | N/A | Post-mortem control | Multiorgan failure | 79 h | N/A | N/A | M | 76 | N/A |
| 21 | 67099 | P33/18 | N/A | Post-mortem control | Renal failure | 47 h | N/A | N/A | M | 89 | N/A |
| 22 | 67103 | P58/10 | N/A | Post-mortem control | Aortic stenosis with left ventricular failure | 51.5 h | N/A | N/A | F | 87 | N/A |
| 23 | 67104 | P47/11 | N/A | Post-mortem control | Pancreatic cancer | 79 h | N/A | N/A | F | 79 | N/A |
| 24 | 67105 | P64/11 | N/A | Post-mortem control | Pancreatic cancer | 49 h | N/A | N/A | F | 80 | N/A |
| 25 | 67106 | P66/11 | N/A | Post-mortem control | Heart failure | 120.5 h | N/A | N/A | F | 86 | N/A |
| 26 | 67109 | P56/18 | N/A | Post-mortem control | Bronchopneumonia | 26 h | N/A | N/A | F | 103 | N/A |

|  | First symptoms | Progression | Examination signs |
|---|---|---|---|
| 1 | social withdrawal, mood disorders, personality change | ataxia, dysphasia, swallowing difficulty | apraxia and ataxia. Marked frontal release sign with positive glabellar tap and bilateral grasp reflexes. |
| 2 | confusion | became withdrawn and quiet, memory impairment, deterioration of confusion, mood alterations | cognitive impairment, myoclonus |
| 3 | weight loss, misplacing things, confusion | disorientation, headache, myoclonus, stiff limbs, weight loss, incontinence | cognitive impairment, visual deficit |
| 4 | N/A | N/A | N/A |
| 5 | N/A | N/A | N/A |
| 6 | N/A | N/A | N/A |
| 7 | difficulty answering the telephone, could not swim, could not calculate | unsteady, speech decline, dyspnoea, falls, low output and slurring speech, poor sleeping, inability to walk, rigidity | cognitive impairment, upper limb paratonia, apraxic gait |
| 8 | dizziness, hazy vision | gait disturbance, visual disturbance, speech decline, | akinetic mutism, no primiive reflexes in the cranial nerves, no grasp reflex in upper limbs, no myoclonus, mild spasticity on the left lower limb |
| 9 | right sided ataxia | progressive gait disturbance, reduced power, coordination disturbance, slurred speech, visual disturbances, dystonia | N/A |
| 10 | tiredness, insomnia, memory issues | distortion of vision, confusion, unsteadiness, became quiet, slurred speech, sleepy | Family did not want examination of the patient |
| 11 | confusion, inability to communicate and complete daily living tasks | N/A | N/A |
| 12 | struggled with reading | deterioration of vision, visual hallucinations, limb rigidity, diminished spontaneous speech, disintergration of episodic memory, myoclonus | akinetic mutism, intermittent rhythmic myoclonus |
| 13 | stuttering, diplopia, ataxic gait, confusion | drowsyness, unsteadiness, agitation, slurred speech, repetitive speech | perseverative behaviour, homonymous hemianopia, mild ataxia on the right, myoclonic jerks in the left arm |
| 14 | confusion, dysarthria, dysphagia, difficulty swallowing | less resposive to questions, more confused, less aware of the surroundings, distressed | N/A |
| 15 | N/A | N/A | N/A |
| 16 | unilateral arm weakness and tremor, startle myoclonus | dysarthria, dysphagia, myoclonus, immobility, dementia | N/A |
| 17 | N/A | N/A | N/A |
| 18 | N/A | N/A | N/A |
| 19 | N/A | N/A | N/A |
| 20 | N/A | N/A | N/A |
| 21 | N/A | N/A | N/A |
| 22 | N/A | N/A | N/A |
| 23 | N/A | N/A | N/A |
| 24 | N/A | N/A | N/A |
| 25 | N/A | N/A | N/A |
| 26 | N/A | N/A | N/A |

| | Investigations | | | |
|---|---|---|---|---|
| | MRI scan | EEG | Cerebrospinal fluid analysis | Prion gene analysis |
| 1 | Patient not tolerant | abnormal with evidence of diffuse cortical dysfunction and occasional sharp components over the left posterior area. No definite epileptiform or periodic complexes were seen. | positive protein 14-3-3 and S100b | no mutations |
| 2 | N/A | N/A | N/A | no mutations |
| 3 | restricted diffusion affecting the cortex of the right hemisphere and the left occipital lobe | background slow activity with sharpened waves, maximal over the left side | positive protein 14-3-3, normal S100b | no mutations |
| 4 | N/A | N/A | N/A | N/A |
| 5 | N/A | N/A | N/A | N/A |
| 6 | N/A | N/A | N/A | N/A |
| 7 | high signal in the caudate and putamen | N/A | 14-3-3 protein not present, s100b elevated | no mutations |
| 8 | first MRI scan showed atrophy of the hippocampi, second showed a considerable amount of vascular disease, hippocampal atrophy, third MRI showed restricted diffusion in the caudate and putamen, cortical ribboning | N/A | N/A | no mutations |
| 9 | showed no cause for symptoms | N/A | positive for 14-3-3, S100b elevated. RT-QuIC positive | no mutations |
| 10 | restricted diffusion in the left caudate nucleus and anterior putamen, cortical ribboning most marked posteriorly involving the the anterior part of the occipital lobe and the precuneus | N/A | acellular with normal protein | no mutations |
| 11 | inconclusive due to movement | showed some abnormalities | normal | no mutations |
| 12 | did not include diffusion weighted imaging, no high signal in the basal ganglia or the cortex. Second MRI showed classical features of sCJD with restriction of diffusion in the basal ganglia and cortical ribboning | generalised periodic complexes | N/A | no mutations |
| 13 | extensive cortical ribboning maximal over the right hemisphere. Restricted diffusion in the right caudate nucleus | periodic complexes which have worsened | CSF protein normal and no cells present | no mutations |
| 14 | no conclusive diagnosis | N/A | N/A | no mutations |
| 15 | N/A | N/A | N/A | no mutations, G124G polymorphism |
| 16 | cortical ribboning and pulvinar sign | N/A | positive for 14-3-3 and S100b. RT-QuIC positive | no mutations |
| 17 | N/A | N/A | N/A | N/A |
| 18 | N/A | N/A | N/A | N/A |
| 19 | N/A | N/A | N/A | N/A |
| 20 | N/A | N/A | N/A | N/A |
| 21 | N/A | N/A | N/A | N/A |
| 22 | N/A | N/A | N/A | N/A |
| 23 | N/A | N/A | N/A | N/A |
| 24 | N/A | N/A | N/A | N/A |
| 25 | N/A | N/A | N/A | N/A |
| 26 | N/A | N/A | N/A | N/A |

|  | Non-contributory neuropathological findings |
|---|---|
|  |  |
| 1 | N/A |
| 2 | N/A |
| 3 | Strong and widespread tau deposition in the cortex and very minimal tau deposition in the cerebellum. No amyloid beta detected in any of the cortical areas. |
| 4 | N/A |
| 5 | N/A |
| 6 | N/A |
| 7 | Thal phase 0, CERAD score 0, B&B NFT stage 1, ABC score: A0, B1, C0: no evidence of AD neuropathological change. Mild hyaline arteriolosclerosis. Mild TDP43 proteinopathy of uncertain significance. |
| 8 | Argyrophilic grain disease, corticobasal degeneration. B&B stage 2. ABC score: A0, B1, C0: no evidence of AD neuropathological change. Hyaline arteriolosclerosis. |
| 9 | Thal phase 1, CERAD score 0. ABC score: A1, B1, C0: low level of AD neuropathological change. Mild hyaline arteriolosclerosis. Focal cerebral amyloid angiopathy. |
| 10 | Thal phase 2, CERAD score 0, Tau B&B NFT stage 0. ABC score: A1, B0, C0: low level of AD neuropathological change. |
| 11 | Thal phase 0, CERAD score 0, B&B NFT stage 1, ABC score: A0, B1, C0: no evidence of AD neuropathological change. Mild hyaline arteriolosclerosis. |
| 12 | Thal phase 4, CERAD score 2, B&B NFT stage 2, ABC score: A3, B1, C2: low level AD neuropathological change. Cerebral amyloid angiopathy, Vonsattel grade 2. |
| 13 | Thal phase 3, CERAD score 3, B&B NFT stage 1, ABC score: A2, B1, C2: low level AD neuropathological changes. Cerebral amyloid angiopathy, Vonsattel grade 2. Hyaline arteriolosclerosis. |
| 14 | Thal Abeta: 4, CERAD score 3, B&B NFT stage 2, ABC score: A3, B1, C3: low level of AD neuropathological change. Cerebral amyloid angiopathy, alpha-synuclein Lewy body pathology, chronic micro-infarcts in basal gangllia, hyaline arteriolosclerosis |
| 15 | N/A |
| 16 | Thal phase 1, CERAD score 0, B&B NFT stage 2. ABC score: A1, B1, C0: low level of AD neuropathological change. Focal Lewy body pathology. Focal leptomeningeal cerebral amyloid angiopathy, Vonsattel grade 2. |
| 17 | Thal Abeta: 1; B&B NFT: 2; CERAD Abeta: Absent; Braak Lewy path: 4 |
| 18 | Thal Abeta: 3; B&B NFT: 1; CERAD Abeta: Sparse; Braak Lewy path: 0 |
| 19 | Thal Abeta: 3; B&B NFT: 2; CERAD Abeta: Moderate; Braak Lewy path: 0 |
| 20 | Thal Abeta: 1; B&B NFT: 2; CERAD Abeta: Absent; Braak Lewy path: 0 |
| 21 | Thal Abeta: 3; B&B NFT: 2; CERAD Abeta: Sparse; Braak Lewy path: 0 |
| 22 | Thal Abeta: 1; B&B NFT: 1; CERAD Abeta: Sparse; Braak Lewy path: 0 |
| 23 | Thal Abeta: 2+; B&B NFT: 1; CERAD Abeta: Sparse; Braak Lewy path: 0 |
| 24 | Thal Abeta: 0; B&B NFT: 2; CERAD Abeta: Absent; Braak Lewy path: 0 |
| 25 | Thal Abeta: 0; B&B NFT: 2; CERAD Abeta: Absent; Braak Lewy path: 0 |
| 26 | Thal Abeta: 5; B&B NFT: 4; CERAD Abeta: Sparse; Braak Lewy path: 0 |

***Supplementary Table 1: Tables of demographic and clinicopathological information of patients included in the human transcriptomics study.*** *PDG: sample identifier, Patient ID: patient identifier, PM No: post-mortem examination identifier, PMI: post-mortem interval between death and sample archiving, PrP type: the PrP type (London classification) evaluated using Western Blotting of proteinase-K-digested frontal cortex brain sample, Codon 129: aminoacid sequence of the PrP protein at codon 129 (MM: methionine homozygous, MV: methionine/Valine heterozygous, VV: Valine homozygous), Clinical duration: disease duration from the first symptoms until death (y: year(s), m: month(s)), N/A: not available/not applicable.*

| Time point | Group | TCIU | Log (TCIU) |
|---|---|---|---|
| 20 dpi | CD1 | N/A | N/A |
| 20 dpi | CD1 | N/A | N/A |
| 20 dpi | CD1 | N/A | N/A |
| 20 dpi | RML | N/A | N/A |
| 20 dpi | RML | 21423.51435 | 4.330890715 |
| 20 dpi | RML | N/A | N/A |
| 40 dpi | CD1 | N/A | N/A |
| 40 dpi | CD1 | N/A | N/A |
| 40 dpi | CD1 | N/A | N/A |
| 40 dpi | RML | N/A | N/A |
| 40 dpi | RML | 8028564.047 | 6.904637876 |
| 40 dpi | RML | 16991086.98 | 7.230221163 |
| 80 dpi | CD1 | N/A | N/A |
| 80 dpi | CD1 | N/A | N/A |
| 80 dpi | CD1 | N/A | N/A |
| 80 dpi | RML | 60898895.24 | 7.784609414 |
| 80 dpi | RML | 61275399.84 | 7.787286154 |
| 80 dpi | RML | 398107170.6 | 8.6 |
| 120 dpi | CD1 | N/A | N/A |
| 120 dpi | CD1 | N/A | N/A |
| 120 dpi | CD1 | N/A | N/A |
| 120 dpi | RML | 68058767.02 | 7.832884077 |
| 120 dpi | RML | 398107170.6 | 8.6 |
| 120 dpi | RML | 501187233.6 | 8.7 |
| end-stage | CD1 | N/A | N/A |
| end-stage | CD1 | N/A | N/A |
| end-stage | CD1 | N/A | N/A |
| end-stage | RML | 51708664.68 | 7.713563323 |
| end-stage | RML | 46211264.52 | 7.664747853 |
| end-stage | RML | 79421324.62 | 7.899937126 |

***Supplementary Table 2: Infectivity values of mouse brain homogenates from the mouse transcriptomics study.*** *Values were calculated using the scrapie cell assay and reported after the 3rd cell* *passage. TCIU: tissue culture infectious units, Log (TCIU): base-10 logarithm of the tissue culture infectious units. N/A values represent zero infectivity.*

| Time point | Animal number | Group | Number of cells | Time point | Animal number | Group | Number of cells |
|---|---|---|---|---|---|---|---|
| 20dpi | 828690 | PBS | 1153 | 80dpi | 829369 | RML | 1630 |
| 20dpi | 828692 | PBS | 2021 | 80dpi | 829370 | RML | 2959 |
| 20dpi | 828693 | PBS | 2201 | 80dpi | 829371 | RML | 2267 |
| 20dpi | 828695 | CD1 | 3448 | 80dpi | 829373 | CD1 | 1792 |
| 20dpi | 828696 | CD1 | 4948 | 80dpi | 829375 | CD1 | 1926 |
| 20dpi | 828698 | CD1 | 2357 | 80dpi | 829380 | RML | 1864 |
| 20dpi | 828699 | CD1 | 2710 | 80dpi | 829381 | RML | 1848 |
| 20dpi | 828700 | CD1 | 1984 | 80dpi | 829382 | RML | 1339 |
| 20dpi | 828701 | CD1 | 1631 | 80dpi | 829387 | RML | 1283 |
| 20dpi | 828702 | CD1 | 1374 | 120dpi | 829389 | PBS | 1207 |
| 20dpi | 828703 | CD1 | 3316 | 120dpi | 829390 | PBS | 1512 |
| 20dpi | 828709 | RML | 4136 | 120dpi | 829392 | PBS | 1508 |
| 20dpi | 828710 | RML | 2104 | 120dpi | 829395 | CD1 | 1307 |
| 20dpi | 828712 | RML | 2502 | 120dpi | 829396 | CD1 | 1997 |
| 20dpi | 828713 | RML | 1970 | 120dpi | 829398 | CD1 | 1964 |
| 20dpi | 828715 | RML | 1834 | 120dpi | 829399 | CD1 | 2112 |
| 20dpi | 828717 | RML | 2057 | 120dpi | 829400 | CD1 | 2244 |
| 20dpi | 828718 | RML | 944 | 120dpi | 829401 | CD1 | 1342 |
| 20dpi | 828719 | RML | 61 | 120dpi | 829402 | CD1 | 1305 |
| 40dpi | 828725 | PBS | 1389 | 120dpi | 829407 | CD1 | 1760 |
| 40dpi | 828727 | PBS | 1165 | 120dpi | 829408 | RML | 2306 |
| 40dpi | 828728 | PBS | 1390 | 120dpi | 829409 | RML | 1681 |
| 40dpi | 828732 | CD1 | 2554 | 120dpi | 829410 | RML | 2188 |
| 40dpi | 828735 | CD1 | 2551 | 120dpi | 829411 | RML | 2352 |
| 40dpi | 828737 | CD1 | 1974 | 120dpi | 829414 | RML | 3390 |
| 40dpi | 828738 | CD1 | 2325 | 120dpi | 829415 | RML | 1732 |
| 40dpi | 828740 | CD1 | 2418 | 120dpi | 829417 | RML | 1751 |
| 40dpi | 828741 | CD1 | 1495 | 120dpi | 829420 | RML | 1438 |
| 40dpi | 828742 | CD1 | 1738 | end | 829972 | PBS | 2262 |
| 40dpi | 828743 | CD1 | 1446 | end | 829973 | PBS | 1888 |
| 40dpi | 828746 | RML | 1915 | end | 829974 | PBS | 1783 |
| 40dpi | 828747 | RML | 2875 | end | 829980 | CD1 | 3245 |
| 40dpi | 828748 | RML | 2016 | end | 829981 | CD1 | 2902 |
| 40dpi | 828750 | RML | 1883 | end | 829982 | CD1 | 1929 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 40dpi | 828751 | RML | 1476 | end | 829985 | CD1 | 3361 |
| 40dpi | 828752 | RML | 1429 | end | 829986 | CD1 | 2978 |
| 40dpi | 828756 | RML | 1283 | end | 829987 | CD1 | 1938 |
| 40dpi | 828757 | RML | 2517 | end | 829988 | CD1 | 2338 |
| 80dpi | 829354 | PBS | 1915 | end | 829989 | CD1 | 2225 |
| 80dpi | 829355 | PBS | 1316 | end | 829993 | RML | 3360 |
| 80dpi | 829356 | PBS | 1763 | end | 829994 | RML | 3698 |
| 80dpi | 829359 | CD1 | 2015 | end | 829995 | RML | 2575 |
| 80dpi | 829360 | CD1 | 2228 | end | 829996 | RML | 1867 |
| 80dpi | 829362 | CD1 | 2674 | end | 829997 | RML | 2247 |
| 80dpi | 829364 | CD1 | 2376 | end | 829998 | RML | 1086 |
| 80dpi | 829365 | CD1 | 2154 | end | 829999 | RML | 4537 |
| 80dpi | 829367 | CD1 | 1381 | end | 830003 | RML | 2637 |
| 80dpi | 829368 | RML | 2887 |  |  |  |  |

***Supplementary Table 3: Numbers of cells identified from each biological sample.***

|  | Cluster name | Counts mean | Counts median | Counts SD | Features mean | Features median | Features SD |
|---|---|---|---|---|---|---|---|
| **20 dpi** | 61 OPC | 809 | 555 | 665 | 564 | 432 | 349 |
|  | 68 Astro Slc7a10 | 805 | 563 | 687 | 560 | 427 | 368 |
|  | 69 Astro Prdm16 | 932 | 605 | 856 | 626 | 459 | 437 |
|  | 46 Migrating Int Cpa6 | 1106 | 685 | 1060 | 672 | 488 | 475 |
|  | 48 Migrating Int Pbx3 | 1115 | 687 | 1113 | 687 | 509 | 495 |
|  | 66 VLMC Slc6a13 | 1147 | 687 | 1060 | 740 | 536 | 519 |
|  | 49 Migrating Int Lgr6 | 1165 | 732 | 1082 | 717 | 534 | 488 |
|  | 57 Oligo MOL | 1325 | 793 | 1236 | 800 | 569.5 | 575 |
|  | 47 Migrating Int Foxp2 | 1148 | 854 | 935 | 709 | 575 | 425 |
|  | 64 Endothelia | 1518 | 1122 | 1201 | 946 | 781 | 596 |
|  | 13 CTX PyrL5 Fezf2 | 1554 | 1230 | 1111 | 906 | 791 | 471 |
|  | 4 Medium Spiny Neurons | 1683 | 1301 | 1243 | 941 | 811 | 524 |
|  | 50 Migrating Int Adarb2 | 1873 | 1484 | 1228 | 1007 | 897.5 | 503 |
|  | 10 CTX PyrL4 Rorb | 1852 | 1546 | 1174 | 1015 | 927 | 476 |
|  | 17 CTX PyrL6 | 2033 | 1633 | 1354 | 1075 | 962 | 534 |
|  | 44 Migrating Int Lhx6 | 2065 | 1715 | 1344 | 1119 | 1019.5 | 553 |
|  | 11 CTX PyrL4/L5 | 2169 | 1815 | 1444 | 1120 | 1034 | 557 |
|  | 14 CTX PyrL6a | 2217 | 1911.5 | 1409 | 1141 | 1072.5 | 541 |
|  | 9 CTX PyrL2/L3/L4 Mef2c | 2363 | 2051 | 1523 | 1195 | 1137 | 575 |
|  | 7 CTX PyrL2/L3 Met | 2443 | 2275 | 1237 | 1241 | 1212.5 | 461 |
|  | 15 CTX PyrL5/L6 Sulf1 | 2584 | 2292 | 1463 | 1278 | 1205 | 548 |
|  | 18 CLAU Pyr | 2667 | 2525 | 1380 | 1308 | 1308 | 493 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 12 CTX PyrL5 Itgb3 | 2971 | 2805.5 | 1549 | 1386 | 1361 | 521 |
| | | | | | | | |
| 40 dpi | 68 Astro Slc7a10 | 869 | 546 | 758 | 577 | 403 | 389 |
| | 61 OPC | 1037 | 689 | 926 | 670 | 496 | 455 |
| | 69 Astro Prdm16 | 1073 | 714 | 908 | 693 | 518 | 456 |
| | 46 Migrating Int Cpa6 | 1397 | 819 | 1337 | 786 | 557 | 563 |
| | 49 Migrating Int Lgr6 | 1319 | 870.5 | 1107 | 788 | 610 | 513 |
| | 47 Migrating Int Foxp2 | 1323 | 897 | 1102 | 761 | 591 | 467 |
| | 72 Ependyma | 1383 | 920 | 1131 | 881 | 651 | 578 |
| | 66 VLMC Slc6a13 | 1507 | 944.5 | 1341 | 900 | 678.5 | 608 |
| | 48 Migrating Int Pbx3 | 1414 | 945 | 1140 | 826 | 648 | 508 |
| | 57 Oligo MOL | 1673 | 1146 | 1370 | 972 | 811 | 608 |
| | 10 CTX PyrL4 Rorb | 1664 | 1315 | 1180 | 911 | 792 | 479 |
| | 4 Medium Spiny Neurons | 1928 | 1513 | 1385 | 1019 | 889 | 552 |
| | 13 CTX PyrL5 Fezf2 | 1869 | 1581 | 1135 | 1024 | 946 | 465 |
| | 64 Endothelia | 1883 | 1607 | 1291 | 1130 | 1069.5 | 624 |
| | 56 Oligo MFOL1 | 2089 | 1661.5 | 1420 | 1132 | 1047.5 | 601 |
| | 50 Migrating Int Adarb2 | 2132 | 1716 | 1447 | 1060 | 937 | 538 |
| | 17 CTX PyrL6 | 2103 | 1741 | 1429 | 1081 | 978.5 | 552 |
| | 9 CTX PyrL2/L3/L4 Mef2c | 2244 | 1824 | 1600 | 1123 | 1036.5 | 586 |
| | 44 Migrating Int Lhx6 | 2204 | 1852.5 | 1427 | 1140 | 1052.5 | 557 |
| | 11 CTX PyrL4/L5 | 2300 | 1946.5 | 1551 | 1151 | 1078.5 | 573 |
| | 7 CTX PyrL2/L3 Met | 2465 | 2185 | 1441 | 1209 | 1167 | 515 |
| | 14 CTX PyrL6a | 2599 | 2322 | 1542 | 1252 | 1212 | 544 |
| | 18 CLAU Pyr | 2969 | 2717.5 | 1569 | 1387 | 1376.5 | 555 |
| | 15 CTX PyrL5/L6 Sulf1 | 3118 | 2984.5 | 1649 | 1425 | 1416.5 | 571 |
| | | | | | | | |
| 80 dpi | 68 Astro Slc7a10 | 710 | 500.5 | 630 | 490 | 377.5 | 322 |
| | 61 OPC | 782 | 540.5 | 688 | 537 | 412 | 349 |
| | 69 Astro Prdm16 | 783 | 560 | 654 | 536 | 415 | 343 |
| | 66 VLMC Slc6a13 | 935 | 567.5 | 1000 | 620 | 451 | 474 |
| | 57 Oligo MOL | 1111 | 645 | 1061 | 693 | 476 | 500 |
| | 46 Migrating Int Cpa6 | 976 | 656 | 858 | 609 | 475.5 | 400 |
| | 72 Ependyma | 1146 | 716 | 963 | 756 | 541 | 496 |
| | 64 Endothelia | 1179 | 836 | 955 | 769 | 595.5 | 500 |
| | 47 Migrating Int Foxp2 | 1180 | 857 | 987 | 717 | 584 | 435 |
| | 4 Medium Spiny Neurons | 1808 | 1435 | 1286 | 977 | 869 | 513 |
| | 50 Migrating Int Adarb2 | 1902 | 1559 | 1237 | 1003 | 912 | 475 |
| | 10 CTX PyrL4 Rorb | 1780 | 1559.5 | 987 | 983 | 919.5 | 401 |
| | 13 CTX PyrL5 Fezf2 | 1861 | 1567 | 1157 | 1032 | 956 | 468 |
| | 17 CTX PyrL6 | 2120 | 1754.5 | 1378 | 1097 | 1011 | 522 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 44 Migrating Int Lhx6 | 2111 | 1787 | 1352 | 1123 | 1032 | 537 |
| | 9 CTX PyrL2/L3/L4 Mef2c | 2296 | 1920.5 | 1523 | 1163 | 1092 | 558 |
| | 11 CTX PyrL4/L5 | 2377 | 2016 | 1551 | 1177 | 1116 | 554 |
| | 14 CTX PyrL6a | 2444 | 2110.5 | 1474 | 1215 | 1152.5 | 524 |
| | 7 CTX PyrL2/L3 Met | 2649 | 2390.5 | 1387 | 1308 | 1266 | 485 |
| | 15 CTX PyrL5/L6 Sulf1 | 2662 | 2446 | 1494 | 1283 | 1257 | 540 |
| | 18 CLAU Pyr | 2873 | 2525.5 | 1525 | 1387 | 1323.5 | 536 |
| | 12 CTX PyrL5 Itgb3 | 3300 | 3308.5 | 1548 | 1474 | 1497 | 493 |
| | | | | | | | |
| **120 dpi** | 63 Microglia | 878 | 555 | 851 | 622 | 459 | 438 |
| | 68 Astro Slc7a10 | 930 | 638 | 824 | 619 | 470 | 416 |
| | 61 OPC | 956 | 694 | 794 | 632 | 507 | 384 |
| | 46 Migrating Int Cpa6 | 1110 | 696.5 | 1044 | 666 | 492 | 452 |
| | 66 VLMC Slc6a13 | 1164 | 747.5 | 1113 | 745 | 563.5 | 505 |
| | 69 Astro Prdm16 | 1043 | 756.5 | 858 | 670 | 535 | 415 |
| | 57 Oligo MOL | 1426 | 825.5 | 1329 | 839 | 592.5 | 585 |
| | 49 Migrating Int Lgr6 | 1403 | 909 | 1166 | 830 | 630 | 523 |
| | 47 Migrating Int Foxp2 | 1490 | 1141 | 1085 | 857 | 727 | 454 |
| | 4 Medium Spiny Neurons | 2119 | 1776 | 1369 | 1102 | 1014 | 526 |
| | 13 CTX PyrL5 Fezf2 | 2211 | 1880 | 1233 | 1181 | 1096 | 473 |
| | 10 CTX PyrL4 Rorb | 2255 | 2000 | 1272 | 1162 | 1094 | 471 |
| | 50 Migrating Int Adarb2 | 2349 | 2028.5 | 1465 | 1169 | 1106 | 537 |
| | 44 Migrating Int Lhx6 | 2491 | 2168 | 1460 | 1265 | 1194 | 549 |
| | 17 CTX PyrL6 | 2523 | 2243 | 1487 | 1243 | 1199 | 535 |
| | 9 CTX PyrL2/L3/L4 Mef2c | 2766 | 2463 | 1692 | 1320 | 1305 | 584 |
| | 14 CTX PyrL6a | 2761 | 2545.5 | 1626 | 1323 | 1309.5 | 567 |
| | 7 CTX PyrL2/L3 Met | 2814 | 2559 | 1533 | 1349 | 1323 | 519 |
| | 11 CTX PyrL4/L5 | 2888 | 2630.5 | 1640 | 1362 | 1334.5 | 555 |
| | 12 CTX PyrL5 Itgb3 | 3340 | 2955 | 1906 | 1493 | 1494 | 586 |
| | 15 CTX PyrL5/L6 Sulf1 | 3294 | 3139.5 | 1601 | 1485 | 1486.5 | 530 |
| | 18 CLAU Pyr | 3336 | 3141 | 1595 | 1521 | 1500 | 503 |
| | | | | | | | |
| **End-stage** | 63 Microglia | 751 | 497 | 704 | 559 | 415.5 | 390 |
| | 68 Astro Slc7a10 | 631 | 516.5 | 517 | 450 | 378 | 247 |
| | 49 Migrating Int Lgr6 | 845 | 584.5 | 726 | 566 | 434.5 | 352 |
| | 61 OPC | 863 | 596.5 | 774 | 583 | 456 | 378 |
| | 48 Migrating Int Pbx3 | 1054 | 649.5 | 945 | 674 | 500 | 456 |
| | 66 VLMC Slc6a13 | 1080 | 650 | 980 | 711 | 499.5 | 490 |
| | 69 Astro Prdm16 | 1044 | 691 | 937 | 662 | 494 | 451 |
| | 46 Migrating Int Cpa6 | 1148 | 725 | 1063 | 675 | 499 | 455 |
| | 57 Oligo MOL | 1333 | 794 | 1226 | 802 | 581.5 | 558 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 64 Endothelia | 1414 | 913 | 1208 | 880 | 634.5 | 587 |
| | 47 Migrating Int Foxp2 | 1373 | 962 | 1088 | 810 | 653 | 467 |
| | 72 Ependyma | 1370 | 985.5 | 1039 | 887 | 705.5 | 537 |
| | 4 Medium Spiny Neurons | 1755 | 1363 | 1261 | 968 | 845 | 508 |
| | 13 CTX PyrL5 Fezf2 | 1722 | 1427 | 1144 | 987 | 892.5 | 461 |
| | 10 CTX PyrL4 Rorb | 1806 | 1511 | 1157 | 1010 | 930 | 470 |
| | 50 Migrating Int Adarb2 | 2020 | 1675 | 1325 | 1061 | 983 | 514 |
| | 17 CTX PyrL6 | 2106 | 1722 | 1370 | 1105 | 1010 | 517 |
| | 9 CTX PyrL2/L3/L4 Mef2c | 2271 | 1846.5 | 1549 | 1159 | 1074 | 561 |
| | 44 Migrating Int Lhx6 | 2234 | 1862 | 1439 | 1181 | 1092 | 559 |
| | 7 CTX PyrL2/L3 Met | 2293 | 1975 | 1457 | 1183 | 1130 | 534 |
| | 14 CTX PyrL6a | 2328 | 1988.5 | 1457 | 1192 | 1133 | 535 |
| | 11 CTX PyrL4/L5 | 2566 | 2218 | 1595 | 1251 | 1203 | 550 |
| | 18 CLAU Pyr | 2721 | 2357.5 | 1508 | 1343 | 1287.5 | 522 |
| | 15 CTX PyrL5/L6 Sulf1 | 2931 | 2679 | 1565 | 1385 | 1334 | 533 |
| | 12 CTX PyrL5 Itgb3 | 3333 | 3307 | 1732 | 1478 | 1512 | 543 |

***Supplementary Table 4: Additional metrics of the cell clusters from the mouse transcriptomics study.*** *SD: standard deviation. Mean and standard deviation have been rounded to the nearest integer.*

## 7.3 External Tables

**External Supplementary Table 1: Differentially expressed genes identified from the comparison between the two controls (CD1 vs PBS) using Seurat.** The tables include genes with an adjusted p-value of less than 0.05. p_val: the p-value of the Wilcoxon rank-sum test; pct1: the percentage of cells in the CD1 group that express the gene; pct2: the percentage of cells in the PBS group that express the gene; avg_log2FC: average log2-fold change of gene expression between groups; p_val_adj: Bonferroni-corrected adjusted p-value. Document worksheets correspond to the 5 time points.

**External Supplementary Table 2: Differentially expressed genes identified from the comparison between RML and CD1 groups using Seurat.** The tables include genes that passed the filtering criteria (adjusted p-value of less than 0.05 and not part of the set of 7 spurious genes identified by the comparison of the two controls). p_val: the p-value of the Wilcoxon rank-sum test; pct1: the percentage of cells in the RML group that express the gene; pct2: the percentage of cells in the CD1 group that express the gene; avg_log2FC: average log2-fold change of gene expression between groups; p_val_adj: Bonferroni- corrected adjusted p-value; gene_unique: set to TRUE if the gene has been found only in a specific cluster of a time point; in_glmGamPoi: set to TRUE if the gene has been identified in the same time point using glmGamPoi for the analysis; in_glmGamPoi_same_cluster: set to TRUE if the gene has been identified in the same time point and cluster using glmGamPoi for the analysis; in_DESeq2: set to TRUE if the gene has been identified in the same time point using DESeq2 for the analysis; in_DESeq2_same_cluster: set to TRUE if the gene has been identified in the same time point and cluster using DESeq2 for the analysis. Document worksheets correspond to the 5 time points.

**External Supplementary Table 3: Differentially expressed genes identified from the comparison between RML and CD1 groups using DESeq2.** The tables include genes that passed the filtering criteria (adjusted p-value of less than 0.05 and not part of the set of 7 spurious genes identified by the comparison of the two controls). log2FoldChange: average log2-fold change of gene expression between groups; pval: the p-values of the Wald test; padj: Benjamini–Hochberg corrected adjusted p-value; baseMean: average of the normalized count values divided by size factors, calculated over all samples; lfcSE: standard error of the log-fold change.

**External Supplementary Table 4: Differentially expressed genes identified from the comparison between RML and CD1 groups using glmGamPoi.** The tables include genes that passed the filtering criteria (adjusted p-value of less than 0.05 and not part of the set of 7 spurious genes identified by the comparison of the two controls). pval: the p-values of the quasi-likelihood ratio test; adj_pval: Benjamini–Hochberg corrected adjusted p-value; f_statistic: statistic of the F-test of overall significance, which indicates whether the linear regression model provides a better fit to the data than a model that contains no independent variables; df1: the degrees of freedom of the test; df2: the degrees of freedom of the test; lfc: average log2-fold change of gene expression between groups.

**External Supplementary Table 5: All set intersections between DEGs across all time points.** Each worksheet is named after the groups that are intersected, for example, worksheet 20dpi_120dpi_end includes genes that were found to be DE at 20 dpi, 120 dpi and end-stage.

**External Supplementary Table 6: All identified gene ontology terms from the over-representation analysis.** An over-representation analysis using clusterProfiler identified perturbed biological pathways in the 20 and 120 dpi time points and the end-stage. No enriched pathways were identified for the 40 or 80 dpi time points. Each worksheet corresponds to a GO classification: BP: biological process, CC: cellular component, MF: molecular function. Cluster: the cell cluster identified from the single-cell analysis; ID: the GO identifier of the relevant gene set; Description: a short description of the GO gene set; GeneRatio: the ratio of the intersection of DE genes in our data with the GO gene set over the intersection of DE genes in our data with all the genes of the GO collection; BgRatio: the ratio of the size of the GO gene set over the size of all identified genes in our analysis (the gene universe), pvalue: the p-value of a one-sided Fisher's exact test; p.adjust: the Benjamini-Hochberg adjusted p-value; qvalue: p-value that has been adjusted for the False Discovery Rate (FDR); geneID: gene symbols of the DE genes in our dataset that are part of the GO gene set; Count: number of the DE genes in our dataset that are part of the GO gene set.

**External Supplementary Table 7: All identified gene ontology terms from the gene set enrichment analysis.** A GSEA using clusterProfiler identified perturbed biological pathways in the 20, 80, 120 dpi and end-stage time points. Each worksheet corresponds to a GO classification: BP: biological process, CC: cellular component, MF: molecular function. ID: the GO identifier of the enriched term; Description: a short description of the GO term; setSize: the number of genes associated with the GO term; enrichmentScore: the primary result of the analysis, which reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes; NES: the normalised enrichment score which accounts for differences in gene set size and correlations between gene sets and the expression dataset; p.adjust: the Benjamini-Hochberg adjusted p-value; qvalues: p-values that have been adjusted for the False Discovery Rate (FDR); rank: the rank of the gene in the sorted gene list when the maximum enrichment score is encountered for a specific gene set; core_enrichment: genes of the leading-edge subset within the gene set. These are the genes that contribute most to the enrichment result.

## 7.4 External Files

**External Supplementary File 1: FastQC reports from sequenced libraries of the mouse experiment.** The file includes reports generated from FastQC for each of the 3 libraries sequenced at each time point. File names ending in _R1 represent the first sequenced read of the pair which contains the gene expression information. File names ending in _R2 represent the second sequenced read of the pair which contains the barcode information for demultiplexing the data and identifying the cell of origin. A more detailed description of the sequencing quality report and more information regarding the interpretation of the plots can be found on the author's website (Andrews, 2010).

**External Supplementary File 2: FastQC reports from sequenced libraries of the human experiment.** The file includes reports generated from FastQC for each of the 6 libraries sequenced. File names ending in _R1 represent the first sequenced read of the pair which contains the gene expression information. File names ending in _R2 represent the second sequenced read of the pair which contains the barcode information for demultiplexing the data and identifying the cell of origin. A more detailed description of the sequencing quality report and more information regarding the interpretation of the plots can be found on the author's website (Andrews, 2010).

## 7.5 Protocols

### 7.5.1 DroNc-seq

The DroNc-seq protocol was downloaded from Protocol Exchange on the 28th of April 2020 (A. Basu et al., 2017).

# DroNc-seq step-by-step

CURRENT STATUS: POSTED

Aviv Regev
Broad Institute

aregev@broadinstitute.org*Corresponding Author*

Feng Zhang
Broad Institute

zhang@broadinstitute.org*Corresponding Author*

Anindita Basu
Broad Institute, Harvard University

Inbal Avraham-Davidi
Broad Institute

Naomi Habib
Broad Institute

Karthik Shekhar
Broad Institute

Matan Hofree

David Weitz
Harvard University

Orit Rozenblatt-Rosen
Broad Institute

Tyler Burks
Broad Institute

Sourav Choudhury
Broad Institute

François Aguet
Broad Institute

Ellen Gelfand

Kristin Ardlie
Broad Institute

**DOI:**

10.1038/protex.2017.094

**SUBJECT AREAS**

*Biotechnology* *Molecular Biology*

**KEYWORDS**

*single nuclei, high throughput, RNA-seq, droplet, microfluidics, DNA barcode, frozen, RNAlater, tissue, hippocampus, PFC*

## Abstract

Currently, most single cell protocols require the preparation of a single cell suspension from fresh tissue, a major roadblock to clinical deployment, to archived materials and to certain tissues such as adult brain. In the adult brain the harsh enzymatic dissociation harms the integrity of the cells and their RNA, and biases toward easily dissociated cell types, and is restricted to young animals.

We developed DroNc-seq, a droplet microfluidic and DNA barcoding technique for analysis of RNA profiles of single nuclei from fresh, frozen or lightly fixed tissues at high throughput and low cost. The utility of DroNc-Seq lies in working with hard-to-dissociate, frozen and/or archived tissues. To demonstrate the utility of this technique, we sequenced over 39 thousand nuclei from mouse and human archived brain samples, including post-mortem human brain tissue from GTEx project.

## Reagents

Reagents:

a. Nuclei EZ lysis buffer (Sigma, #EZ PREP NUC-101)

b. RNAlater (ThermoFisher Scientific, Cat # AM7020)

c. PBS buffer (ThermoFisher Scientific, Cat # 10010023)

d. DNAse/RNAse free distilled water (ThermoFisher Scientific, Cat # 10977023)

e. BSA, molecular biology grade, 20 mg/ml (New England Biolabs, Cat # B9000S)

f. Ficoll PM-400 (Sigma, Cat # F5415-50ML)

g. Sarkosyl (Teknova, Inc., Cat # S3377)

h. 0.5 M EDTA (Life Technologies)

i. 1M Tris pH 7.5 (Sigma)

j. 1M DTT (Teknova, Inc., Cat # D9750)

k. 20% PEG solution (Teknova, Inc., Cat # P4137)

l. 10% SDS solution (Teknova, Cat #S0287)

m. 10% Tween 20 solution (Teknova, Cat # T0710)

n. Carrier oil (BioRad Sciences, Cat # 186-4006)

o. DAPI (ThermoFisher Scientific, Cat # D1306)

p. 6x SSC (Teknova, Inc., Cat # S0282)

q. 1H,1H,2H,2H-Perfluorooctan- 1-ol (SynQuest Laboratories, Cat # 647-42- 7)

r. 1x Maxima H- RT buffer (Fisher, Cat # EP0753)

s. dNTP (Takara Bio, Cat # 639125)

t. RNase Inhibitor (Lucigen, Cat # 30281-2)

u. Maxima H-RT enzyme (Fisher, Cat # EP0753)

v. Exonuclease I kit (New England Biolabs, Cat # M0293L)

w. 2x Kapa HiFi Hotstart Readymix (Kapa Biosystems, Cat # KK2602)

x. Nextera XT sample prep kit, 96 samples (Illumina, Cat # FC-131- 1096)

Primers:

a. Barcoded bead, sequence: TTTTTTTAAGCAGTGGTATCAACGCAGAGTACJJJJJJJJJJJJ NNNNNNNNT(30); where J=split-pool oligo; N=random oligo (Chemgenes, Cat # Macosko-2011- 10)

b. Template Switch Oligo, AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG (IDT, custom RNA oligo, HPLC purified)

c. SMART PCR primer, AAGCAGTGGTATCAACGCAGAGT (IDT, custom DNA oligo, standard desalting)

d. P5-PCR hybrid oligo AATGATACGGCGACCACCGAGATCTACACGCCTGTC CGCGGAAGCAGTGGTATCAACGCAGAGT**A**C, (IDT, custom DNA oligo)

e. Custom Read1 primer, GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC (IDT, custom DNA oligo, standard desalting)

Consumables:

a. Cell strainer, 35 µm (Corning, Cat # 352235)

b. Cell strainer, 40 µm (PluriSelect, Cat # 43-50040- 03)

c. Cell strainer, 100 µm (VWR, Cat #08-771- 19)

d. Dounce homogenizers (Sigma, Cat # D8938-1SET)

e. Fuchs-Rosenthal (FR) hemocytometer (VWR, Cat # 22-600- 102)

f. Neubauer Improved (NI) Hemocytometer (Life Technologies, Cat # 22-600- 100)

g. 3ml syringe (BD Scientific, Cat # BD309657)

h. 10 ml syringe (BD Scientific, Cat # BD309695)

i. 26G1/2 sterile needles (BD Scientific, Cat # BD305111)

j. PE tubing (Scientific Commodities, Inc. Cat # BB31695-PE/2)

k. Flea magnet (VP Scientific, cat # 782N-6- 150)

l. 1.5 ml micro-centrifuge tube (Ambion, Cat # AM12450)

m. Ampure XP beads (Beckman Coulter, Cat # A63881)

n. Qubit dsDNA HS Assay kit (ThermoFisher, Cat # Q32854)

o. BioAnalyzer High Sensitivity Chip (Agilent, Cat # 5067-4626)

p. Illumina NextSeq 75

## Equipment

a. Microfluidic chip (see CAD file). The unit in the CAD provided is 1 unit = 1 µm; channel depth on device is 75 µm.

b. Drop-seq microfluidic setup (see reference):

optical microscope (Olympus IX83)
Fast camera (Photron SA5)
Three syringe pumps (KD Scientific, KDS910)
Magnetic Stirrer (VP Scientific, #710D2)

c. Invitrogen Qubit 3.0 Fluorometer

d. Agilent 2100 Bioanalyzer

e. Illumina NextSeq 500

## Procedure

Protocol:

1. Beads preparation: a. Wash and filter barcoded beads (Chemgenes, Cat # Macosko-2011-10) as previously described6. Isolate beads smaller than 40 µm, using a 40 µm cell strainer (PluriSelect, Cat # 43-50040-03).

b. Suspend barcoded beads in Drop-seq Lysis Buffer (DLB6; a 10 ml stock consists of 4 ml of nuclease-free H2O, 3 ml 20% Ficoll PM-400 (Sigma, Cat # F5415-50ML), 100 µl 20% Sarkosyl (Teknova, Inc., Cat # S3377), 400 µl 0.5M EDTA (Life Technologies), 2 ml 1M Tris pH 7.5 (Sigma), and 500 µl 1M DTT (Teknova, Inc., Cat # D9750), where the DTT is added fresh before every experiment). Count beads at

1:1 dilution in 20% PEG solution, using a disposable Fuchs-Rosenthal hemocytometer (VWR, Cat # 22-600-102) and resuspend beads at concentrations ranging between 325,000 and 350,000 per ml.

2. Cell culture: Cell lines are cultured according to ATCC's instructions. For DroNc-seq, wash cells once with 1x PBS, scrape them with 2 ml nuclease- and protease-free Nuclei EZ lysis or EZ PREP buffer (Sigma, Cat # EZ PREP NUC-101) and process as tissues, described below.

3. Tissue preservation: Tissue samples may be flash-frozen on dry ice and stored at -80°C until they are processed for nuclei isolation. To preserve tissue in RNAlater, samples are placed in ice-cold RNAlater (ThermoFisher Scientific, Cat # AM7020) and stored at 4°C overnight. RNAlater is removed the following day and samples are then stored at -80°C until processing.

4. Nuclei isolation: a. Use either fresh, frozen or RNAlater fixed tissue or fresh cells as input material.

b. Prepare Nuclei Suspension Buffer (NSB; consisting of 1x PBS, 0.01% BSA (New England Biolabs, Cat # B9000S) and 0.1% RNAse inhibitor (Clontech, Cat #2313A)).

c. Dounce homogenize tissue samples (smaller than 0.5 cm) or cell pellets in 2 ml of ice-cold Nuclei EZ lysis buffer (Sigma, #EZ PREP NUC-101). For brain tissue: grind 20-25 times with pestle A, followed by 20-25 times with pestle B (This may need to be modified for other tissues). Move sample to a 15 ml conical tube, add 2 ml of ice-cold Nuclei EZ lysis buffer and incubate on ice for 5 minutes.

d. Collect nuclei by centrifugation at 500 x g for 5 minutes at 4°C. Discard supernatant and carefully resuspend nuclei in 4 ml of ice-cold Nuclei EZ lysis buffer. Incubate on ice for 5 minutes. Collect nuclei by centrifugation at 500 x g for 5 minutes at 4°C.

e. Resuspend isolated nuclei in 4 ml of NSB and collect nuclei by centrifugation at 500 x g for 5 minutes at 4°C.

f. Resuspend isolated nuclei in 1 ml of NSB, and filter through a 35 µm cell strainer (Corning, Cat # 352235). Stain 10 µl of the single nuclei suspension with DAPI (Fisher, Cat # D1306), load on an NI hemocytometer, and count under a microscope. A final concentration of 300,000 nuclei/ml is used for DroNc-seq experiments. Proceed immediately to microfluidic droplet co-encapsulation.

5. Microfluidics: a. Load the nuclei and barcoded bead suspension into 3 ml syringes (BD Scientific, Cat # BD309695) and connect to DroNc-seq microfluidic chip via 26G1/2 sterile needles (BD Scientific, Cat # BD305111) and PE2 tubing (Scientific Commodities, Inc. Cat # BB31695-PE/2). Note that the bead syringe is loaded onto the syringe pump in an upside down position, along with a flea magnet inside the syringe and constant stirring, using external magnetic stirrer. Flow both bead and nuclei suspensions at 1.5 ml/hr each, along with carrier oil (BioRad Sciences, Cat # 186-4006) loaded in 10 ml syringes (BD Scientific, Cat # BD309695) and flown at 16 ml/hr to co-encapsulate single nuclei and beads in ~75 µm drops at 4,500 drops/sec and double Poisson loading concentrations.

b. Collect resulting emulsion via PE2 tubing into a 50 ml Falcon tube for a period of ~22 min each, and incubate at room temperature for up to 45 min before proceeding to break droplets.

6. Droplet breakage, washes and reverse transcription (RT): a. Emulsion collected after microfluidic co-encapsulation has the droplets cream to the top with clear oil collected under the droplets. Carefully remove the excess clear oil, add 30 ml of 6x SSC (Teknova, Inc., Cat # S0282) into each 50 ml Falcon collection tube, agitate it vigorously, and add 1 ml of 1H,1H,2H,2H-Perfluorooctan-1-ol (SynQuest Laboratories, Cat # 647-42-7). It is recommended that all washes following this step be performed and the beads temporarily stored on ice.

b. Vigorously shake the tubes by hand and centrifuge at 1,000 x g for 1 min.

c. Carefully remove the supernatant from each tube and squirt an additional 30 ml of 6x SSC to kick up the beads from the oil-water interface into the aqueous phase.

d. Remove the beads that were kicked up momentarily into the SSC with a 25 ml pipette and transfer them into a clean 50 ml Falcon tube, leaving the heavier oil behind.

e. Centrifuge the newly transferred beads and SSC mix again at 1,000 x g for 1 min; carefully remove the supernatant leaving ~1 ml of SSC and bead sediment behind.

f. Carefully transfer remaining SSC and bead mix into a 1.5 ml micro-centrifuge tube (Ambion, Cat # AM12450) and spin it down on a desktop micro-centrifuge for ~10 sec to generate a noticeable bead pellet.

g. Remove any residual oil that got transferred into the 1.5 ml tube with a p200 pipette with low-retention pipette tip.

h. Wash the beads again in 1.5 ml of 6x SSC and then again in 300 µl of 5x Maxima H- RT buffer (Fisher, Cat # EP0753). A pellet of barcoded beads in each micro-centrifuge tube should have ~130,000 beads.

i. Make a fresh batch of 200 µl RT mix for each barcoded bead aliquot, consisting of: 80 µl H2O, 40 µl Maxima 5x RT Buffer, 40 µl 20% Ficoll PM-400 (Sigma, Cat # F5415-50ML), 20 µl 10 mM dNTP (Takara Bio, Cat # 639125), 5 µl RNase Inhibitor (Lucigen, Cat # 30281-2), 10 µl Maxima H-RT enzyme (Fisher, Cat # EP0753), and 5 µl 100 µM Template Switch Oligo, AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG (IDT, custom RNA oligo, HPLC purification). After the supernatant is carefully removed from each bead pellet, add 200 µl of the above RT mix into each tube, and incubate it under gentle rocking or tumbling for 30 min at room temperature, and then at 42°C for 1.5 hr in a rotisserie-style hybridization oven, for a total of two hours.

7. Post RT wash, exonuclease I treatment and PCR: a. Post RT, each bead has cDNA barcoded with the bead's unique barcode (BC) bound onto it, also referred to as a STAMP6. Wash each STAMP pellet with (1) 1 ml of TE buffer containing 0.5% SDS (TE-SDS), once; (2) 1 ml of TE buffer containing 0.01% Tween-20 (TE-TW), twice; and (3) 1 ml of 10 mM Tris pH 8.0, once.

b. Spin down to remove all supernatant and treat the STAMPs with exonuclease I (New England Biolabs, Cat # M0293L) as follows: add 20 µl of Exo I buffer, 170 µl of RNAse free water, 10 µl of Exo I enzyme, mix well by pipetting up and down, and incubate for 45 min at 37°C under rotation to remove all unextended primers.

c. Wash the pellet with TE-SDS and TE-TW washes (as described in a), followed by a round of wash in 1 ml of RNAse free water. You may pool beads from multiple collections of a given sample at this point.

d. Resuspend pellet in 1 mL of H2O, and count them, by mixing 10 µl of bead suspension with an equal volume of 20% PEG solution.

e. Resuspend aliquots of 5,000 beads in a PCR mix each consisting of 24.6 µl H2O, 0.4 µl 100 µM SMART PCR primer, AAGCAGTGGTATCAACGCAGAGT (IDT, custom DNA oligo, standard desalting purification), and 25 µl 2x Kapa HiFi Hotstart Readymix (Kapa Biosystems, Cat # KK2602).

f. Amplify the samples in separate wells on a skirted PCR plate, using the Eppendorf Thermocycler (Part # EP-950030020).

i. Mouse PCR samples were amplified using the following PCR steps: 95°C for 3 min; then 4 cycles of: 98°C for 20 sec, 65°C for 45 sec, 72°C for 3 min; then 10 cycles of: 98°C for 20 sec, 67°C for 20 sec, 72°C for 3 min; and finally, 72°C for 5 min. Amplified mouse PCR products were pooled in batches of 4 wells or 16 wells.

ii. Human PCR samples were amplified with either the previously mentioned PCR steps, or the following PCR steps: 95°C for 3 min; then 4 cycles of: 98°C for 20 sec, 65°C for 45 sec, 72°C for 3 min; then 12 cycles of: 98°C for 20 sec, 67°C for 20 sec, 72°C for 3 min; and finally, 72°C for 5 min. Amplified human PCR products were pooled in batches of 4 wells (16 total PCR cycles) or 16 wells (14 total PCR cycles).

g. Combine the 5,000 STAMP aliquots of each well in a 1.5 ml Eppendorf tube and clean with 0.6X SPRI beads (Ampure XP beads, Beckman Coulter, Cat # A63881).

Note that the total number of PCR wells from a single sample depends on the number of STAMPs collected in a DroNc-seq run from a given input of nuclei. A user may access the pool of STAMPs in different ways, depending on the number of nuclei they wish to retrieve and their sequencing setup. In particular, a user would typically access the pool of STAMPs once or more, each time taking only a portion of the STAMPs to generate a library, and repeat the process if more nuclei are desired. For our mouse and human brain samples, it was optimal to pool 20,000 STAMPs in each PCR reaction and then to pool 4 PCR wells together for the library preparation step. Depending on the amount of desired reads per nucleus and the sequencing yield, a user may pool a higher number of PCR wells in a single Nextera library, as we demonstrate here using 16-32 wells.

8. WTA library QC and Nextera library prep: a. Quantify purified cDNA using Qubit dsDNA HS Assay kit (ThermoFisher Scientific, Cat # Q32854) and BioAnalyzer High Sensitivity Chip (Agilent, Cat # 5067-4626).

b. Use 550 pg of each sample library for fragmentation, tagging and amplification using the Nextera XT sample prep kit, 96 samples (Illumina, Cat # FC-131-1096), and custom primer, AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT**A**C, (IDT, custom DNA oligo, HPLC purification) that enable selective amplification of the 3’ end, according to manufacturer’s instructions.

c. Quantify Nextera libraries again with Qubit dsDNA HS Assay kit and BioAnalyzer High Sensitivity Chip.

9. Sequencing: a. The libraries (at 2.2 pM (mouse, 16 wells pool), 2.7 pM (mouse, 4 wells pool) and 2.3 pM (human)) were sequenced on an Illumina NextSeq 500. We used NextSeq 75 cycle kits to sequence paired-end reads as follows: 20 bp (Read 1), 60 bp (Read 2), and 8 bp for Index 1, with Custom Read1 primer, GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC (IDT, custom DNA oligo, standard desalting), according to Illumina loading instructions.

b. The sequencing cluster density and percent passing filter number from different experiments vary according to the quality of nuclei samples used, but were optimized at around a cluster density of 220 and a 90% passing filter.

## References

Macosko, E. Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using

Nanoliter Droplets. Cell 161, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

## Acknowledgements

We thank Rhiannon Macare, Assaf Rotem, Christoph Muus and Eugene Drokhlyansky for helpful discussions, Talia Habib for babysitting, Timothy Tickle and Asma Bankapur for technical support, Leslie Gaffney and Ania Hupalowska for help with graphics. Work was supported by Klarman Cell Observatory, NIMH grant U01MH105960, NCI grant 1R33CA202820-1 (to A.R.), and Koch Institute Support (core) grant P30-CA14051 from the NCI. Microfluidic devices were fabricated at the Center for Nanoscale Systems, Harvard University, member of NNIN, supported by NSF award no. 1541959. A.R. is an Investigator of Howard Hughes Medical Institute (HHMI). A.R. is a member of Scientific Advisory Boards for Thermo Fisher Scientific, Syros Pharmaceuticals and Driver Genomics. F.Z. is a New York Stem Cell Foundation-Robertson Investigator. F.Z. is supported by NIH through NIMH (5DP1-MH100706 and 1R01-MH110049), NSF, HHMI, the New York Stem Cell, Simons, Paul G. Allen Family, and Vallee Foundations; and James and Patricia Poitras, Robert Metcalfe, and David Cheng. D.A.W. thanks NSF DMR-1420570, NSF DMR-1310266 and NIH P01HL120839 grants for their support. NH is a HHMI fellow for Helen Hey Whitney Foundation. N.H., A.B., I.A.D., D.A.W., F.Z. and A.R. are inventors on international patent application PCT/US16/59239 filed by the Broad Institute, Harvard and MIT, relating to method of this manuscript. GTEx is supported by the Common Fund of the Office of the Director of NIH, through Contract HHSN268201000029C (to K.A LDACC, Broad Institute).

## Massively parallel single-nucleus RNA-seq with DroNc-seq

by Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, +11
Nature Methods (29 August, 2017)

### 7.5.2 SPLiT-seq

The SPLiT-seq protocol v3 was downloaded from https://sites.google.com/uw.edu/splitseq/protocol on the 28th of April 2020. Available at https://www.seeliglab.org/tools.html as of the 11th of January 2022.

## SPLiT-seq Protocol, Version 3.0

Projected Experimental Time: 2 Days
Recommended time on day 1 to start: morning

### Addition of RNase inhibitor to buffers:

When any buffer has “+RI” next to it, this indicates that enzymatics RNase inhibitor should be added to a final concentration of 0.1 U/uL.

### Centrifugation Steps:

All centrifugation steps should be performed with a swinging bucket rotor. Using a fixed angle centrifuge may lead to more cell loss. Depending on the tissue type, centrifugation speeds may need to be changed to optimize cell retention (e.g. smaller cells = higher speeds).

### DNA Barcoding Plate Generation

What you need:

- Three 96 well plates from IDT - Reverse Transcription Barcode Primers, Ligation Round 1,and Ligation Round 2 Stock DNA Oligo plates (100 uM)
- Two linker oligos - BC_0215, BC_0060 (*Note: these are assumed to be in stock concentration of 1mM, be sure to correct for volume if only have 100 uM stocks)*
- Six 96 well PCR plates (3 stock plates that will last at least 10 experiments, and 3 plates for 1st experiment)

*Note: This will generate 100 uL of DNA barcodes for each well. Each SPLiT-seq experiment requires only 4 uL/well of the reverse transcription primer solution which will last for 25 experiments. Each SPLiT-seq experiment requires only 10 uL/well of the barcode/linker solutions, so these plates will last a total of 10 experiments.*

*Round 1 reverse transcription barcoded primers (final concentrations of 12.5 uM random hexamer and 12.5 uM 15dT primers in each of 48 wells)*

1. Using multichannel pipette, add 12.5 uL of rows A-D in the IDT Reverse Transcription Barcode Primers to rows A-D of the BC Stock 96 well PCR plate.
2. Using multichannel pipette, add 12.5 uL of rows E-H in the IDT Reverse Transcription Barcode Primers to rows A-D of the BC stock 96 well PCR plate (mixing polydT with random hexamer primer here)
3. Add 75ul of water to rows A-D of the BC stock 96 well PCR plate.

*Round 2 ligation round (Final concentrations of 12uM barcodes, 11uM linker-BC_0215)*

1. Using multichannel pipette, add 12uL of IDT Round 2 Barcodes to R1 Stock 96 well PCR plate
2. Add 138.6ul of BC_0215(1mM) to 10.9494mL water in a basin (BC_0215_dil)
3. Using multichannel pipette, add 88uL BC_0215_dil to each well of R2 Stock 96 well PCR plate

*Ligation Round 3 (Final concentrations of 14uM barcodes, 13uM linker-BC_0060)*

1. Using multichannel pipette, add 14uL of Round 3 Barcodes to R3 Stock 96 well PCR plate
2. Add 163.8ul of BC_0060(1mM) to 10.6722mL water in a basin (BC_0060_dil)
3. Using multichannel pipette, add 86uL BC_0060 to each well R3 Stock 96 well PCR plate

For each ligation plate (R2 and R3, not including reverse transcription barcodes), anneal the barcode and linker oligos with the following thermocycling protocol:

1. Heat to 95C for 2 minutes
2. Ramp down to 20C for at a rate of -0.1C/s
3. 4C

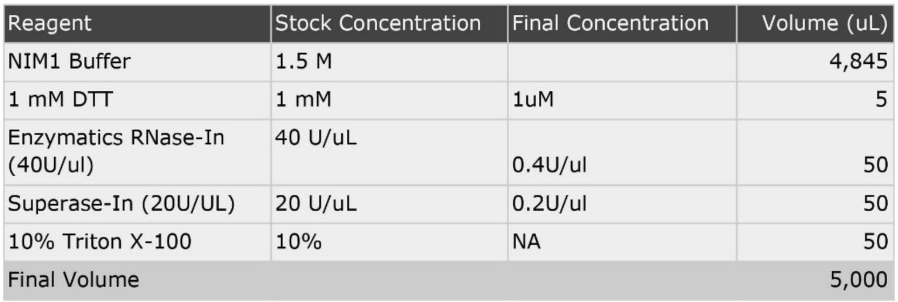
Aliquot out 10 uL of each barcode/linker stock plate into 3 new 96 well PCR plates. These are the plates that should be used for DNA barcoding in the split-pool ligation steps in the protocol.

## Nuclei Extraction (Optional):

1. Prepare the following items:
   - Keep dounce at 4C until use
   - 15ml of 1xPBS + 37.5 Superase-in + 19ul Enzymatics Rnase inhibitor. (kept on ice)
   - Precool centrifuge to 4C
2. Make **NIM1 buffer**:

| Reagent | Stock Concentration | Final Concentration | Volume (uL) |
|---|---|---|---:|
| Sucrose | 1.5 M | 250mM | 2,500 |
| KCl | 1 M | 25mM | 375 |
| MgCl2 | 1 M | 5mM | 75 |
| Tris buffer, pH 8 | 1 M | 10mM | 150 |
| Water | NA | NA | 11,900 |
| Final Volume | | | 15,000 |

3. Make the **homogenization buffer:**



| Reagent | Stock Concentration | Final Concentration | Volume (uL) |
|---|---|---|---|
| NIM1 Buffer | 1.5 M | | 4,845 |
| 1 mM DTT | 1 mM | 1uM | 5 |
| Enzymatics RNase-In (40U/ul) | 40 U/uL | 0.4U/ul | 50 |
| Superase-In (20U/UL) | 20 U/uL | 0.2U/ul | 50 |
| 10% Triton X-100 | 10% | NA | 50 |
| Final Volume | | | 5,000 |

4. Dounce
    - ○ Add tissue/cells sample to dounce. If cells, resuspend in 700ul of homogenization buffer.
    - ○ Add homogenization buffer to ~700ul
    - ○ Perform 5 strokes of loose pestle
    - ○ Perform 10 - 15 of tight pestle
    - ○ Add homogenization buffer up to 1ml
    - ○ Check cell lysis with 5ul trypan blue and 5ul cells on haemocytometer to see if nuclei have been released
5. Filter homogenates with 40um strainer into 5ml eppendorf tubes (or 15mL falcon). Tilting the filter 45° while straining over the tube ensures that the lysate passes through as intended.

    Note: This straining process is different from every other one below.
6. Spin for 4min at 600g (4C) and remove supernatant (can leave about 20uL to avoid aspirating pellet)
7. Resuspend in 1ml of 1x PBS + RI
8. Add 10ul of BSA
9. Centrifuge at 600g for 4min.
10. Resuspend in 200ul 1x PBS + RI.
11. Take 50ul of the resuspended cells from step 4 and add 150ul of 1xPBS + RI. Count sample on hemocytometer and/or flow-cytometer.
    - ○ The volume of resuspended cells from the step 4 can be changed based on the considerations of the user.
12. Pass cells through a 40um strainer into a fresh 15mL Falcon tube and place on ice.
    - ○ See note on step 4 of Fixation and Permeabilization.
13. Resuspend the desired number of nuclei (typically 2M) in 1mL 1x PBS + RI and proceed with step 5 in the following *Fixation and Permeabilization* protocol.

## Fixation and Permeabilization

1. Prepare the following buffers (calculated for two experiments):
    - A 1.33% formalin (360 uL of 37% formaldehyde solution (Sigma)+ 9.66 ml PBS) solution and store at 4C.
    - 6 mL of 1X PBS+RI (15 uL of SUPERase In and 7.5 uL of Enzymatics RNase inhibitor)
    - 2 mL of 0.5X PBS+RI (5 uL of SUPERase in and 2.5 of Enzymatics RNase inhibitor)
    - 500uL of 5% Triton X-100 + RI (2 uL of SUPERase In)
    - 1100uL of 100mM Tris pH 8.0 + 4 uL SUPERase In
    - Set the centrifuge to 4C
2. Pellet cells by centrifuging at 500g for 3 mins at 4C. (Some cells may require faster centrifugation.)
3. Resuspend cells in 1mL of cold PBS+RI. Keep cells on ice between these steps.
4. Pass cells through a 40um strainer into a fresh 15mL Falcon tube and place on ice.

    Note: The cell resuspension is not likely to passively go through the strainer, which can cause cell loss. Instead, with a 1ml pipette filled with the resuspension, press the end of the tip directly onto the strainer and actively push the liquid through. The motion should take ~1 second.
5. Add 3 mL of cold 1.33% formaldehyde (final concentration of 1% formaldehyde). Fix cells on ice for 10 mins.
6. Add 160uL of 5% Triton-X100+RI to fixed cells and mix by gently pipetting up and down 5x with a 1mL pipette. Permeabilize cells for 3 mins on ice.
7. Centrifuge cells at 500g for 3 mins at 4C.
8. Aspirate carefully and resuspend cells in 500 uL of cold PBS+RI.
9. Add 500uL of cold 100 mM Tris-HCl, pH 8.0.
10. Add 20 uL of 5% Triton X-100.
11. Centrifuge cells at 500g for 3 mins at 4C.
12. Aspirate and resuspend cells in 300 ul of cold **0.5x** PBS+RI.
13. Run cells through a 40uM strainer into a new 1.7mL tube.
    - See note on step 4 of Fixation and Permeabilization.
14. Count cells using a hemacytometer or a flow-cytometer and dilute the cell suspension to 1,000,000 cells/mL. While counting cells, keep cell suspension on ice.

    *Note: This step will dictate how many cells enter the split-pool rounds. It will be possible to sequence only a subset of the cells that enter the split-pool rounds (can be done during sublibrary generation at lysis step). The total number of barcode combinations you will be using should be calculated to determine the maximum number of cells you can sequence with minimal barcode collisions. As a rule of thumb, the number of cells you process should not exceed more than 5% of total barcode combinations. We usually have a dilution between 500k to 1M cells/mL here (equates to 4-8k cells going into each well for reverse transcription barcoding rounds).*

**Reverse Transcription**

1. Aliquot out 4 uL of the RT barcodes stock plate into the top 4 rows (48 wells) of a new 96 well plate. Cover the this plate with an adhesive plate seal until ready for use.
2. Create the following reverse transcription (RT) mix on ice:

| Reagent | Stock Concentration | Desired Concentration | Per Reaction | Volume in Mix (48 wells + 10%) |
|---|---|---|---|---|
| 5X RT Buffer | 5x | 1x | 4 | 211.2 |
| Enzymatics Rnase Inhibitor | 40u/uL | 0.25u/uL | 0.125 | 6.6 |
| Superase In Rnase Inhibitor | 20U/uL | 0.25U/uL | 0.25 | 13.2 |
| dNTPs | 10mM (per base) | 500uM | 1 | 52.8 |
| Maxima H Minus Reverse Transcriptase | 200u/uL | 20u/ul | 2 | 105.6 |
| H2O | NA | NA | 0.625 | 33 |
| **Total Volume** | | | **8** | **422.4** |

3. Add 8uL of the RT mix to each of the top 48 wells. Each well should now contain a volume of 12uL.
4. Add 8uL of cells in 0.5x PBS+RI to each of the top 48 wells. Each well should now contain a volume of 20uL.
5. Add the plate into a thermocycler with the following protocol
    a. 50 C for 10 minutes
    b. Cycle 3 times:
        i. 8C for 12s
        ii. 15C for 45s
        iii. 20C for 45s
        iv. 30C for 30s
        v. 42C for 2 min
        vi. 50C for 3 min
    c. 50C for **5 min**
    d. 4C forever
6. Place the RT plate on ice.
7. Prepare 2 mL of 1x NEB buffer 3.1 with 20uL of Enzymatics RNase Inhibitor.
8. Transfer each RT reaction to a 15mL falcon tube (also on ice).
9. Add 9.6uL of 10% Triton-X100 to get a final concentration of 0.1%.
10. Centrifuge pooled RT reaction for 3 min at 500G.
11. Aspirate supernatant and resuspend into 2 mL of 1x NEB buffer 3.1 + 20uL Enzymatics RNase Inhibitor.

**Ligation Barcoding**

Make the following ligation master mix on ice:
*Note: Final concentration takes added volume of DNA barcodes into account. Concentrations of this mix is not the final concentration at time of barcoding*

| Reagent | Stock Concentration | Final Concentration | Volume (uL) |
|---|---|---|---:|
| Water | NA | NA | 1337.5 |
| T4 Ligase Buffer 10x | 10X | 1X | 500 |
| Enzymatics Rnase Inhibitor | 40 U/uL | 0.32 U/uL | 40 |
| Superase In | 20 U/uL | 0.05 U/uL | 12.5 |
| BSA | 20 mg/mL | 0.2 mg/mL | 50 |
| T4 DNA Ligase | 400 U/uL | 8 U/uL | 100 |
| **Total Volume** | | | **2040** |

1. Add the 2mL of cells in NEB buffer 3.1 into the ligation mix. The mix should now have a volume of 4.04 mL
2. Add the mix into a basin
3. Using a multichannel pipet, add 40 uL of ligation mix (with cells) into each well of the round 1 DNA barcode plate.
4. Cover the round 1 DNA barcode plate with an adhesive plate seal and incubate for **30 minutes at 37C** with gentle rotation (50 rpm).
5. Make the round 1 blocking solution and add it to a new basin

| Reagent | Stock Concentration | Final Concentration | Volume (uL) |
|---|---:|---:|---:|
| BC_0216 | 100 uM | 26.4 uM | 316.8 |
| 10x Ligase Buffer | 10X | 2.5X | 300 |
| Water | NA | NA | 583.2 |
| **Final Volume** | | | **1200 uL** |

6. Remove the round 1 DNA barcoding plate from the incubator and remove the cover.
7. Using a multichannel pipet, add 10 uL of the round 1 blocking solution to each of the 96 wells in the round 1 DNA barcoding plate.
8. Cover the round 1 DNA barcode plate with an adhesive plate seal and incubate for **30 minutes at 37C** with gentle rotation (50 rpm).

9. Remove round 1 DNA barcoding plate from the incubator, remove cover, and pool all cells into a new basin.
10. Pass all the cells from this basin through a 40 um strainer into another basin.
    - See note on step 4 of Fixation and Permeabilization.
11. Add 100 uL of T4 DNA ligase to the basin and mix by pipetting ~20 times.
12. Using a multichannel pipette, add 50 uL of cell/ligase solution into each well of the round 2 DNA barcode plate.
13. Cover the round 2 DNA barcode plate with an adhesive plate seal and incubate for **30 minutes at 37C** with gentle rotation (50 rpm).
14. Make the round 2 blocking solution and add it to a new basin

| Reagent | Stock Concentration | Final Concentration | Volume (uL) |
|---|---:|---:|---:|
| BC_0066 | 100 uM | 11.5 uM | 369 |
| EDTA | 0.5 M | 125 mM | 800 |
| Water | NA | NA | 2031 |
| **Final Volume** | | | **3200 uL** |

15. Remove the round 2 DNA barcoding plate from the incubator and remove the cover.
16. Using a multichannel pipet, add 20 uL of the round 2 blocking and termination solution to each of the 96 wells in the round 2 DNA barcoding plate.
17. Pool all cells into a new basin. (no incubation for the final blocking step)
18. Pass all the cells from this basin through a 40 um strainer into a 15 mL falcon tube.
    - See the note for step 4.
19. Count cells on a flow cytometer. Make sure cells are well mixed before aliquoting sample for counting.

**Lysis**

1. Make the 2X lysis buffer:

| Reagent | Stock Concentration | Final Concentration (2X) | Volume (mL) |
|---|---|---|---:|
| Tris, pH 8.0 | 1 M | 20 mM | 0.5 |
| NaCl | 5 M | 400 mM | 2 |
| EDTA, pH 8.0 | 0.5 M | 100 mM | 5 |
| SDS | 10% | 4.4 % | 11 |
| Water | NA | NA | 6.5 |
| **Final Volume** | | | **25** |

2. If white precipitate appears, warm at 37C until precipitate is back in solution (roughly 10-15 min).
3. Make the following wash buffer:

| Reagent | Volume (uL) |
|---|---|
| 1X PBS | 4000 |
| 10 % Triton X-100 | 40 |
| Superase In Rnase Inhibitor | 10 |
| **Final Volume** | **4050** |

4. Add 70ul of 10% triton to the cells. (~0.1% final conc.)
5. Centrifuge for 5 min at 1000G in 15ml tube.

   Note: The pellet for the steps below will be very small and it may not be visible.
6. Aspirate supernatant, leave ~30ul to avoid removing pellet.
   a. If possible, remove as much supernatant as possible with 20uL pipet.
7. Resuspend with 4 mL of wash buffer.
8. Centrifuge for 5 min at 1000G.
9. Aspirate supernatant and resuspend in 50ul 1x PBS + RI.
10. Dilute 5ul into 195uL of 1x PBS and count via flow cytometry.
    - Or take 5ul into 5ul of 1x PBS and count on hemocytometer (it can be hard to distinguish debris from cells).
11. Determine how many sublibraries you would like to generate (# sublibraries= # tubes needed), and how many cells you would like to have for each of these sublibraries.
12. Aliquot the desired number of cells for each sublibrary into new 1.7mL tubes. Add 1x PBS to each tube to a final volume of 50uL.
13. Add 50uL of 2x Lysis buffer to each tube.
14. Add 10uL of Proteinase K (20mg/mL) to each lysate.
15. Incubate at 55C for 2 hrs with shaking at 200rpm.
16. Stopping point: Freeze lysate(s) at -80C.

**Prepare buffers**

First make the following stock solutions:

100mM PMSF (resuspended in isopropanol)

| 2x B&W | |
|---|---|
| **Reagents** | **Volume** |
| 1M Tris-HCl pH 8.0 | 500uL |
| 5M NaCl | 20ml |
| EDTA, 0.5M | 100ul |
| Nuclease Free Water | 29.4ml |
| **Total** | **50mL** |

| 1x B&W-T | |
|---|---|
| **Reagents** | **Volume** |
| 1M Tris-HCl pH 8.0 | 100uL |
| 5M NaCl | 4ml |
| EDTA, 0.5M | 20ul |
| Tween 20 10% | 100ul |
| Nuclease Free Water | 15.78ml |
| **Total** | **20mL** |

Then make the following smaller aliquots (with added RNase inhibitor):

**1x B&W-T + RI:**

| | Volume per Number of Samples (uL) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Reagent** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| 1xB&W-T | 3600.0 | 4200.0 | 4800.0 | 5400.0 | 6000.0 | 6600.0 | 7200.0 | 7800.0 |
| SUPERase In | 5.0 | 5.8 | 6.7 | 7.5 | 8.3 | 9.2 | 10.0 | 10.8 |
| **Final Volume** | **3605.0** | **4205.8** | **4806.7** | **5407.5** | **6008.3** | **6609.2** | **7210.0** | **7810.8** |

**2x B&W + RI:**

| | Volume per Number of Samples (uL) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Reagent** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| 2xB&W | 110.0 | 220.0 | 330.0 | 440.0 | 550.0 | 660.0 | 770.0 | 880.0 |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| SUPERase In | 2.0 | 4.0 | 6.0 | 8.0 | 10.0 | 12.0 | 14.0 | 16.0 |
| **Final Volume** | **112.0** | **224.0** | **336.0** | **448.0** | **560.0** | **672.0** | **784.0** | **896.0** |

**Tris-T + RI:**

|  | Volume per Number of Samples (uL) |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| **Reagent** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| 10mM Tris-HCl (pH 8.0) | 600.0 | 1200.0 | 1800.0 | 2400.0 | 3000.0 | 3600.0 | 4200.0 | 4800.0 |
| Tween-20 (10%) | 6.0 | 12.0 | 18.0 | 24.0 | 30.0 | 36.0 | 42.0 | 48.0 |
| SUPERase In | 1.5 | 3.0 | 4.5 | 6.0 | 7.5 | 9.0 | 10.5 | 12.0 |
| **Final Volume** | **607.5** | **1215.0** | **1822.5** | **2430.0** | **3037.5** | **3645.0** | **4252.5** | **4860.0** |

**Purification of cDNA**

*Note: We performed agitation steps on a vortexer with a foam 1.7mL tube holder on a low setting (2/10).*

*Wash MyOne C1 Dynabeads*

1. For each lysate to be processed, add 44uL of MyOne C1 Dynabeads to a 1.5 mL tube (eg, 1 lysate=44uL, 2 lysates = 88uL, 3 lysates = 132ul etc)
2. Add 800uL of 1xB&W-T buffer
3. Place sample against a magnetic rack and wait until liquid becomes clear (1-2 min).
4. Remove supernatant and resuspend beads in 800uL of 1xB&W-T buffer.
5. Repeat steps 3-4 two more times for a total of 3 washes.
6. Place sample against a magnetic rack and wait until liquid becomes clear.
7. Resuspend beads in 100uL (per sample) 2xB&W buffer + RI.

**Sample Binding to Streptavidin:**

1. Add 5uL of 100uM PMSF (resuspended in isopropanol) to each sample and leave at room temperature for 10 min.
2. Add 100ul of resupended C1 beads to each tube.
3. To bind cDNA to C1 beads, agitate at room temperature for 60 min.
4. Place sample against a magnetic rack and wait until liquid becomes clear (1-2 min).
5. Remove supernatant and resuspend beads in 250uL of 1xB&W-T +RI
6. Agitate beads for 5 min at room temperature.
7. Repeat steps 5 and 6.
8. Remove supernatant and resuspend beads in 250 uL of 10mM Tris-T + RI
9. Agitate beads for 5 min at room temperature.
10. Leave beads in final wash solution on ice.

*Template Switch*

Prepare the following mix depending on the number of samples:

| | Volume per Number of Samples (uL) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Reagent** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| Water | 88.0 | 176.0 | 264.0 | 352.0 | 440.0 | 528.0 | 616.0 | 704.0 |
| Maxima RT Buffer | 44.0 | 88.0 | 132.0 | 176.0 | 220.0 | 264.0 | 308.0 | 352.0 |
| Ficoll PM-400 (20%) | 44.0 | 88.0 | 132.0 | 176.0 | 220.0 | 264.0 | 308.0 | 352.0 |
| 10mM dNTPs (each, total is 40mM) | 22.0 | 44.0 | 66.0 | 88.0 | 110.0 | 132.0 | 154.0 | 176.0 |
| RNase Inhibitor | 5.5 | 11.0 | 16.5 | 22.0 | 27.5 | 33.0 | 38.5 | 44.0 |
| TSO (BC_0127) | 5.5 | 11.0 | 16.5 | 22.0 | 27.5 | 33.0 | 38.5 | 44.0 |
| Maxima RT RnaseH Minus Enzyme | 11.0 | 22.0 | 33.0 | 44.0 | 55.0 | 66.0 | 77.0 | 88.0 |
| **Total** | **220.0** | **440.0** | **660.0** | **880.0** | **1100.0** | **1320.0** | **1540.0** | **1760.0** |

1. Place sample against a magnetic rack and wait until liquid becomes clear.
2. With sample still on magnetic rack, remove supernatant and wash with 250uL of water (do not resuspend beads this time).
3. Resuspend sample in 200ul of Template Switch Mix.
4. Incubate at room temp for 30 min with agitation or rolling.
5. Incubate at 42C for 90 min with agitation or rolling (we shook in incubator at 100 rpm).
6. Potential Stopping Point. If stopping perform the following (otherwise skip to next section):
    a. Place sample against a magnetic rack and wait until liquid becomes clear.
    b. Resuspend in 250uL Tris-T.

**cDNA Amplification**

Prepare the following PCR mix depending on the number of samples:

| | Volume per Number of Samples (uL) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Reagent** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| Kapa Hifi 2x Master Mix | 121.00 | 242.00 | 363.00 | 484.00 | 605.00 | 726.00 | 847.00 | 968.00 |
| BC_0108 (10uM) | 9.68 | 19.36 | 29.04 | 38.72 | 48.40 | 58.08 | 67.76 | 77.44 |
| BC_0062 (10uM) | 9.68 | 19.36 | 29.04 | 38.72 | 48.40 | 58.08 | 67.76 | 77.44 |
| Water | 101.64 | 203.28 | 304.92 | 406.56 | 508.20 | 609.84 | 711.48 | 813.12 |
| **Total** | **242.0** | **484.0** | **726.0** | **968.0** | **1210.0** | **1452.0** | **1694.0** | **1936.0** |

1. Place sample against a magnetic rack and wait until liquid becomes clear.
2. With sample against magnet wash with 250uL nuclease-free water (do not resuspend).
3. Resuspend sample with 220uL PCR mix and split equally into 4 different PCR tubes.
4. Run the following thermocycling program:
    a. 95C 3 min
    b. 98C 20s
    c. 65C 45s
    d. 72C 3min
    e. Repeat (b-d) 4x (5 total cycles)
    f. 4C hold.
5. Combine all 4 reactions into a single 1.7mL tube. Make sure to resuspend any beads that may be stuck to the bottom or sides of the PCR tubes before combining reactions.
6. Place sample against a magnetic rack and wait until liquid becomes clear.
7. Transfer 200uL of supernatant to 4 optical grade qPCR tubes (50uL in each tube).
8. Add 2.5uL of 20x evagreen to each qPCR tube.
9. Run the following qPCR program (make sure to remove samples, once signal starts to leave exponential phase to prevent overamplification).
    a. 95C 3 min
    b. 98C 20s
    c. 67C 20s
    d. 72C 3min
    e. Repeat (b-d) until signal plateaus out of exponential amplification
    f. 72C 5 min
    g. 4C hold
10. Optional: Run an agarose gel or bioanalyze resulting qPCR. There will likely be a combination of cDNA and dimer present.

*SPRI size selection (0.8x)*

1. Combine qPCR reactions into a single tube.
2. Take out 180 uL of the pooled qPCR reaction and place in new 1.7 mL tube
3. Add 144uL of Kapa Pure Beads to tube and vortex briefly to mix. Wait 5 min to bind DNA.
4. Place tube against magnetic rack and wait until liquid becomes clear.
5. Remove the supernatant.
6. With tubes still on magnetic rack, wash with 750uL 85% ethanol. Do not resuspend beads.
7. Repeat step 6.
8. Remove ethanol and air dry bead (~5min). To not let beads overdry and crack.
9. Resuspend beads from each tube in 20uL of water. Once beads are fully resuspended in the water, incubate the tube at 37C for 10 min.
10. Bind tubes against magnetic rack and wait until liquid becomes clear.
11. Transfer 18.5uL of elutant into a new optical grade PCR tube.

12. Run a bioanalyzer trace on 10 uL of the elutant
13. If no dimer is present after size selection, jump directly to “Tagmentation and Illumina Amplicon Generation” section. If dimer is still present, proceed to step 14 to perform a second amplification and size selection step. This may be necessary for cells with low RNA content, but should not be necessary for cells with high RNA content (eg, HeLa-S3, NIH/3T3, etc.).

**Tagmentation and Illumina Amplicon Generation**

1. Quibit amplified cDNA and dilute to 0.12ng/uL.
2. Preheat a thermocycler to 55 degrees.
3. For each sample, combine 600 pg of purified cDNA with H2O in a total volume of 5 ul.
4. To each tube, add 10 ul of Nextera TD buffer and 5 ul of Amplicon Tagment enzyme (the total volume of the reaction is now 20 ul). Mix by pipetting ~5 times. Spin down.
5. Incubate at 55 C for 5 minutes.
6. Add 5 ul of Neutralization Buffer. Mix by pipetting ~5 times. Spin down. Bubbles are normal.
7. Incubate at room temperature for 5 minutes.
8. Add to each PCR tube in the following order:
    1. 15 ul of Nextera PCR mix
    2. 8 ul H2O
    3. 1 ul of 10 uM (N7 indexed primer, one of BC_0076-BC_0083)
    4. 1 ul of 10 uM Nextera (BC_0118) N501 oligo
9. Run the following thermocycling program:
    1. 95 C 30 sec
    2. 12 cycles of:
        1. 95 C 10 seconds
        2. 55 C 30 seconds
        3. 72 C 30 seconds
    3. Then: 72 C 5 minutes 4 C forever
10. Transfer 40ul out of the 50uL reaction to a 1.7mL tube.
11. Add 28uL of Kapa Pure beads to do a 0.7x cleanup. Elute in 20ul.
12. Bioanalyze resulting sample and quibit before sequencing. See lane 1 on figure 1 for expected size distribution.

**Illumina Sequencing**

1. Use a paired-end sequencing run with a 150 bp kit.
2. Set read1 to 66 nt (transcript sequence)
3. Set read2 to 94 nt (cell-specific barcodes and UMI)
4. Include a 6nt read 1 index to ready sublibrary indices (this is 4th round of barcodes).

### 7.5.3 Evercode Whole Transcriptome

Evercode cell fixation and single-cell whole-transcriptome kit protocols. Protocols were downloaded from the Parse Biosciences user portal on the 11th of January 2022.

### 7.5.4 Drop-seq alignment cookbook

The Drop-seq Alignment Cookbook was downloaded from https://github.com/broadinstitute/Drop-seq on the 28th of April 2020.

# Drop-seq Core Computational Protocol

version 2.0.0 (9/28/18)

**James Nemesh**
**Steve McCarroll's lab, Harvard Medical School**

## Introduction

The following is a manual for using the software we have written for processing Drop-seq sequence data into a "digital expression matrix" that will contain integer counts of the number of transcripts for each gene, in each cell. This software pipeline performs many analyses including massive de-multiplexing of the data, alignment of reads to a reference genome, and processing of cellular and molecular barcodes.

Drop-seq sequencing libraries produce paired-end reads: read 1 contains both a cell barcode and a molecular barcode (also known as a UMI); read 2 is aligned to the reference genome. This document provides step-by-step instructions for using the software we have developed to convert these sequencing reads into a digital expression matrix that contains integer counts of the number of transcripts for each gene, in each cell.

We may release updates to this manual as we learn from users' experiences. If a revision simply contains additional hints or advice or detail, then we will update the date on the protocol but not the version number. Whenever we implement a substantive change to the software or protocol, we will increment the version number.

We hope this is helpful and that you are soon generating exciting data with Drop-seq.

## Introduction V2:

There are a number of enhancements to the Drop-seq platform that come with version 2.0: new methods to clean up the cell barcodes from bead synthesis errors and PCR errors result in less clutter in the data when trying to decide which cell barcodes are truly cells. We've also enhanced Digital expression to be more flexible in how it interprets gene annotations, allowing the program to extract both intronic DGE data as well as the typical coding+utr data. Read on to find out about new Drop-seq program capabilities in-line with the rest of the documentation.

## Drop-seq Software and Hardware Requirements

The Drop-seq software provided is implemented entirely in Java. This means it will run on a huge number of devices that are capable of running Java, from large servers to laptops. We require 4 gigabytes of memory for each program to run, which is also sufficient for Picard programs we use as part of alignment and analysis. Disk space will be determined by your data size plus the meta-data and aligner index. 50 gigabytes of disk space will be sufficient to store our meta data plus a STAR index.

**Overview of Alignment**

The raw reads from the sequencer must be converted into a Picard-queryname-sorted BAM file for each library in the sequencer run. Since there are many sequencers and pipelines available to do this, we leave this step to the user. For example, we use either Picard IlluminaBasecallsToSam (preceded by Picard ExtractIlluminaBarcodes for a library with sample barcodes); or Illumina's bcl2fastq followed by Picard FastqToSam**.** Once you have an unmapped, queryname-sorted BAM, you can follow this set of steps to align your raw reads and create a BAM file that is suitable to produce digital gene expression (DGE) results.

1. Unmapped BAM -> aligned and tagged BAM
    a. Tag cell barcodes
    b. Tag molecular barcodes
    c. Trim 5' primer sequence
    d. Trim 3' polyA sequence
    e. SAM -> Fastq
    f. STAR alignment
    g. Sort STAR alignment in queryname order
    h. Merge STAR alignment tagged SAM to recover cell/molecular barcodes
    i. Add gene/exon and other annotation tags
    j. Barcode Repair
        i. Repair substitution errors (DetectBeadSubstitutionErrors)
        ii. Repair indel errors (DetectBeadSynthesisErrors)

**A walkthrough of the alignment process**

Let's walk through these steps to help you build intuition about how reads are manipulated - later parts of this document will detail the software and invocations necessary to carry out these operations. First, you'll take your Drop-seq experiment and put it on a sequencer. The sequencer will gather data from both reads of the read pair. Read one is a *barcoded read,* containing the cell and molecular barcodes that will later identify this read as coming from a particular transcript on a particular cell. Read two is the *biological read*, which contains a portion of the sequence of the transcript observed.

First, you'll make a BAM file out of this data so that these two reads are in the same place. Then, we'll transfer information from the barcoded read over to the BAM record containing the genome read as a set of BAM tags. The first 12 bases of the barcoded read contain the cell barcode, so we'll copy those bases over to a BAM tag (XC) on the genome read. Then we'll take the next 8 bases containing the molecular barcode and copy them over as another BAM tag (XM). Since we're now extracted all the information out of the barcoded read, we discard the read, converting the BAM to single-ended reads.

If a barcoded read has low quality base, both the barcoded read and genome read are purged at this point. This makes life a lot easier for us in the future, as we don't have to track the barcoded read to know the origins of any genome read.

After this, we clean up the genome read with a few processes. The 5' adapter is detected and trimmed, as are 3' poly A tails. We call this final cleaned-up BAM the unaligned BAM. Then, we want to align these single-ended genome reads to the genome using STAR. To do this, we extract the fastq file containing single-ended reads from our genome read BAM file and run STAR. After STAR is done aligning reads, we now know where the genome reads align, but we've lost track of what cell and molecular barcodes these reads have. This information is recovered by merging the BAM tags from the unaligned BAM to the aligned reads from STAR. We then add additional annotation to the reads that is dependent on the genome read, such as any genes or exons that the read overlaps. Finally, we check for bead synthesis errors and repair them if possible.

The next sections will explain the metadata needed to follow this workflow, as well as explain each of the programs that have been developed to run these steps. Some of these programs are developed by us, and others take advantage of existing Picard Tools or aligners like STAR.

**Metadata**

To follow this set of processes from raw unaligned reads to an aligned BAM, it's necessary to have a number of different metadata files. These provide information about the sequence of the organism(s) you're running your experiment on, as well as genomic features like genes,transcripts, and exons that help extract DGE data from the reads.

We organize our metadata using a set of conventions we suggest you follow, as it makes it easier to keep track of what files are used for particular processes. In the software section, we'll refer to these files using these conventions.

The first convention is that we establish a root name for all of our files that encodes information about the organism and the genome build used to derive that metadata. For example, mm10 is the Dec. 2011 *Mus musculus* assembly. All files for mouse use this as the root name, followed by a ".", then the type of file.

**metadata file types:**

- *fasta*: The reference sequence of the organism. Needed for most aligners.
- *dict*: A dictionary file as generated by Picard's CreateSequenceDictionary. Needed for Picard Tools.

- gtf: The principle file to determine the location of genomic features like genes, transcripts, and exons. Many other metadata files we use derive from this original file. We download our GTF files from ensembl, which has a handy description of the file format here. Ensembl has a huge number of prepared GTF files for a variety of organisms here.
- refFlat: This file contains a subset of the the same information in the GTF file in a different format. Picard tools like the refFlat format, so we require this as well. To make life easy, we provide a program ConvertToRefFlat that can convert files from GTF format to refFlat for you.
- genes.intervals: The genes from the GTF file in interval list format. This file is optional, and useful if you want to go back to your BAM later to see what gene(s) a read aligns to.
- exons.intervals: The exons from the GTF file in interval list format. This file is optional, and useful if you want to go back to your BAM and view what exon(s) a read aligns to.
- rRNA.intervals: The locations of ribosomal RNA in interval list format. This file is optional, but we find it useful to later assess how much of a dropseq library aligns to rRNA.
- reduced.gtf: This file contains a subset of the information in the GTF file, but in a far more human readable format. This file is optional, but can be generated easily by the supplied ReduceGTF program that will take a GTF file as input.

On the Drop-Seq website you will find a set of pre-made meta data for human, mouse and human/mouse experiments.

**Premade Meta Data links @GEO.**
MIXED MOUSE HUMAN

**MetaData Creation Programs**
A few files and required to generate meta data: a GTF file, and a fastq file. From these two files we can derive various other files needed by the the Drop-seq software.

**CreateSequenceDictionary**
The first file needed is the sequence dictionary. This is a list of the contigs in the fastq file and their lengths.

```
java -jar /path/to/picard/picard.jar CreateSequenceDictionary
REFERENCE=my.fasta
OUTPUT= my.dict
SPECIES=species_name
```

**ConvertToRefFlat**
The next file is the refFlat file, which is generated using the sequence dictionary generated above.

```
ConvertToRefFlat
ANNOTATIONS_FILE=my.gtf
SEQUENCE_DICTIONARY=my.dict
OUTPUT=my.refFlat
```

### ReduceGTF

The may be useful if you need an easy to parse version of your annotations in a language like R, and is also used to generate the other metadata.

```
ReduceGTF
SEQUENCE_DICTIONARY=my.dict
GTF=my.gtf
OUTPUT=my.reduced.gtf
```

### CreateIntervalsFiles

As a last step, we create interval files needed for various programs in the Drop-seq pipeline. This program generates a number of interval files for genes, exons, consensus introns, rRNA, and mt. The example below uses the human MT contig name, but if you use a different organism you should set that argument appropriately.

```
CreateIntervalsFiles
SEQUENCE_DICTIONARY=my.dict
REDUCED_GTF=my.reduced.gtf
PREFIX=my
OUTPUT=/path/to/output/files
MT_SEQUENCE=MT
```

### MetaData Generation Pipeline

We've provided a shell script to generate new meta data sets for single organism data in the distribution. This script is called create_Drop-seq_reference_metadata.sh, and the options for the program can be accessed by running with the -h option:

```
/path/to/dropseq_tools/create_Drop-seq_reference_metadata.sh -h
```

### Alignment Pipeline Programs

On the Drop-seq website you will find a zipfile containing the programs described below. The zipfile also contains a script Drop-seq_alignment.sh that executes the process described below. Because of differences in computing environments, this script is not guaranteed to work for all users. However, we hope it will serve as an example of how the various programs should be invoked.

### TagBamWithReadSequenceExtended

This Drop-seq program extracts bases from the cell/molecular barcode encoding read (BARCODED_READ), and creates a new BAM tag with those bases on the *genome read*. By default, we use the BAM tag XM for molecular barcodes, and XC for cell barcodes, using the TAG_NAME parameter.

This program is run once per barcode extraction to add a tag. On the first iteration, the cell barcode is extracted from bases 1-12. This is controlled by the BASE_RANGE option. On the second iteration, the molecular barcode is extracted from bases 13-20 of the barcode read. This program has an option to drop a read (DISCARD_READ), which we use after both barcodes have been extracted, which makes the output BAM have unpaired reads with additional tags.

Additionally, this program has a BASE_QUALITY option, which is the minimum base quality of all bases of the barcode being extracted. If more than NUM_BASES_BELOW_QUALITY bases falls below this quality, the read pair is discarded.

**Example Cell Barcode:**

```
TagBamWithReadSequenceExtended
INPUT=my_unaligned_data.bam
OUTPUT=unaligned_tagged_Cell.bam
SUMMARY=unaligned_tagged_Cellular.bam_summary.txt
BASE_RANGE=1-12
BASE_QUALITY=10
BARCODED_READ=1
DISCARD_READ=False
TAG_NAME=XC
NUM_BASES_BELOW_QUALITY=1
```

**Example Molecular Barcode:**

```
TagBamWithReadSequenceExtended
INPUT=unaligned_tagged_Cell.bam
OUTPUT=unaligned_tagged_CellMolecular.bam
SUMMARY=unaligned_tagged_Molecular.bam_summary.txt
BASE_RANGE=13-20
BASE_QUALITY=10
BARCODED_READ=1
DISCARD_READ=True
TAG_NAME=XM
NUM_BASES_BELOW_QUALITY=1
```

**FilterBam:**

This Drop-seq program is used to remove reads where the cell or molecular barcode has low quality bases. During the run of TagBamWithReadSequenceExtended, an XQ tag is added to each read to represent the number of bases that have quality scores below the BASE_QUALITY threshold. These reads are then removed from the BAM.

**Example:**

```
FilterBam
TAG_REJECT=XQ
```

```
INPUT=unaligned_tagged_CellMolecular.bam
OUTPUT=unaligned_tagged_filtered.bam
```

### TrimStartingSequence

This Drop-seq program is one of two sequence cleanup programs designed to trim away any extra sequence that might have snuck it's way into the reads. In this case, we trim the SMART Adapter that can occur 5' of the read. In our standard run, we look for at least 5 contiguous bases (NUM_BASES) of the SMART adapter (SEQUENCE) at the 5' end of the read with no errors (MISMATCHES) , and hard clip those bases off the read.

**Example:**

```
TrimStartingSequence
INPUT=unaligned_tagged_filtered.bam
OUTPUT=unaligned_tagged_trimmed_smart.bam
OUTPUT_SUMMARY=adapter_trimming_report.txt
SEQUENCE=AAGCAGTGGTATCAACGCAGAGTGAATGGG
MISMATCHES=0
NUM_BASES=5
```

### PolyATrimmer

This Drop-seq program is the second sequence cleanup program designed to trim away trailing polyA tails from reads. It searches for at least 6 (NUM_BASES) contiguous A's in the read with 0 mismatches (MISMATCHES), and hard clips the read to remove these bases and all bases 3' of the polyA run.

**Example:**

```
PolyATrimmer
INPUT=unaligned_tagged_trimmed_smart.bam
OUTPUT=unaligned_mc_tagged_polyA_filtered.bam
OUTPUT_SUMMARY=polyA_trimming_report.txt
MISMATCHES=0
NUM_BASES=6
USE_NEW_TRIMMER=true
```

### SamToFastq

Now that your data has had the cell and molecular barcodes extracted, the reads have been cleaned of SMARTSeq primer and polyA tails, and the data is now unpaired reads, it's time to align. To do this, we extract the FASTQ files using Picard's SamToFastq program.

**Example:**

```
java -Xmx4g -jar /path/to/picard/picard.jar SamToFastq
INPUT=unaligned_mc_tagged_polyA_filtered.bam
FASTQ=unaligned_mc_tagged_polyA_filtered.fastq
```

### Alignment - STAR

We use STAR as our RNA aligner. The manual for STAR can be found here. There are many potential aligners one could use at this stage, and it's possible to substitute in your lab's favorite. We haven't tested other aligners in methodical detail, but all should produce valid BAM files that can be plugged into the rest of the process detailed here.

If you're unsure how to create an indexed reference for STAR, please read the STAR manual.
Below is a minimal invocation of STAR. Since STAR contains a huge number of options to tailor alignment to a library and trade off sensitivity vs specificity, you can alter the default settings of the algorithm to your liking, but we find the defaults work reasonably well for Drop-seq. Be aware that STAR requires roughly 30 gigabytes of memory to align a single human sized genome, and 60 gigabytes for our human/mouse reference.

**Example:**

```
/path/to/STAR/STAR
--genomeDir /path/to/STAR_REFERENCE
--readFilesIn unaligned_mc_tagged_polyA_filtered.fastq
--outFileNamePrefix star
```

### SortSam

This picard program is invoked after alignment, to guarantee that the output from alignment is sorted in queryname order. As a side bonus, the output file is a BAM (compressed) instead of SAM (uncompressed.)

**Example:**

```
java -Xmx4g -jar /path/to/picard/picard.jar SortSam
I=starAligned.out.sam
O=aligned.sorted.bam
SO=queryname
```

### MergeBamAlignment

This Picard program merges the sorted alignment output from STAR (ALIGNED_BAM) with the unaligned BAM that had been previously tagged with molecular/cell barcodes (UNMAPPED_BAM). This recovers the BAM tags that were "lost" during alignment. The REFERENCE_SEQUENCE argument refers to the fasta metadata file.

We ignore secondary alignments, as we want only the best alignment from STAR (or another aligner), instead of assigning a single sequencing read to multiple locations on the genome.

**Example:**

```
java -Xmx4g -jar /path/to/picard/picard.jar MergeBamAlignment
REFERENCE_SEQUENCE=my_fasta.fasta
UNMAPPED_BAM=unaligned_mc_tagged_polyA_filtered.bam
```

```
ALIGNED_BAM=aligned.sorted.bam
OUTPUT=merged.bam
INCLUDE_SECONDARY_ALIGNMENTS=false
PAIRED_RUN=false
```

**TagReadWithGeneExon**

This is a Drop-seq program that adds a BAM tag “GE” onto reads when the read overlaps the exon of a gene. This tag contains the name of the gene, as reported in the annotations file. You can use either a GTF or a RefFlat annotation file with this program, depending on what annotation data source you find most useful. This is used later when we extract digital gene expression (DGE) from the BAM.

**Example:**

```
TagReadWithGeneExon
I=merged.bam
O=star_gene_exon_tagged.bam
ANNOTATIONS_FILE=${refFlat}
TAG=GE
```

**Updates to TagReadWithGeneExon (V2)**

We have updated and re-written how reads are tagged with functional annotations in V 2.0 of the dropseq toolkit. In V1, reads received two BAM tags when a read overlapped the exon of a gene. The GE tag specified the gene that overlapped the read, while GS specified which strand the gene was on. This information allows DigitalExpression and other programs to decide if they want to consider reads that are on the same strand as the gene, or run without regard to strand.

A typical read on that overlaps a gene might have the following tags, indicating the read overlapped an exon of GENE_A, and was on the positive strand:

```
H53FWBGXX150403:1:11307:13550:9549	0	1	29658	1	60M	*	0
0	CTGCCTTCCCCTCAAGCTCAGGGCCAAGCTGTCCGCCAACCTCGGCTCCTCCGGGCAGCC
7FFFFFFFFFFFFFFFFFFFF.FFFFFFFFFFFAFFFFFFFFFFA.FFFF<FFFFAAAAA	XC:Z:TTGTCATGTCAC
GE:Z:GENE_A	XF:Z:CODING	PG:Z:STAR.1	RG:Z:H53FW.1	H:i:4	NM:i:0	XM:Z:GCAAACCT	UQ:i:0
AS:i:59 GS:Z:+
```

This functionality has been retained exactly as it was implemented in a newly distributed program TagReadWithGeneExonFunction. We’ve done this in case other users need to retain backwards compatibility with any analysis they may have implemented.

**TagReadWithGeneFunction** (replacement for TagReadWithGeneExon)

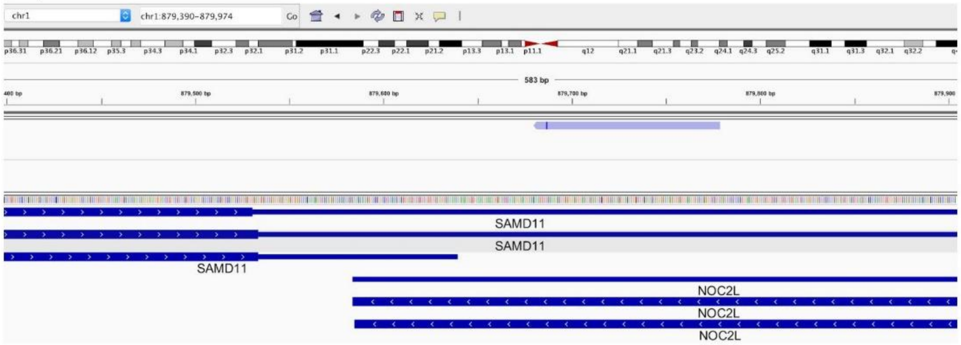
Our replacement for TagReadWithGeneExon is TagReadWithGeneFunction. This program provides a more flexible and informative set of tags for reads that allow downstream programs to measure not only digital expression of reads that overlap exons, but can leverage reads that introns as well. This program provides 3 tags for each read, **gn** [gene name], **gs** [gene strand] and **gf** [gene function]. These tags can have more than one value, and the values are comma separated. These tags can also co-exist with the original tagger (TagReadWithGeneExon) as the tag names are different, so if you use those tags for other purposes, you can tag your BAM with both taggers.

**Example Invocation (**The call to TagReadWithGeneFunction is the same as TagReadWithGeneExon)

```
TagReadWithGeneFunction
I=merged.bam
O=star_gene_exon_tagged.bam
ANNOTATIONS_FILE=${refFlat}
```

**Below is an example read using the new tagger:**

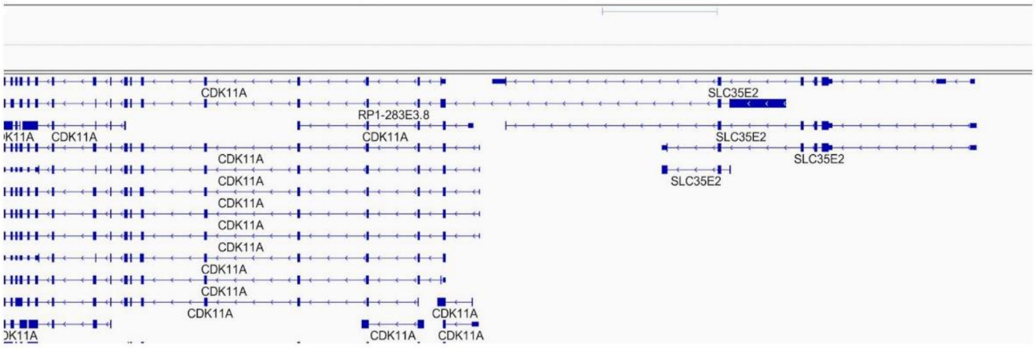
```
HFWN3DMXX:1:2133:1949:24283       16    1    879682  255    98M           *           0     0
AATTTCCAAAGACTTGGGGGAGTGAAGGCAGAGCCTGGTGCAGATGGACGAGGTCTGCAGACGGAGGGCAGAGGTGGTGGAAGGGGCCA
GGGGCCTGC     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
          XC:Z:CTACGTCCACATGACT   MD:Z:5T92            XF:Z:CODING        PG:Z:STAR.3W       RG:Z:HFWN3.1
XG:Z:SAMD11,NOC2L          NH:i:1  NM:i:1  XM:Z:GAAGGATAAA UQ:i:37 AS:i:94 gf:Z:CODING,UTR gn:Z:NOC2L,SAMD11
gs:Z:-,+
```



The gs, gn, and gf tags all have the same number of values. They are interpreted as a trio of values that describe the gene the read overlaps, the strand the gene is on, and the functional annotation of the gene at that position. In the example above, the read overlaps both **NOC2L** on the negative strand and completely overlaps an exon. The read also overlaps **SAMD11** on the positive strand. The read is on the negative strand (from the bitflag on the read of 16), so standard DGE would interpret this as expression of **NOC2L.**

**Example 2:**

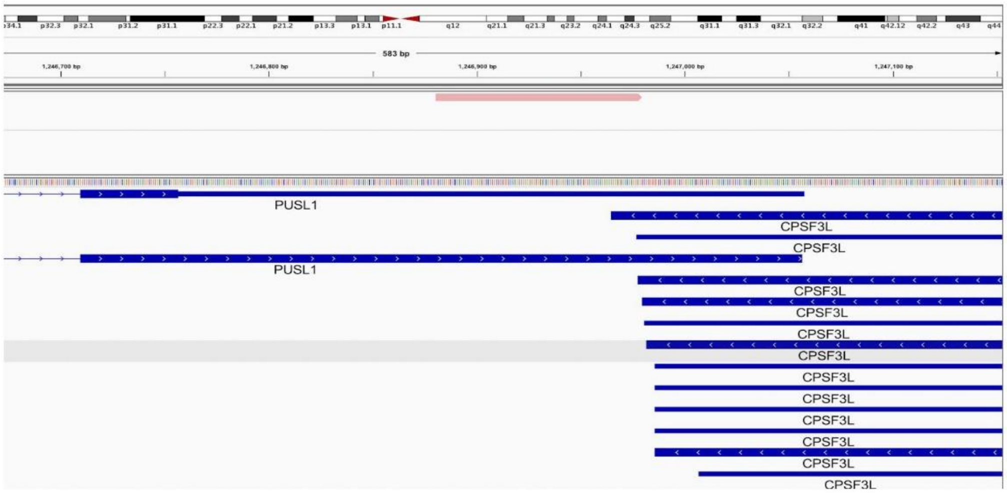
```
HFWNNDMXX:2:2220:15085:21292      16     1   1661100 255          2S20M5009N76M  *      0     0
CCGAGCCACCGCAGCCGGTCTTCTGAAAGTCACCGGGGAGATTTTCCCCATGAGGGCGTACGCCGTGACGCTCTGAAGGTGGAACAGGACT
CCGTCTG     FFFFFFFFFFFFFFFFFFF,FFFFFF:FFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFF:FFF:FFFFFFFFFFF
XC:Z:TCAGGATCAGCAGTTT   MD:Z:4T91              XF:Z:CODING
          PG:Z:STAR.3O        RG:Z:HFWNN.2.B  XG:Z:RP1-283E3.8,SLC35E2      NH:i:1  NM:i:1  XM:Z:ACATGCCGCG UQ:i:37
AS:i:91 gf:Z:CODING,INTRONIC,CODING,INTRONIC    gn:Z:RP1-283E3.8,RP1-283E3.8,SLC35E2,SLC35E2         gs:Z:-,-,-,-
```



This read is a bit more "interesting", due to the overlapping gene annotations. Both SLC35E2 and RP1-283E.8 appear to share the same exon, though in different splicing contexts. The read is mapped as a split read, where part of the read is gapped and splices to a different location, which is common when mapping exon-exon junctions. The other part of the read appears to splice in the middle of the intron for both genes. DGE will interpret this read as ambiguous, as it can be assigned to either gene.

**Example 3:**

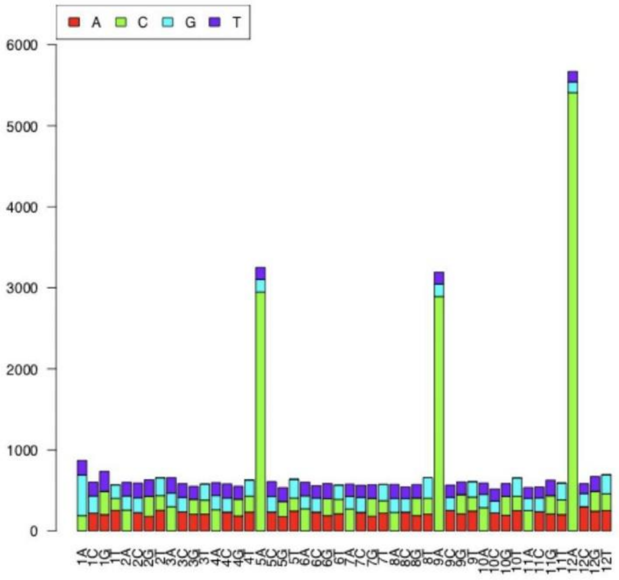
```
HFWNNDMXX:2:2114:15917:25019     0     1     1246881 255         98M    *                0            0
GCCCGGGTCCCAGCACCCTGGATGCCCGTCTCTGTCCCAGGCGGGATGGGGCACAGTGCAGGACACAGCCATGTACACCAAGAAGAGAGTA
CCAAGTA     F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
XC:Z:TGCCAAAGTCGCTTTC   MD:Z:98 XF:Z:CODING            PG:Z:STAR.15
          RG:Z:HFWNN.2.A  XG:Z:CPSF3L,PUSL1               NH:i:1  NM:i:0  XM:Z:GTATGATTGA UQ:i:0  AS:i:96
gf:Z:CODING,INTERGENIC,CODING   gn:Z:CPSF3L,CPSF3L,PUSL1         gs:Z:-,-,+
```



This read maps to two genes, CPSF3L and PUSL1. The read is on the positive strand, as is PUSL1, so the read would be assigned to that gene. If you were extracting expression on the opposite strand from the gene, then the read would be assigned both coding and intergenic (outside the bounds of the gene) portions of CPSF3L. Since the read wasn't assigned to only CODING+UTR portions of the transcript, under the standard functional types fort DGE, the read would be ignored. Reads must be entirely contained within the requested functional types to be counted. If you wanted to go wild, (and who doesn't?), you could request DGE to extract the antisense transcript and include intergenic regions by using STRAND_STRATEGY=ANTISENSE LOCUS_FUNCTION_LIST=INTERGENIC. That would generate a UMI for this cell barcode on the CPSF3L gene. Hopefully this demonstrates the flexibility of our new approach to tagging reads.

**DetectBeadSubstitutionErrors - Detecting and repairing substitution errors in cell barcodes**
In previous chemgenes bead lots, we have observed non-random patterns of substitution changes at hamming edit distance=1 between pairs of barcodes that appear to be related. Given many of these cell barcodes may in fact be multiple cell barcodes that reside on a single physical bead, it makes sense to combine the reads across these barcodes. However, barcodes can appear to be related to each other at hamming distance=1 by chance due to sequencing/PCR errors, or because we sample a significant subset of the total available cell barcode space and happen to observe two barcodes that are truly independent, but very similar in sequence. The challenge is to combine related cell barcodes together that have arisen from the same bead, while avoiding capricious collapse of other pairs of cell barcodes that are related by chance.

Errors that occur at the synthesis level ought to be systemic - the same barcode position and base substitution pattern from an intended sequence to a related sequence (for example: position 5 of many cell barcodes change from A to C) should be consistent across an entire experiment at some substitution rate. The higher the rate of substitution, the more frequently the related sequence will be observed. To determine where these events take place, we survey the set of cell barcodes with at least 20 transcripts, and exhaustively look for pairs of barcodes that are related at hamming distance=1. For each pair of barcodes, we assume the more frequently observed barcode to be the "intended" sequence. We filter pairs of barcodes so that smaller barcodes are unambiguously related to one and only one intended sequence. By building up this set of barcodes, we can observe patterns in the substitution events that are biased to certain bases and positions in the synthesis reaction.



The X axis describes the position of the substitution and the intended base. The color of the stacked column indicates the base that is substituted at that position.

There are clear patterns of substitution events that occur in the data above at some positions, as well as a lower level of stochastic changes that occur at every position. By looking for a dominant base change (>80% of events) at each position and intended sequence, a subset of all possible base substitution events can be selected. In this case, there are A->C substitution events at bases 1,2,3,4,5,9,12. We select barcodes that contain these specific substitution bases and positions to perform repair at. We remove the smaller neighbors that are related to multiple intended sequences, or do not fit the substitution pattern observed above. These patterns are discovered on an experiment by experiment basis as part of the cleanup process.

These results are repeatable across many experiments using the same bead lot, and differ across bead lots.

**Example:**
**DetectBeadSubstitutionErrors**
I=my.bam
O=my_clean_subtitution.bam
OUTPUT_REPORT=my_clean.substitution_report.txt

**DetectBeadSynthesisErrors - Detecting and repairing barcode indel synthesis errors**

In June 2015, we noticed that a recently purchased batch of ChemGenes beads generated a population of cell barcodes (about 10-20%) with sequences that shared the first 11 bases, but differed at the last base. These same cell barcodes also had a very high percentage of the base "T" at the last position of the UMI. Based on these observations, we concluded that a percentage of beads in the lot had not undergone all twelve split-and-pool bases (perhaps they had stuck to some piece of equipment or container, and the been re-introduced after the missing synthesis cycle). Thus, the 20-bp Read 1 contained a mixed base at base 12 (in actuality, the first base of the UMI) and a fixed T-base at base 20 (in actuality, the first base of the polyT segment).

To correct for this, we generated DetectBeadSynthesisErrors, which identifies cell barcodes with aberrant "fixed" UMI bases. If only the last UMI base is fixed as a T, the cell barcode is corrected (the last base is trimmed off) and all cell barcodes with identical sequence at the first 11 bases are merged together. If any other UMI base is fixed, the reads with that cell barcode are discarded.

**UPDATE**

More recently (Fall 2017), we observed that it was possible to not only discover cell barcodes where the UMIs were biased to T at the last base, but in many cases to discover what the original barcode sequence was. During synthesis, these incorporation errors often occur incompletely - a base is missing at a certain position, and elsewhere in the experiment, it's possible to recover the original "intended" sequence as another cell barcode where the UMIs do not experience the T bias, and the 2 cell barcodes are related by a 1 base pair insertion/deletion event. The intended sequence has 4 "neighbor" barcodes at an indel distance of 1 (and if they have few UMIs, not all 4 neighbors will be detected), and those neighbors are all related to each other by a substitution edit distance of 1 at the last base of their sequence.. Each of these neighbors also has high T bias at the last base of their UMIs. By looking at the number of UMIs observed by the intended and neighbor sequences, it's possible to calculate the rate at which the base was not incorporated into the sequence. For example, if the intended sequence is quite small, and the neighbors are relatively large, then the base was not incorporated at a high rate.

This allows us to validate which UMI biased cell barcodes should properly be merged into an intended sequence, which is a more stringent repair process than what we'd previously employed. Because of this, we've removed the requirement to estimate the number of cells in the experiment. We repair all cell barcodes with at least 20 UMis.

Below is an example of a few intended sequence / neighbor barcodes and their relationships. When this software is run, these parameters (and many others) are emitted as a result. The neighbor sequences are colon separated.

| Example | Intended Sequence | Neighbor Sequences | Deleted base | Deleted base position | Non incorporation rate |
|---|---|---|---|---|---|
| 1 | TGATGCACGAGG | TGATCACGAGGA:TGATCACGAGGC:TGATCACGAGGG:TGATCACGAGGT | G | 5 | 0.01 |
| 2 | CTCCGAACTGCC | CTCCGAACGCCA:CTCCGAACGCCC:CTCCGAACGCCG:CTCCGAACGCCT | T | 9 | 0.95 |
| 3 | CCCTCGTTAGAT | CCTCGTTAGATA:CCTCGTTAGATC:CCTCGTTAGATT | C | 3 | 0 |
| 4 | <NA> | CCCGCAGCTTGA:CCCGCAGCTTGC:CCCGCAGCTTGG:CCCGCAGCTTGT | <NA> | <NA> | <NA> |

1. An intended sequence where all 4 neighbors are discovered. Non-incorporation rate is low, so the intended cell barcode has many UMIs relative to the neighbor cell barcodes.
2. An intended sequence where all 4 neighbors are discovered. Non-incorporation rate is high, so the intended cell barcode has few UMIS relative to the neighbor cell barcodes.
3. Only 3 neighbors are discovered. Fewer than 4 neighbors being discovered can occur when the neighbors have ~ 25 UMIs (the smallest cell barcode we look at is 20), and one of the neighbors has fewer UMIs so is not reported. The missing sequence is CCTCGTTAGATG, as A/C/T are found.
4. 4 related neighbors are found, but the intended sequence is not discovered. This occurs when the base non-incorporation rate is very high , and occurs in almost every copy of the barcode on the bead. Because the intended sequence is not discovered, the other properties can not be determined.

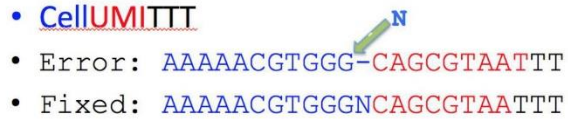
**Example:**

```
DetectBeadSynthesisErrors
I=my_clean_subtitution.bam
O=my_clean.bam
REPORT=my_clean.indel_report.txt
OUTPUT_STATS=my.synthesis_stats.txt
SUMMARY=my.synthesis_stats.summary.txt
PRIMER_SEQUENCE=AAGCAGTGGTATCAACGCAGAGTAC
```

This program reads in the BAM file, and looks at the distribution of bases at each position of all UMIs for a cell barcode. It detects unusual distributions of base frequency, where a base with >=80% frequency at any position is detected as an error. Barcodes with less than 20 total UMIs are ignored. There are a number of different errors that are categorized:

1. SYNTHESIS_MISSING_BASE - 1 or more bases missing from cell barcode, resulting in fixed T's at the end of UMIs. This counts the maximum number of fixed sequential T's in the UMIs at the end. This error type is cleaned up by the software for situations where there is a single base

missing, and is by far the most common error. The fix involves inserting an "N" base before the last cell barcode base, effectively shifting the reading frame back to where it should be. This will both collapse these beads back together in further analysis, as well as repair the UMIs for these bead barcodes. **[Note as of V2, we no longer insert an N into sequences to repair them if we're able to determine the intended sequence. In those cases we use the intended sequence instead.]**



- CellUMITTT
- Error: AAAAACGTGGG-CAGCGTAATTT
- Fixed: AAAAACGTGGGNCAGCGTAATTT

2. SINGLE_UMI_ERROR - At each position of the UMIs, the base distribution is highly skewed, i.e. at each position, a single base appears in >= 80% of the UMIs for that cell. There's no fix for this currently. Cell barcodes with this property are dropped. These cells have the interesting property that the number of genes and transcripts are at a close to 1:1 ratio, as there's generally only 1 UMI for every gene.

3. PRIMER_MATCH - Same as SINGLE_UMI_ERROR, but in addition the UMI perfectly matches one of the PCR primers. These cell barcodes are dropped. These errors are only detected if a PRIMER_SEQUENCE argument is supplied.

4. 4) OTHER - UMIs are extremely skewed towards at least one base (and not T at the last base), but not at all 8 positions. These cell barcodes are dropped.

The file my.synthesis_stats.txt contains a bunch of useful information:

1. CELL_BARCODE - the 12 base cell barcode
2. NUM_UMI - the number of total umis observed
3. FIRST_BIASED_BASE - the first base position where any bias is observed. -1 for no detected bias
4. SYNTH_MISSING_BASE - as #3 but specific to runs of T's at the end of the UMI
5. ERROR_TYPE - see error type definitions above
6. For bases 1-8 of the UMI, the observed base counts across all UMIs. This is a "|" delimited field, with counts of the A,C,G,T,N bases.

The file my.synthesis_stats.summary.txt contains a histogram of the SYNTHESIS_MISSING_BASE errors, as well as the counts of all other errors, the number of total barcodes evaluated, and the number of barcodes ignored.

**End of Alignment**

At this point, the alignment is completed, and your raw reads have been changed from paired reads to single end reads with the cell and molecular barcodes extracted, cleaned up, aligned, and prepared for DGE extraction.

**Going with the flow - using Unix pipes to simplify alignment**

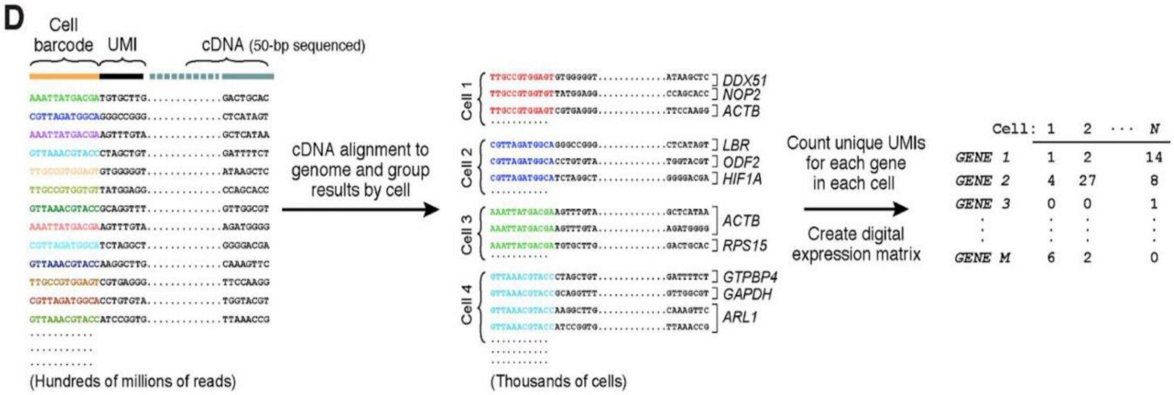
If you're on a Unix or OS X operating system, you may be familiar with pipes. Drop-seq programs extend the Picard API, and so like Picard are able to use pipes to redirect output from one program to the next. Why is this useful? It's a little bit faster, but more importantly it saves a significant amount of disk space by not generating a large number of temporary files, as the examples above have. It also simplifies writing pipelines, as there are fewer named files - intermediate data flows through the pipeline without being saved. The tradeoff is that executing several programs in a pipeline requires more RAM and more processing power, so if your computer does not have a lot of RAM and lots of processors, this might not be useful.

There are some limitations to the amount of pipelining that can be done, because some files must be read more than once, and because STAR does not have the ability to write to standard output. The following steps may be pipelined:

- Pre-alignment tagging and trimming. The final result of these steps must be saved to be input to both SamToFastq and MergeBamAlignment.
    - TagBamWithReadSequenceExtended (one or more times)
    - FilterBam
    - TrimStartingSequence
    - PolyATrimmer
- Alignment (STAR does not support output to a pipe):
    - SamToFastq
    - STAR
- Merging and tagging aligned reads:
    - MergeBamAlignment
    - TagReadWithGeneFunction

The bead-repair programs make multiple passes over the input so they can't be pipelined.

**Overview of DGE extraction**


D
Cell barcode
UMI
cDNA (50-bp sequenced)
(Hundreds of millions of reads)
cDNA alignment to genome and group results by cell
Cell 1
Cell 2
Cell 3
Cell 4
DDX51
NOP2
ACTB
LBR
ODF2
HIF1A
ACTB
RPS15
GTPBP4
GAPDH
ARL1
(Thousands of cells)
Count unique UMIs for each gene in each cell
Create digital expression matrix

| Cell: | 1 | 2 | ··· | N |
|---|---|---|---|---|
| GENE 1 | 1 | 2 | | 14 |
| GENE 2 | 4 | 27 | | 8 |
| GENE 3 | 0 | 0 | | 1 |
| ⋮ | ⋮ | ⋮ | | ⋮ |
| GENE M | 6 | 2 | | 0 |

To digitally count gene transcripts, a list of UMIs in each gene, within each cell, is assembled, and UMIs within edit distance = 1 are merged together. The total number of unique UMI sequences is counted, and this number is reported as the number of transcripts of that gene for a given cell.

**Digital Gene Expression**

Extracting Digital Gene Expression (DGE) data from an aligned library is done using the Drop-seq program DigitalExpression. The input to this program is the aligned BAM from the alignment workflow. There are two outputs available: the primary is the DGE matrix, with each a row for each gene, and a column for each cell. The secondary analysis is a summary of the DGE matrix on a per-cell level, indicating the number of genes and transcripts observed.

**Primary Output Example:**

| GENE | ATCAGGGACAGA | AGGGAAAATTGA | TTGCCTTACGCG | TGGCGAAGAGAT | TACAATTAAGGC |
|---|---|---|---|---|---|
| LOXL4 | 0 | 0 | 0 | 0 | 0 |
| PYROXD2 | 1 | 0 | 1 | 1 | 0 |
| HPS1 | 23 | 12 | 9 | 8 | 3 |
| CNNM1 | 0 | 2 | 1 | 0 | 0 |
| GOT1 | 22 | 6 | 7 | 9 | 3 |

**Summary Output Example:**

| CELL_BARCODE | NUM_GENES | NUM_TRANSCRIPTS |
|---|---|---|
| ATCAGGGACAGA | 12128 | 232831 |
| AGGGAAAATTGA | 12161 | 185418 |
| TTGCCTTACGCG | 10761 | 173547 |
| TGGCGAAGAGAT | 10036 | 108545 |
| TACAATTAAGGC | 9889 | 99771 |
| CTAAGTAGCTTT | 9244 | 91563 |

**Long output Example (new for V2):**

| CELL | GENE | UMI_COUNT |
|---|---|---|
| ATCAGGGACAGA | HPS1 | 23 |
| ATCAGGGACAGA | GOT1 | 22 |
| ATCAGGGACAGA | PYROXD2 | 1 |
| AGGGAAAATTGA | HPS1 | 12 |
| AGGGAAAATTGA | GOT1 | 6 |

This file is ordered by the list of cell barcodes (input cell barcode file or based on number of reads per cell), then the number of UMIs per gene, then alphabetically by gene when they have the same number of UMIs. There are no entries when a cell does not have expression of a gene.

**DGE Extraction Options:**

There are a large number of options in the DGE program, as we've performed large amounts of experimentation with the outputs to this program. Most of these parameters have default settings, and are the correct setting for a standard Drop-seq experiment. Outlined below are some of the parameters that you might change.

**READ_MQ** The minimum map quality of a read to be used in the DGE calculation. For aligners like STAR, the default (10) is higher than what's needed to eliminate all multi-mapping reads. If you use a different aligner, you might want to set a different threshold.

**EDIT_DISTANCE.** By default we collapse UMI barcodes with a hamming distance of 1.

**RARE_UMI_FILTER_THRESHOLD** This is an implementation of the rare UMI filter implemented by Islam, et al. We leave this off by default, and use edit distance collapse instead. If desired, one can set EDIT_DISTANCE=0 and enable this filter instead at some threshold, like 0.01.

**Options for selecting sets of cells**
When running DGE, we don't select every cell barcode observed. This is because the aligned BAM can contain hundreds of thousands of cell barcodes; most reads will be on either STAMPs (beads exposed to a cell in droplets) or "empties" (beads that were exposed only to ambient RNA in droplets). There will also be a lot of cell barcodes with just a handful of reads. Because a huge matrix might be difficult to work with, these options limit the number of cell barcodes that are emitted by DGE extraction. *You must use one of these options.*

**MIN_NUM_GENES_PER_CELL**. DigitalExpression runs a single iteration across all data, and selects cells that have at least this many genes.

**MIN_NUM_TRANSCRIPTS_PER_CELL**. DigitalExpression runs a single iteration across all data, and selects cells that have at least this many transcripts. (Finally bugfixed and working in V2.0.0!)

**NUM_CORE_BARCODES.** DigitalExpression counts the number of reads per cell barcode (thresholded by READ_MQ), and only includes cells that have at least this number of reads.

**CELL_BC_FILE.** Instead of iterating over the BAM and discovering what cell barcodes should be used, override this with a specific subset of cell barcodes in a text file. This file has no header and a single column, containing one cell barcode per line. Since this option doesn't have to iterate through the BAM to select barcodes, DGE extraction is significantly faster when using this option.

**Functional annotations and strand selection, NEW for V2.0.0:**
Along with the changes to how reads are tagged with functional annotations, DGE and similar programs are now able to extract these enhanced sets of tags. There are two main parameters to use:

**STRAND_STRATEGY**:
The strand strategy decides which reads will be used by analysis based on the strand of the read and the strand of the gene. The SENSE strategy requires the read and annotation to be on the same strand. The

ANTISENSE strategy requires the read and annotation to be on opposite strands. The BOTH strategy is permissive, and allows the read to be on either strand.

**LOCUS_FUNCTION_LIST**:
This is a list of functional annotations that should be used to include reads in analysis. The default is include reads in DGE analysis where the read entirely overlaps the CODING and UTR portions of a gene. This is slightly more conservative than DGE V1, which allowed reads that only partially overlap an exon to be counted. Changing the list of annotations allows for different sorts of expression data to be extracted.

**Example:**
In this example, we extract the DGE for the top 100 most commonly occurring cell barcodes in the aligned BAM, using CODING+UTR regions on the SENSE strand.

```
DigitalExpression
I=out_gene_exon_tagged.bam
O=out_gene_exon_tagged.dge.txt.gz
SUMMARY=out_gene_exon_tagged.dge.summary.txt
NUM_CORE_BARCODES=100
```

**Example INTRONIC+CODING:**
If you want to simply add additional annotations to CODING+UTR, specifying LOCUS_FUNCTION_LIST adds to the list. For example, we add intronic expression, and coding is already specified as the default.

```
DigitalExpression
I=out_gene_exon_tagged.bam
O=out_gene_exon_tagged.dge.txt.gz
SUMMARY=out_gene_exon_tagged.dge.summary.txt
NUM_CORE_BARCODES=100
LOCUS_FUNCTION_LIST=INTRONIC.
```

**Example INTRONIC ONLY:**
There's a bit of a "gotcha" in how this is specified by to the program (this comes from the Picard's API for command line argument interpretation.) If you want to specify INTRONIC only expression, you first need to clear the list of functional annotations by giving a value of null, then add your additional values:

```
DigitalExpression
I=out_gene_exon_tagged.bam
O=out_gene_exon_tagged.dge.txt.gz
SUMMARY=out_gene_exon_tagged.dge.summary.txt
NUM_CORE_BARCODES=100
LOCUS_FUNCTION_LIST=null LOCUS_FUNCTION_LIST=INTRONIC.
```

**Cell Selection**

A key question to answer for your data set is how many cells you want to extract from your BAM. One way to estimate this is to extract the number of reads per cell, then plot the cumulative distribution of reads and select the "knee" of the distribution.

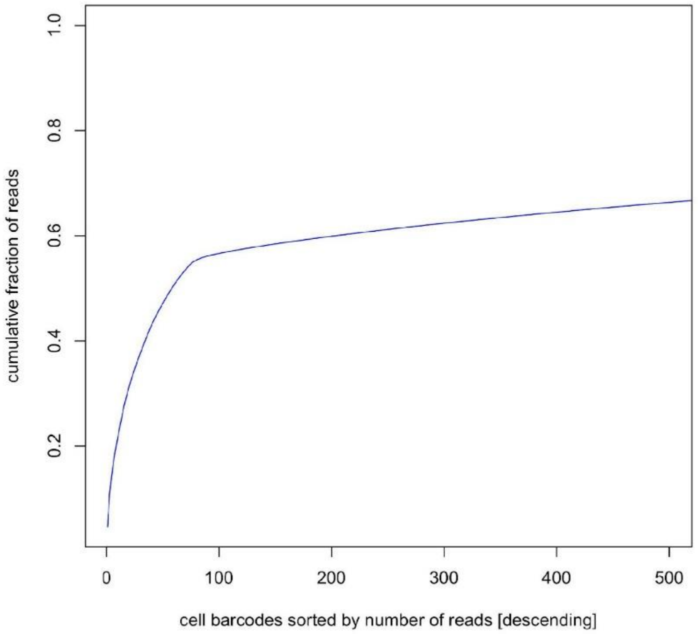
We provide a tool to extract the reads per cell barcode in the Drop-seq software called BAMTagHistogram. This extracts the number of reads for any BAM tag in a BAM file, and is a general purpose tool you can use for a number of purposes. For this purpose, we extract the cell tag "XC":

**Example:**

```
BAMTagHistogram
I=out_gene_exon_tagged.bam
O=out_cell_readcounts.txt.gz
TAG=XC
```

Once we run this program, a little bit of R code can create a cumulative distribution plot. Here's an example using the 100 cells data from the Drop-seq initial publication (Figures 3C and 3D):

```
a=read.table("100cells_numReads_perCell_XC_mq_10.txt.gz", header=F, stringsAsFactors=F)
x=cumsum(a$V1)
x=x/max(x)
plot(1:length(x), x, type='l', col="blue", xlab="cell barcodes sorted by number of reads [descending]",
ylab="cumulative fraction of reads", xlim=c(1,500))
```



In this example, the number of STAMPs are the number of cell barcodes to the left of the inflection point; to the right of the inflection point are the empty beads that have only been exposed to ambient RNA. Figure S3A of Macosko et al., 2015 provides additional justification and explanation for how we identify the number of cells sequenced.

**Mixed-species plots**

To create the mixed species plots used in the paper, we suggest the following steps:

1. Align your data to a mixed species reference. There is metadata at the bottom of one of the pages of our GEO submission
2. Determine how many cells are in your BAM. See **Cell Selection** in this document and the BAMTagHistogram program. Put that list of cell barcodes in a file that has a single column of cell barcodes, 1 per line.
3. At this point, if you are using our human/mouse metadata, your BAM has chromosomes that are prepended with HUMAN or MOUSE, i.e.: HUMAN_11. Filter your BAM into 2 organism specific BAMs using FilterBam with the argument REF_SOFT_MATCHED_RETAINED=HUMAN or REF_SOFT_MATCHED_RETAINED=MOUSE (this is about as fancy as running grep on your BAM.)

4. Run DigitalExpression on each organism specific BAM with the CELL_BC_FILE argument, using the file generated in step #2.
5. You now have two summary files that have the number of genes/transcripts contained by each cell, in an organism specific manner. Merge them into one file and plot.

**Conclusion**

With successful execution of our software you have hopefully transformed a pile of hundreds of millions of sequence reads into a digital expression matrix that has genome-wide expression measurements (digital counts) for each gene in each individual cell.

What to do next? We expect analysis of massive single-cell expression data to become a lively field. We think very highly of the Seurat package developed by our colleague Rahul Satija. We used Seurat to perform all of the downstream analyses (cell clustering, etc) in the Cell paper. Seurat is available on Rahul's web site (http://www.satijalab.org/seurat.html), where Rahul will also have protocols for the specific analyses in the paper

**But what if everything doesn't go perfectly?**

One of the big challenges with releasing a new software toolkit to the world is that people will always do things you didn't anticipate, with data sets you never imagined. While we feel the Drop-seq software produces the computationally correct (at least to our intentions) answers, it's possible that you will discover a bug, or documentation of a particular software parameter will be unclear.

If you find part of this document unclear, let us know and we'll do our best to update it and add clarity. If parameters of our software have unclear documentation, let us know which ones are unclear, and we'll do our best to buff up those descriptions.

If you run into software behavior you think is a bug, then you can help to be part of the solution. To do this, you'll need to give us the following information

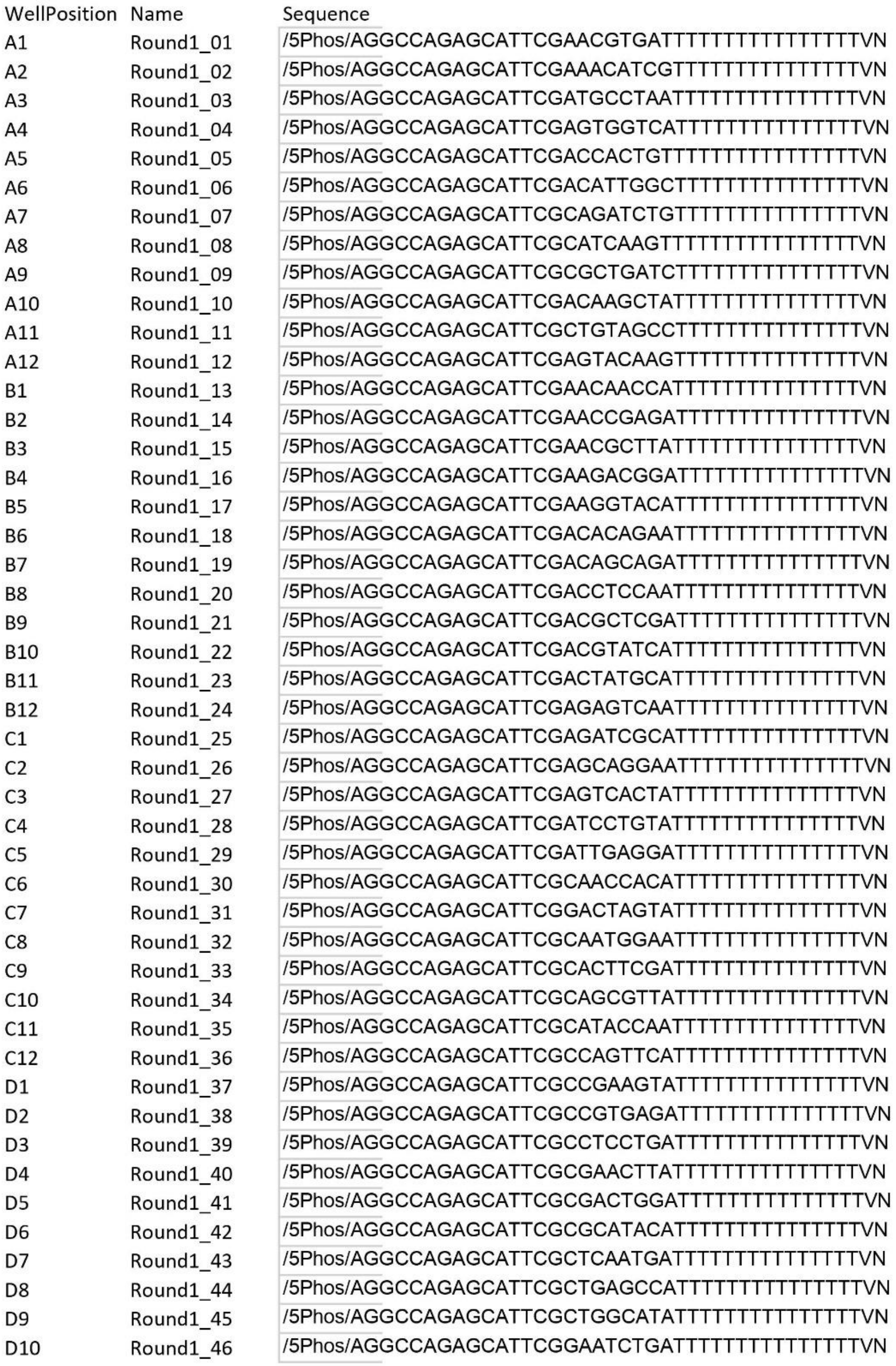
- The program you were running, and the exact command line arguments you supplied to that program
- The console output of the program invocation
- A small test data set that can replicate the problem you observed
- The behavior that you think was faulty, and if possible what you expected to see. This can be very useful when a computation produces an answer that doesn't make sense.
- We have a public githib repository at https://github.com/broadinstitute/Drop-seq. If you're a programmer, you can submit pull requests to us to fix bugs or add additional capabilities.

## 7.6 Oligonucleotides

### 7.6.1 SPLiT-seq barcodes

The SPLiT-seq barcodes were downloaded from https://sites.google.com/uw.edu/splitseq/protocol on the 28th of April 2020. Available on https://www.seeliglab.org/tools.html as of the 5th of July 2022.

| **Number** | **Sequence** |
|---|---|
| BC_0215 | CGAATGCTCTGGCCTCTCAAGCACGTGGAT |
| BC_0216 | ATCCACGTGCTTGAGAGGCCAGAGCATTCG |
| BC_0060 | AGTCGTACGCCGATGCGAAACATCGGCCAC |
| BC_0066 | GTGGCCGATGTTTCGCATCGGCGTACGACT |
| BC_0062 | CAGACGTGTGCTCTTCCGATCT |
| BC_0108 | AAGCAGTGGTATCAACGCAGAGT |
| BC_0076 | CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| BC_0077 | CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| BC_0078 | CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| BC_0079 | CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |



| WellPosition | Name | Sequence |
|---|---|---|
| A1 | Round1_01 | /5Phos/AGGCCAGAGCATTCGAACGTGATTTTTTTTTTTTTTTTVN |
| A2 | Round1_02 | /5Phos/AGGCCAGAGCATTCGAAACATCGTTTTTTTTTTTTTTTVN |
| A3 | Round1_03 | /5Phos/AGGCCAGAGCATTCGATGCCTAATTTTTTTTTTTTTTTVN |
| A4 | Round1_04 | /5Phos/AGGCCAGAGCATTCGAGTGGTCATTTTTTTTTTTTTTTVN |
| A5 | Round1_05 | /5Phos/AGGCCAGAGCATTCGACCACTGTTTTTTTTTTTTTTTTVN |
| A6 | Round1_06 | /5Phos/AGGCCAGAGCATTCGACATTGGCTTTTTTTTTTTTTTTVN |
| A7 | Round1_07 | /5Phos/AGGCCAGAGCATTCGCAGATCTGTTTTTTTTTTTTTTTVN |
| A8 | Round1_08 | /5Phos/AGGCCAGAGCATTCGCATCAAGTTTTTTTTTTTTTTTTVN |
| A9 | Round1_09 | /5Phos/AGGCCAGAGCATTCGCGCTGATCTTTTTTTTTTTTTTTVN |
| A10 | Round1_10 | /5Phos/AGGCCAGAGCATTCGACAAGCTATTTTTTTTTTTTTTTVN |
| A11 | Round1_11 | /5Phos/AGGCCAGAGCATTCGCTGTAGCCTTTTTTTTTTTTTTTVN |
| A12 | Round1_12 | /5Phos/AGGCCAGAGCATTCGAGTACAAGTTTTTTTTTTTTTTTVN |
| B1 | Round1_13 | /5Phos/AGGCCAGAGCATTCGAACAACCATTTTTTTTTTTTTTTVN |
| B2 | Round1_14 | /5Phos/AGGCCAGAGCATTCGAACCGAGATTTTTTTTTTTTTTTVN |
| B3 | Round1_15 | /5Phos/AGGCCAGAGCATTCGAACGCTTATTTTTTTTTTTTTTTVN |
| B4 | Round1_16 | /5Phos/AGGCCAGAGCATTCGAAGACGGATTTTTTTTTTTTTTTVN |
| B5 | Round1_17 | /5Phos/AGGCCAGAGCATTCGAAGGTACATTTTTTTTTTTTTTTVN |
| B6 | Round1_18 | /5Phos/AGGCCAGAGCATTCGACACAGAATTTTTTTTTTTTTTTVN |
| B7 | Round1_19 | /5Phos/AGGCCAGAGCATTCGACAGCAGATTTTTTTTTTTTTTTVN |
| B8 | Round1_20 | /5Phos/AGGCCAGAGCATTCGACCTCCAATTTTTTTTTTTTTTTVN |
| B9 | Round1_21 | /5Phos/AGGCCAGAGCATTCGACGCTCGATTTTTTTTTTTTTTTVN |
| B10 | Round1_22 | /5Phos/AGGCCAGAGCATTCGACGTATCATTTTTTTTTTTTTTTVN |
| B11 | Round1_23 | /5Phos/AGGCCAGAGCATTCGACTATGCATTTTTTTTTTTTTTTVN |
| B12 | Round1_24 | /5Phos/AGGCCAGAGCATTCGAGAGTCAATTTTTTTTTTTTTTTVN |
| C1 | Round1_25 | /5Phos/AGGCCAGAGCATTCGAGATCGCATTTTTTTTTTTTTTTVN |
| C2 | Round1_26 | /5Phos/AGGCCAGAGCATTCGAGCAGGAATTTTTTTTTTTTTTTVN |
| C3 | Round1_27 | /5Phos/AGGCCAGAGCATTCGAGTCACTATTTTTTTTTTTTTTTVN |
| C4 | Round1_28 | /5Phos/AGGCCAGAGCATTCGATCCTGTATTTTTTTTTTTTTTTVN |
| C5 | Round1_29 | /5Phos/AGGCCAGAGCATTCGATTGAGGATTTTTTTTTTTTTTTVN |
| C6 | Round1_30 | /5Phos/AGGCCAGAGCATTCGCAACCACATTTTTTTTTTTTTTTVN |
| C7 | Round1_31 | /5Phos/AGGCCAGAGCATTCGGACTAGTATTTTTTTTTTTTTTTVN |
| C8 | Round1_32 | /5Phos/AGGCCAGAGCATTCGCAATGGAATTTTTTTTTTTTTTTVN |
| C9 | Round1_33 | /5Phos/AGGCCAGAGCATTCGCACTTCGATTTTTTTTTTTTTTTVN |
| C10 | Round1_34 | /5Phos/AGGCCAGAGCATTCGCAGCGTTATTTTTTTTTTTTTTTVN |
| C11 | Round1_35 | /5Phos/AGGCCAGAGCATTCGCATACCAATTTTTTTTTTTTTTTVN |
| C12 | Round1_36 | /5Phos/AGGCCAGAGCATTCGCCAGTTCATTTTTTTTTTTTTTTVN |
| D1 | Round1_37 | /5Phos/AGGCCAGAGCATTCGCCGAAGTATTTTTTTTTTTTTTTVN |
| D2 | Round1_38 | /5Phos/AGGCCAGAGCATTCGCCGTGAGATTTTTTTTTTTTTTTVN |
| D3 | Round1_39 | /5Phos/AGGCCAGAGCATTCGCCTCCTGATTTTTTTTTTTTTTTVN |
| D4 | Round1_40 | /5Phos/AGGCCAGAGCATTCGCGAACTTATTTTTTTTTTTTTTTVN |
| D5 | Round1_41 | /5Phos/AGGCCAGAGCATTCGCGACTGGATTTTTTTTTTTTTTTVN |
| D6 | Round1_42 | /5Phos/AGGCCAGAGCATTCGCGCATACATTTTTTTTTTTTTTTVN |
| D7 | Round1_43 | /5Phos/AGGCCAGAGCATTCGCTCAATGATTTTTTTTTTTTTTTVN |
| D8 | Round1_44 | /5Phos/AGGCCAGAGCATTCGCTGAGCCATTTTTTTTTTTTTTTVN |
| D9 | Round1_45 | /5Phos/AGGCCAGAGCATTCGCTGGCATATTTTTTTTTTTTTTTVN |
| D10 | Round1_46 | /5Phos/AGGCCAGAGCATTCGGAATCTGATTTTTTTTTTTTTTTVN |

| | | |
|---|---|---|
| D11 | Round1_47 | /5Phos/AGGCCAGAGCATTCGCAAGACTATTTTTTTTTTTTTTTVN |
| D12 | Round1_48 | /5Phos/AGGCCAGAGCATTCGGAGCTGAATTTTTTTTTTTTTTTVN |
| E1 | Round1_49 | /5Phos/AGGCCAGAGCATTCGGATAGACANNNNNN |
| E2 | Round1_50 | /5Phos/AGGCCAGAGCATTCGGCCACATANNNNNN |
| E3 | Round1_51 | /5Phos/AGGCCAGAGCATTCGGCGAGTAANNNNNN |
| E4 | Round1_52 | /5Phos/AGGCCAGAGCATTCGGCTAACGANNNNNN |
| E5 | Round1_53 | /5Phos/AGGCCAGAGCATTCGGCTCGGTANNNNNN |
| E6 | Round1_54 | /5Phos/AGGCCAGAGCATTCGGGAGAACANNNNNN |
| E7 | Round1_55 | /5Phos/AGGCCAGAGCATTCGGGTGCGAANNNNNN |
| E8 | Round1_56 | /5Phos/AGGCCAGAGCATTCGGTACGCAANNNNNN |
| E9 | Round1_57 | /5Phos/AGGCCAGAGCATTCGGTCGTAGANNNNNN |
| E10 | Round1_58 | /5Phos/AGGCCAGAGCATTCGGTCTGTCANNNNNN |
| E11 | Round1_59 | /5Phos/AGGCCAGAGCATTCGGTGTTCTANNNNNN |
| E12 | Round1_60 | /5Phos/AGGCCAGAGCATTCGTAGGATGANNNNNN |
| F1 | Round1_61 | /5Phos/AGGCCAGAGCATTCGTATCAGCANNNNNN |
| F2 | Round1_62 | /5Phos/AGGCCAGAGCATTCGTCCGTCTANNNNNN |
| F3 | Round1_63 | /5Phos/AGGCCAGAGCATTCGTCTTCACANNNNNN |
| F4 | Round1_64 | /5Phos/AGGCCAGAGCATTCGTGAAGAGANNNNNN |
| F5 | Round1_65 | /5Phos/AGGCCAGAGCATTCGTGGAACAANNNNNN |
| F6 | Round1_66 | /5Phos/AGGCCAGAGCATTCGTGGCTTCANNNNNN |
| F7 | Round1_67 | /5Phos/AGGCCAGAGCATTCGTGGTGGTANNNNNN |
| F8 | Round1_68 | /5Phos/AGGCCAGAGCATTCGTTCACGCANNNNNN |
| F9 | Round1_69 | /5Phos/AGGCCAGAGCATTCGAACTCACCNNNNNN |
| F10 | Round1_70 | /5Phos/AGGCCAGAGCATTCGAAGAGATCNNNNNN |
| F11 | Round1_71 | /5Phos/AGGCCAGAGCATTCGAAGGACACNNNNNN |
| F12 | Round1_72 | /5Phos/AGGCCAGAGCATTCGAATCCGTCNNNNNN |
| G1 | Round1_73 | /5Phos/AGGCCAGAGCATTCGAATGTTGCNNNNNN |
| G2 | Round1_74 | /5Phos/AGGCCAGAGCATTCGACACGACCNNNNNN |
| G3 | Round1_75 | /5Phos/AGGCCAGAGCATTCGACAGATTCNNNNNN |
| G4 | Round1_76 | /5Phos/AGGCCAGAGCATTCGAGATGTACNNNNNN |
| G5 | Round1_77 | /5Phos/AGGCCAGAGCATTCGAGCACCTCNNNNNN |
| G6 | Round1_78 | /5Phos/AGGCCAGAGCATTCGAGCCATGCNNNNNN |
| G7 | Round1_79 | /5Phos/AGGCCAGAGCATTCGAGGCTAACNNNNNN |
| G8 | Round1_80 | /5Phos/AGGCCAGAGCATTCGATAGCGACNNNNNN |
| G9 | Round1_81 | /5Phos/AGGCCAGAGCATTCGATCATTCCNNNNNN |
| G10 | Round1_82 | /5Phos/AGGCCAGAGCATTCGATTGGCTCNNNNNN |
| G11 | Round1_83 | /5Phos/AGGCCAGAGCATTCGCAAGGAGCNNNNNN |
| G12 | Round1_84 | /5Phos/AGGCCAGAGCATTCGCACCTTACNNNNNN |
| H1 | Round1_85 | /5Phos/AGGCCAGAGCATTCGCCATCCTCNNNNNN |
| H2 | Round1_86 | /5Phos/AGGCCAGAGCATTCGCCGACAACNNNNNN |
| H3 | Round1_87 | /5Phos/AGGCCAGAGCATTCGCCTAATCCNNNNNN |
| H4 | Round1_88 | /5Phos/AGGCCAGAGCATTCGCCTCTATCNNNNNN |
| H5 | Round1_89 | /5Phos/AGGCCAGAGCATTCGCGACACACNNNNNN |
| H6 | Round1_90 | /5Phos/AGGCCAGAGCATTCGCGGATTGCNNNNNN |
| H7 | Round1_91 | /5Phos/AGGCCAGAGCATTCGCTAAGGTCNNNNNN |
| H8 | Round1_92 | /5Phos/AGGCCAGAGCATTCGGAACAGGCNNNNNN |
| H9 | Round1_93 | /5Phos/AGGCCAGAGCATTCGGACAGTGCNNNNNN |

| | | |
|---|---|---|
| H10 | Round1_94 | /5Phos/AGGCCAGAGCATTCGGAGTTAGCNNNNNN |
| H11 | Round1_95 | /5Phos/AGGCCAGAGCATTCGGATGAATCNNNNNN |
| H12 | Round1_96 | /5Phos/AGGCCAGAGCATTCGGCCAAGACNNNNNN |

| WellPosition | Name | Sequence |
|---|---|---|
| A1 | Round2_01 | /5Phos/CATCGGCGTACGACTAACGTGATATCCACGTGCTTGAG |
| A2 | Round2_02 | /5Phos/CATCGGCGTACGACTAAACATCGATCCACGTGCTTGAG |
| A3 | Round2_03 | /5Phos/CATCGGCGTACGACTATGCCTAAATCCACGTGCTTGAG |
| A4 | Round2_04 | /5Phos/CATCGGCGTACGACTAGTGGTCAATCCACGTGCTTGAG |
| A5 | Round2_05 | /5Phos/CATCGGCGTACGACTACCACTGTATCCACGTGCTTGAG |
| A6 | Round2_06 | /5Phos/CATCGGCGTACGACTACATTGGCATCCACGTGCTTGAG |
| A7 | Round2_07 | /5Phos/CATCGGCGTACGACTCAGATCTGATCCACGTGCTTGAG |
| A8 | Round2_08 | /5Phos/CATCGGCGTACGACTCATCAAGTATCCACGTGCTTGAG |
| A9 | Round2_09 | /5Phos/CATCGGCGTACGACTCGCTGATCATCCACGTGCTTGAG |
| A10 | Round2_10 | /5Phos/CATCGGCGTACGACTACAAGCTAATCCACGTGCTTGAG |
| A11 | Round2_11 | /5Phos/CATCGGCGTACGACTCTGTAGCCATCCACGTGCTTGAG |
| A12 | Round2_12 | /5Phos/CATCGGCGTACGACTAGTACAAGATCCACGTGCTTGAG |
| B1 | Round2_13 | /5Phos/CATCGGCGTACGACTAACAACCAATCCACGTGCTTGAG |
| B2 | Round2_14 | /5Phos/CATCGGCGTACGACTAACCGAGAATCCACGTGCTTGAG |
| B3 | Round2_15 | /5Phos/CATCGGCGTACGACTAACGCTTAATCCACGTGCTTGAG |
| B4 | Round2_16 | /5Phos/CATCGGCGTACGACTAAGACGGAATCCACGTGCTTGAG |
| B5 | Round2_17 | /5Phos/CATCGGCGTACGACTAAGGTACAATCCACGTGCTTGAG |
| B6 | Round2_18 | /5Phos/CATCGGCGTACGACTACACAGAAATCCACGTGCTTGAG |
| B7 | Round2_19 | /5Phos/CATCGGCGTACGACTACAGCAGAATCCACGTGCTTGAG |
| B8 | Round2_20 | /5Phos/CATCGGCGTACGACTACCTCCAAATCCACGTGCTTGAG |
| B9 | Round2_21 | /5Phos/CATCGGCGTACGACTACGCTCGAATCCACGTGCTTGAG |
| B10 | Round2_22 | /5Phos/CATCGGCGTACGACTACGTATCAATCCACGTGCTTGAG |
| B11 | Round2_23 | /5Phos/CATCGGCGTACGACTACTATGCAATCCACGTGCTTGAG |
| B12 | Round2_24 | /5Phos/CATCGGCGTACGACTAGAGTCAAATCCACGTGCTTGAG |
| C1 | Round2_25 | /5Phos/CATCGGCGTACGACTAGATCGCAATCCACGTGCTTGAG |
| C2 | Round2_26 | /5Phos/CATCGGCGTACGACTAGCAGGAAATCCACGTGCTTGAG |
| C3 | Round2_27 | /5Phos/CATCGGCGTACGACTAGTCACTAATCCACGTGCTTGAG |
| C4 | Round2_28 | /5Phos/CATCGGCGTACGACTATCCTGTAATCCACGTGCTTGAG |
| C5 | Round2_29 | /5Phos/CATCGGCGTACGACTATTGAGGAATCCACGTGCTTGAG |
| C6 | Round2_30 | /5Phos/CATCGGCGTACGACTCAACCACAATCCACGTGCTTGAG |
| C7 | Round2_31 | /5Phos/CATCGGCGTACGACTGACTAGTAATCCACGTGCTTGAG |
| C8 | Round2_32 | /5Phos/CATCGGCGTACGACTCAATGGAAATCCACGTGCTTGAG |
| C9 | Round2_33 | /5Phos/CATCGGCGTACGACTCACTTCGAATCCACGTGCTTGAG |
| C10 | Round2_34 | /5Phos/CATCGGCGTACGACTCAGCGTTAATCCACGTGCTTGAG |
| C11 | Round2_35 | /5Phos/CATCGGCGTACGACTCATACCAAATCCACGTGCTTGAG |
| C12 | Round2_36 | /5Phos/CATCGGCGTACGACTCCAGTTCAATCCACGTGCTTGAG |
| D1 | Round2_37 | /5Phos/CATCGGCGTACGACTCCGAAGTAATCCACGTGCTTGAG |
| D2 | Round2_38 | /5Phos/CATCGGCGTACGACTCCGTGAGAATCCACGTGCTTGAG |
| D3 | Round2_39 | /5Phos/CATCGGCGTACGACTCCTCCTGAATCCACGTGCTTGAG |
| D4 | Round2_40 | /5Phos/CATCGGCGTACGACTCGAACTTAATCCACGTGCTTGAG |
| D5 | Round2_41 | /5Phos/CATCGGCGTACGACTCGACTGGAATCCACGTGCTTGAG |
| D6 | Round2_42 | /5Phos/CATCGGCGTACGACTCGCATACAATCCACGTGCTTGAG |
| D7 | Round2_43 | /5Phos/CATCGGCGTACGACTCTCAATGAATCCACGTGCTTGAG |
| D8 | Round2_44 | /5Phos/CATCGGCGTACGACTCTGAGCCAATCCACGTGCTTGAG |
| D9 | Round2_45 | /5Phos/CATCGGCGTACGACTCTGGCATAATCCACGTGCTTGAG |
| D10 | Round2_46 | /5Phos/CATCGGCGTACGACTGAATCTGAATCCACGTGCTTGAG |

| | | |
|---|---|---|
| D11 | Round2_47 | /5Phos/CATCGGCGTACGACTCAAGACTAATCCACGTGCTTGAG |
| D12 | Round2_48 | /5Phos/CATCGGCGTACGACTGAGCTGAAATCCACGTGCTTGAG |
| E1 | Round2_49 | /5Phos/CATCGGCGTACGACTGATAGACAATCCACGTGCTTGAG |
| E2 | Round2_50 | /5Phos/CATCGGCGTACGACTGCCACATAATCCACGTGCTTGAG |
| E3 | Round2_51 | /5Phos/CATCGGCGTACGACTGCGAGTAAATCCACGTGCTTGAG |
| E4 | Round2_52 | /5Phos/CATCGGCGTACGACTGCTAACGAATCCACGTGCTTGAG |
| E5 | Round2_53 | /5Phos/CATCGGCGTACGACTGCTCGGTAATCCACGTGCTTGAG |
| E6 | Round2_54 | /5Phos/CATCGGCGTACGACTGGAGAACAATCCACGTGCTTGAG |
| E7 | Round2_55 | /5Phos/CATCGGCGTACGACTGGTGCGAAATCCACGTGCTTGAG |
| E8 | Round2_56 | /5Phos/CATCGGCGTACGACTGTACGCAAATCCACGTGCTTGAG |
| E9 | Round2_57 | /5Phos/CATCGGCGTACGACTGTCGTAGAATCCACGTGCTTGAG |
| E10 | Round2_58 | /5Phos/CATCGGCGTACGACTGTCTGTCAATCCACGTGCTTGAG |
| E11 | Round2_59 | /5Phos/CATCGGCGTACGACTGTGTTCTAATCCACGTGCTTGAG |
| E12 | Round2_60 | /5Phos/CATCGGCGTACGACTTAGGATGAATCCACGTGCTTGAG |
| F1 | Round2_61 | /5Phos/CATCGGCGTACGACTTATCAGCAATCCACGTGCTTGAG |
| F2 | Round2_62 | /5Phos/CATCGGCGTACGACTTCCGTCTAATCCACGTGCTTGAG |
| F3 | Round2_63 | /5Phos/CATCGGCGTACGACTTCTTCACAATCCACGTGCTTGAG |
| F4 | Round2_64 | /5Phos/CATCGGCGTACGACTTGAAGAGAATCCACGTGCTTGAG |
| F5 | Round2_65 | /5Phos/CATCGGCGTACGACTTGGAACAAATCCACGTGCTTGAG |
| F6 | Round2_66 | /5Phos/CATCGGCGTACGACTTGGCTTCAATCCACGTGCTTGAG |
| F7 | Round2_67 | /5Phos/CATCGGCGTACGACTTGGTGGTAATCCACGTGCTTGAG |
| F8 | Round2_68 | /5Phos/CATCGGCGTACGACTTTCACGCAATCCACGTGCTTGAG |
| F9 | Round2_69 | /5Phos/CATCGGCGTACGACTAACTCACCATCCACGTGCTTGAG |
| F10 | Round2_70 | /5Phos/CATCGGCGTACGACTAAGAGATCATCCACGTGCTTGAG |
| F11 | Round2_71 | /5Phos/CATCGGCGTACGACTAAGGACACATCCACGTGCTTGAG |
| F12 | Round2_72 | /5Phos/CATCGGCGTACGACTAATCCGTCATCCACGTGCTTGAG |
| G1 | Round2_73 | /5Phos/CATCGGCGTACGACTAATGTTGCATCCACGTGCTTGAG |
| G2 | Round2_74 | /5Phos/CATCGGCGTACGACTACACGACCATCCACGTGCTTGAG |
| G3 | Round2_75 | /5Phos/CATCGGCGTACGACTACAGATTCATCCACGTGCTTGAG |
| G4 | Round2_76 | /5Phos/CATCGGCGTACGACTAGATGTACATCCACGTGCTTGAG |
| G5 | Round2_77 | /5Phos/CATCGGCGTACGACTAGCACCTCATCCACGTGCTTGAG |
| G6 | Round2_78 | /5Phos/CATCGGCGTACGACTAGCCATGCATCCACGTGCTTGAG |
| G7 | Round2_79 | /5Phos/CATCGGCGTACGACTAGGCTAACATCCACGTGCTTGAG |
| G8 | Round2_80 | /5Phos/CATCGGCGTACGACTATAGCGACATCCACGTGCTTGAG |
| G9 | Round2_81 | /5Phos/CATCGGCGTACGACTATCATTCCATCCACGTGCTTGAG |
| G10 | Round2_82 | /5Phos/CATCGGCGTACGACTATTGGCTCATCCACGTGCTTGAG |
| G11 | Round2_83 | /5Phos/CATCGGCGTACGACTCAAGGAGCATCCACGTGCTTGAG |
| G12 | Round2_84 | /5Phos/CATCGGCGTACGACTCACCTTACATCCACGTGCTTGAG |
| H1 | Round2_85 | /5Phos/CATCGGCGTACGACTCCATCCTCATCCACGTGCTTGAG |
| H2 | Round2_86 | /5Phos/CATCGGCGTACGACTCCGACAACATCCACGTGCTTGAG |
| H3 | Round2_87 | /5Phos/CATCGGCGTACGACTCCTAATCCATCCACGTGCTTGAG |
| H4 | Round2_88 | /5Phos/CATCGGCGTACGACTCCTCTATCATCCACGTGCTTGAG |
| H5 | Round2_89 | /5Phos/CATCGGCGTACGACTCGACACACATCCACGTGCTTGAG |
| H6 | Round2_90 | /5Phos/CATCGGCGTACGACTCGGATTGCATCCACGTGCTTGAG |
| H7 | Round2_91 | /5Phos/CATCGGCGTACGACTCTAAGGTCATCCACGTGCTTGAG |
| H8 | Round2_92 | /5Phos/CATCGGCGTACGACTGAACAGGCATCCACGTGCTTGAG |
| H9 | Round2_93 | /5Phos/CATCGGCGTACGACTGACAGTGCATCCACGTGCTTGAG |

| | | |
|---|---|---|
| H10 | Round2_94 | /5Phos/CATCGGCGTACGACTGAGTTAGCATCCACGTGCTTGAG |
| H11 | Round2_95 | /5Phos/CATCGGCGTACGACTGATGAATCATCCACGTGCTTGAG |
| H12 | Round2_96 | /5Phos/CATCGGCGTACGACTGCCAAGACATCCACGTGCTTGAG |

| Well | Name | Sequence |
|---|---|---|
| A1 | Round3_01 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACGTGATGTGGCCGATGTTTCG |
| A2 | Round3_02 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAACATCGGTGGCCGATGTTTCG |
| A3 | Round3_03 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATGCCTAAGTGGCCGATGTTTCG |
| A4 | Round3_04 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGTGGTCAGTGGCCGATGTTTCG |
| A5 | Round3_05 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACCACTGTGTGGCCGATGTTTCG |
| A6 | Round3_06 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACATTGGCGTGGCCGATGTTTCG |
| A7 | Round3_07 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAGATCTGGTGGCCGATGTTTCG |
| A8 | Round3_08 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCATCAAGTGTGGCCGATGTTTCG |
| A9 | Round3_09 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGCTGATCGTGGCCGATGTTTCG |
| A10 | Round3_10 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACAAGCTAGTGGCCGATGTTTCG |
| A11 | Round3_11 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGTAGCCGTGGCCGATGTTTCG |
| A12 | Round3_12 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGTACAAGGTGGCCGATGTTTCG |
| B1 | Round3_13 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACAACCAGTGGCCGATGTTTCG |
| B2 | Round3_14 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACCGAGAGTGGCCGATGTTTCG |
| B3 | Round3_15 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACGCTTAGTGGCCGATGTTTCG |
| B4 | Round3_16 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGACGGAGTGGCCGATGTTTCG |
| B5 | Round3_17 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGGTACAGTGGCCGATGTTTCG |
| B6 | Round3_18 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACACAGAAGTGGCCGATGTTTCG |
| B7 | Round3_19 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACAGCAGAGTGGCCGATGTTTCG |
| B8 | Round3_20 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACCTCCAAGTGGCCGATGTTTCG |
| B9 | Round3_21 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACGCTCGAGTGGCCGATGTTTCG |
| B10 | Round3_22 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACGTATCAGTGGCCGATGTTTCG |
| B11 | Round3_23 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACTATGCAGTGGCCGATGTTTCG |
| B12 | Round3_24 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGAGTCAAGTGGCCGATGTTTCG |
| C1 | Round3_25 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGATCGCAGTGGCCGATGTTTCG |
| C2 | Round3_26 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCAGGAAGTGGCCGATGTTTCG |
| C3 | Round3_27 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGTCACTAGTGGCCGATGTTTCG |
| C4 | Round3_28 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATCCTGTAGTGGCCGATGTTTCG |
| C5 | Round3_29 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATTGAGGAGTGGCCGATGTTTCG |
| C6 | Round3_30 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAACCACAGTGGCCGATGTTTCG |
| C7 | Round3_31 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGACTAGTAGTGGCCGATGTTTCG |
| C8 | Round3_32 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAATGGAAGTGGCCGATGTTTCG |
| C9 | Round3_33 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCACTTCGAGTGGCCGATGTTTCG |
| C10 | Round3_34 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAGCGTTAGTGGCCGATGTTTCG |
| C11 | Round3_35 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCATACCAAGTGGCCGATGTTTCG |
| C12 | Round3_36 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCAGTTCAGTGGCCGATGTTTCG |
| D1 | Round3_37 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGAAGTAGTGGCCGATGTTTCG |
| D2 | Round3_38 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGTGAGAGTGGCCGATGTTTCG |
| D3 | Round3_39 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTCCTGAGTGGCCGATGTTTCG |
| D4 | Round3_40 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGAACTTAGTGGCCGATGTTTCG |
| D5 | Round3_41 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGACTGGAGTGGCCGATGTTTCG |
| D6 | Round3_42 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGCATACAGTGGCCGATGTTTCG |
| D7 | Round3_43 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTCAATGAGTGGCCGATGTTTCG |
| D8 | Round3_44 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGAGCCAGTGGCCGATGTTTCG |
| D9 | Round3_45 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGGCATAGTGGCCGATGTTTCG |
| D10 | Round3_46 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAATCTGAGTGGCCGATGTTTCG |

| | | |
|---|---|---|
| D11 | Round3_47 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAAGACTAGTGGCCGATGTTTCG |
| D12 | Round3_48 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAGCTGAAGTGGCCGATGTTTCG |
| E1 | Round3_49 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGATAGACAGTGGCCGATGTTTCG |
| E2 | Round3_50 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCCACATAGTGGCCGATGTTTCG |
| E3 | Round3_51 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCGAGTAAGTGGCCGATGTTTCG |
| E4 | Round3_52 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCTAACGAGTGGCCGATGTTTCG |
| E5 | Round3_53 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCTCGGTAGTGGCCGATGTTTCG |
| E6 | Round3_54 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGGAGAACAGTGGCCGATGTTTCG |
| E7 | Round3_55 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGGTGCGAAGTGGCCGATGTTTCG |
| E8 | Round3_56 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTACGCAAGTGGCCGATGTTTCG |
| E9 | Round3_57 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTCGTAGAGTGGCCGATGTTTCG |
| E10 | Round3_58 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTCTGTCAGTGGCCGATGTTTCG |
| E11 | Round3_59 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTGTTCTAGTGGCCGATGTTTCG |
| E12 | Round3_60 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTAGGATGAGTGGCCGATGTTTCG |
| F1 | Round3_61 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTATCAGCAGTGGCCGATGTTTCG |
| F2 | Round3_62 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTCCGTCTAGTGGCCGATGTTTCG |
| F3 | Round3_63 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTCTTCACAGTGGCCGATGTTTCG |
| F4 | Round3_64 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGAAGAGAGTGGCCGATGTTTCG |
| F5 | Round3_65 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGAACAAGTGGCCGATGTTTCG |
| F6 | Round3_66 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGCTTCAGTGGCCGATGTTTCG |
| F7 | Round3_67 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGTGGTAGTGGCCGATGTTTCG |
| F8 | Round3_68 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTTCACGCAGTGGCCGATGTTTCG |
| F9 | Round3_69 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACTCACCGTGGCCGATGTTTCG |
| F10 | Round3_70 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGAGATCGTGGCCGATGTTTCG |
| F11 | Round3_71 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGGACACGTGGCCGATGTTTCG |
| F12 | Round3_72 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAATCCGTCGTGGCCGATGTTTCG |
| G1 | Round3_73 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAATGTTGCGTGGCCGATGTTTCG |
| G2 | Round3_74 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACACGACCGTGGCCGATGTTTCG |
| G3 | Round3_75 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACAGATTCGTGGCCGATGTTTCG |
| G4 | Round3_76 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGATGTACGTGGCCGATGTTTCG |
| G5 | Round3_77 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCACCTCGTGGCCGATGTTTCG |
| G6 | Round3_78 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCCATGCGTGGCCGATGTTTCG |
| G7 | Round3_79 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGGCTAACGTGGCCGATGTTTCG |
| G8 | Round3_80 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATAGCGACGTGGCCGATGTTTCG |
| G9 | Round3_81 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATCATTCCGTGGCCGATGTTTCG |
| G10 | Round3_82 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATTGGCTCGTGGCCGATGTTTCG |
| G11 | Round3_83 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAAGGAGCGTGGCCGATGTTTCG |
| G12 | Round3_84 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCACCTTACGTGGCCGATGTTTCG |
| H1 | Round3_85 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCATCCTCGTGGCCGATGTTTCG |
| H2 | Round3_86 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGACAACGTGGCCGATGTTTCG |
| H3 | Round3_87 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTAATCCGTGGCCGATGTTTCG |
| H4 | Round3_88 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTCTATCGTGGCCGATGTTTCG |
| H5 | Round3_89 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGACACACGTGGCCGATGTTTCG |
| H6 | Round3_90 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGGATTGCGTGGCCGATGTTTCG |
| H7 | Round3_91 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTAAGGTCGTGGCCGATGTTTCG |
| H8 | Round3_92 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAACAGGCGTGGCCGATGTTTCG |
| H9 | Round3_93 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGACAGTGCGTGGCCGATGTTTCG |

|  |  |  |
|---|---|---|
| H10 | Round3_94 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAGTTAGCGTGGCCGATGTTTCG |
| H11 | Round3_95 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGATGAATCGTGGCCGATGTTTCG |
| H12 | Round3_96 | /5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCCAAGACGTGGCCGATGTTTCG |

### 7.6.2 Real-time PCR primers

| Gene | Company | Product name | GeneGlobe Id | Catalog Number |
|---|---|---|---|---|
| Gfap | Qiagen | Mm_Gfap_1_SG QuantiTect Primer Assay | QT00101143 | 249900 |
| Aldh1 | Qiagen | Mm_Aldh1l1_va.1_SG QuantiTect Primer Assay | QT01565382 | 249900 |
| Rbfox3 | Qiagen | Mm_Rbfox3_1_SG QuantiTect Primer Assay | QT01054326 | 249900 |
| Pdgfra | Qiagen | Mm_Pdgfra_1_SG QuantiTect Primer Assay | QT00140021 | 249900 |
| Ywhaz | Qiagen | Mm_Ywhaz_1_SG QuantiTect Primer Assay | QT00105350 | 249900 |
| Tubb4a | Qiagen | Mm_Tubb4a_1_SG QuantiTect Primer Assay | QT00251664 | 249900 |
| Sdha | Qiagen | Mm_Sdha_1_SG QuantiTect Primer Assay | QT00265237 | 249900 |
| Abi3bp | Qiagen | Mm_Abi3bp_1_SG QuantiTect Primer Assay | QT01074199 | 249900 |
| Auts2 | Qiagen | Mm_Auts2_1_SG QuantiTect Primer Assay | QT00147000 | 249900 |
| Gphn | Qiagen | Mm_Gphn_1_SG QuantiTect Primer Assay | QT00170275 | 249900 |
| Il31ra | Qiagen | Mm_Il31ra_1_SG QuantiTect Primer Assay | QT00144382 | 249900 |
| Ndst4 | Qiagen | Mm_Ndst4_1_SG QuantiTect Primer Assay | QT01066268 | 249900 |
| Pde10a | Qiagen | Mm_Pde10a_1_SG QuantiTect Primer Assay | QT00151151 | 249900 |
| Pdzrn4 | Qiagen | Mm_Pdzrn4_1_SG QuantiTect Primer Assay | QT00299467 | 249900 |
| Rora | Qiagen | Mm_Rora_1_SG QuantiTect Primer Assay | QT00158053 | 249900 |
| Plp1 | Qiagen | Mm_Plp1_1_SG QuantiTect Primer Assay | QT00096096 | 249900 |
| Ptn | Qiagen | Mm_Ptn_1_SG QuantiTect Primer Assay | QT00167076 | 249900 |
| Trf | Qiagen | Mm_Trf_1_SG QuantiTect Primer Assay | QT00198072 | 249900 |
| Pard3 | Qiagen | Mm_Pard3_1_SG QuantiTect Primer Assay | QT00161875 | 249900 |
| Gjc3 | Qiagen | Mm_Gjc3_1_SG QuantiTect Primer Assay | QT00168581 | 249900 |
| Rph3a | Qiagen | Mm_Rph3a_1_SG QuantiTect Primer Assay | QT00135695 | 249900 |
| Nrp1 | Qiagen | Mm_Nrp1_1_SG QuantiTect Primer Assay | QT00157381 | 249900 |
| Igsf9b | Qiagen | Mm_Igsf9b_1_SG QuantiTect Primer Assay | QT01050567 | 249900 |
| Homer1 | Qiagen | Mm_Homer1_1_SG QuantiTect Primer Assay | QT00129983 | 249900 |
| Ptk2 | Qiagen | Mm_Ptk2_1_SG QuantiTect Primer Assay | QT01059891 | 249900 |
| Grin2a | Qiagen | Mm_Grin2a_1_SG QuantiTect Primer Assay | QT00093562 | 249900 |
| Apoe | Qiagen | Mm_Apoe_1_SG QuantiTect Primer Assay | QT01043889 | 249900 |
| Mef2c | Qiagen | Mm_Mef2c_1_SG QuantiTect Primer Assay | QT00103733 | 249900 |
| Lcn2 | Qiagen | Mm_Lcn2_1_SG QuantiTect Primer Assay | QT00113407 | 249900 |
| Hspb1 | Qiagen | Mm_Hspb1_1_SG QuantiTect Primer Assay | QT00100632 | 249900 |
| Vim | Qiagen | Mm_Vim_1_SG QuantiTect Primer Assay | QT00159670 | 249900 |
| Osmr | Qiagen | Mm_Osmr_1_SG QuantiTect Primer Assay | QT00104433 | 249900 |
| Serping1 | Qiagen | Mm_Serping1_1_SG QuantiTect Primer Assay | QT00126252 | 249900 |
| Fkbp5 | Qiagen | Mm_Fkbp5_1_SG QuantiTect Primer Assay | QT00166390 | 249900 |
| Ggta1 | Qiagen | Mm_Ggta1_1_SG QuantiTect Primer Assay | QT00165788 | 249900 |
| Srgn | Qiagen | Mm_Srgn_1_SG QuantiTect Primer Assay | QT01064273 | 249900 |
| C3 | Qiagen | Mm_C3_1_SG QuantiTect Primer Assay | QT00109270 | 249900 |
| Gbp2 | Qiagen | Mm_Gbp2_1_SG QuantiTect Primer Assay | QT00106050 | 249900 |

|  |  |  |  |  |
|---|---|---|---|---|
| Ptx3 | Qiagen | Mm_Ptx3_1_SG QuantiTect Primer Assay | QT01063587 | 249900 |
| S100a10 | Qiagen | Mm_S100a10_1_SG QuantiTect Primer Assay | QT00103894 | 249900 |
| Cd109 | Qiagen | Mm_Cd109_1_SG QuantiTect Primer Assay | QT00127638 | 249900 |
| Emp1 | Qiagen | Mm_Emp1_1_SG QuantiTect Primer Assay | QT00137774 | 249900 |
| Slc10a6 | Qiagen | Mm_Slc10a6_va.1_SG QuantiTect Primer Assay | QT01548750 | 249900 |
| Tm4sf1 | Qiagen | Mm_Tm4sf1_1_SG QuantiTect Primer Assay | QT00097076 | 249900 |

QuantiTect lyophilized primers were reconstituted in 1.1 mL of TE, pH 8.0, to prepare QuantiTect Primer Assays at 10X stock concentration. The exact concentration of the primers is proprietary.

## 7.7 Scripts

### 7.7.1 Count matrices to Seurat objects

```r
# Load packages
library("Seurat")
library("openxlsx")

# Samples info
samples <- read.xlsx("samples.xlsx", detectDates = T)

# Create the directory for the Seurat objects
dir.create('./seurat_objects', showWarnings = F)

# Function to create the Seurat objects and add metadata
create_seurat_object <- function(samples) {
  timepoint <- samples[1]
  animal <- samples[2]
  inocula <- samples[3]
  date <- samples[4]

  project_folder <- paste0("mice_", timepoint)
  sample_folder <- paste0(animal, '_DGE_filtered')
  dge_path <- file.path('..',project_folder, 'splitseq_pipeline', 'libs_merged', sample_folder)

  # Read the 3 files as a sparse matrix
  sr.data <- Seurat::ReadMtx(
    mtx = file.path(dge_path, "DGE.mtx"),
    cells = file.path(dge_path, "cell_metadata.csv"),
    features = file.path(dge_path, "genes.csv"),
    feature.column = 2,
    cell.sep = ",",
    feature.sep = ",",
    mtx.transpose = T,
    skip.cell = 1,
    skip.feature = 1
  )

  # Create the Seurat object
  sr <- CreateSeuratObject(counts = sr.data, project = "mouse_sc", min.cells = 3)

  # Add metadata
  sr[["timepoint"]] <- timepoint
  sr[["animal"]] <- animal
  sr[["inocula"]] <- inocula
  sr[["date"]] <- date

  # Set identities for each cell
  Idents(sr) <- paste(timepoint, inocula, animal, sep = "_")

  # Save the object
  saveRDS(sr, file.path("seurat_objects", paste0(animal, ".rds")))

  # Return the object to be saved in a list
  sr
}

# Run the function to save the objects in the directory
sr_objects <- apply(samples, 1, create_seurat_object)

# Now we need to merge the objects in a new object
```

```
sr_merged <- merge(x = sr_objects[[1]], y = sr_objects[-1])

# Save the merged object
saveRDS(sr_merged, "./seurat_objects/sr_merged.rds")
```

### 7.7.2 Reference dataset pre-processing in Seurat

```
# Load libraries
library("Seurat")
library("dplyr")
library("ggplot2")
library("R.matlab")
library("openxlsx")

# Load Matlab object
data <- readMat("./GSM3017261_150000_CNS_nuclei.mat")

# Keep the matrix that contains the expression values
expression.mat <- t(data$DGE)

# Rename the clusters to remove trailing space
cluster.assignment <- sapply(data$cluster.assignment[,1], trimws)
organ <- sapply(data$sample.type[,1], trimws)
barcode <- paste0("Cell-", data$barcodes[1,])

# Prepare the cell metadata
coldata <- data.frame(barcode = barcode, organ = organ, cluster_full_name =
cluster.assignment)
coldata$cluster_number <- unlist(lapply(sapply(coldata$cluster_full_name, strsplit, split =
" ", fixed = T), function(x) as.integer(x[1])))

# Add extra info to the coldata object
extra.info <- read.xlsx("./splitseq_clusters_no_unknown.xlsx")
extra.info$cluster_full_name <- NULL
coldata <- coldata %>% left_join(extra.info, by = "cluster_number")
row.names(coldata) <- coldata$barcode

# Prepare the genes
genes <- sapply(data$genes[,1], trimws)

# Add info to matrix
colnames(expression.mat) <- coldata$barcode
rownames(expression.mat) <- genes

# Create Seurat object
sr <- CreateSeuratObject(counts = expression.mat, project = "splitseq_paper", min.cells = 3)
sr <- AddMetaData(sr, coldata)

# Cleanup
rm(coldata, data, expression.mat, extra.info, cluster.assignment, genes, organ, barcode)

# Remove clusters from tissue not relevant to my study
sr <- subset(sr, subset = keep == "yes")
sr$keep <- NULL

# Split the object and keep the P2 and P11 brain only
sr.p2 <- subset(sr, subset = organ == "p2_brain")
sr.p11 <- subset(sr, subset = organ == "p11_brain")
sr.list <- list(sr.p2, sr.p11)

# Cleanup
rm(sr, sr.p2, sr.p11)
```

```

# Normalize datasets individually by SCTransform()
sr.list <- lapply(X = sr.list, FUN = SCTransform, method = "glmGamPoi")

# Select the integration features
features <- SelectIntegrationFeatures(object.list = sr.list, nfeatures = 3000)

# Run the PrepSCTIntegration() function prior to identifying anchors
sr.list <- PrepSCTIntegration(object.list = sr.list, anchor.features = features)

# When running FindIntegrationAnchors(), and IntegrateData(),
# set the normalization.method parameter to the value SCT.
int.anchors <- FindIntegrationAnchors(
  object.list = sr.list,
  normalization.method = "SCT",
  anchor.features = features)

sr_integrated <- IntegrateData(
  anchorset = int.anchors,
  normalization.method = "SCT")

# Cleanup
sr_ref <- sr_integrated
rm(sr.list, features, int.anchors, sr_integrated)

# Run PCA and UMAP on the data
sr_ref <- sr_ref %>%
  RunPCA() %>%
  RunUMAP(dims = 1:30, return.model = TRUE)

# Save the pre-processed integrated reference
saveRDS(sr_ref, "./sr_integrated_reference.rds")
```

### 7.7.3 Main analysis in Seurat

```
# Load packages
library("Seurat")
library("openxlsx")
library("ggplot2")
library("ggrepel")
library("RColorBrewer")
library("tidyverse")
library("ensembldb")
library("AnnotationHub")
library("scProportionTest")
library("clusterProfiler")
library("cowplot")

# Set the time point for the whole script
TIMEPOINT <- "20dpi"

#### Create Seurat object ####

# Samples info
samples <- read.xlsx("samples.xlsx", detectDates = T)

# Subset the samples
samples <- subset(samples, subset = timepoint == TIMEPOINT)

# Function to load the Seurat objects and add metadata
load_seurat_object <- function(samples) {
  timepoint <- samples[1]
```

```r
  animal <- samples[2]
  inocula <- samples[3]
  date <- samples[4]

  # Create the Seurat object
  sr <- readRDS(paste0("./seurat_objects/", animal, ".rds"))

  # Add metadata
  sr[["timepoint"]] <- timepoint
  sr[["animal"]] <- animal
  sr[["inocula"]] <- inocula
  sr[["date"]] <- date

  # Set identities for each cell
  Idents(sr) <- paste(timepoint, inocula, animal, sep = "_")

  # Return the object to be saved in a list
  sr
}

# Run the function to save the objects in the directory
sr_objects <- apply(samples, 1, load_seurat_object)

# Now we need to merge the objects in a new object
sr_merged <- merge(x = sr_objects[[1]], y = sr_objects[-1])

# Cleanup
sr <- sr_merged
rm(samples, load_seurat_object, sr_objects, sr_merged)


#### Rename features ####

## Load the annotation resource.
ah <- AnnotationHub()

# fetch one of the databases
# ahDb <- query(ah, pattern = c("Mus musculus", "EnsDb", 104))
ahEdb <- ah[["AH95775"]]

# Create one vector with the Ensembl IDs of all the genes from the experiment
ensembl.genes <- row.names(sr[["RNA"]])

# Convert the Ensembl IDs to Gene symbols
gene_ids <- ensembldb::select(ahEdb, keys= ensembl.genes, keytype = "GENEID", columns = c("SYMBOL","GENEID"))

# Some Ensembl IDs don't have corresponding gene symbols and will be removed
empty_genes <- which(gene_ids$SYMBOL == "")
gene_ids <- gene_ids[-empty_genes,]
sr <- sr[-empty_genes,]

# There might be duplicate names in the symbols. Add a suffix to make them unique
unique_gene_ids <- make.unique(gene_ids$SYMBOL)

# Replace underscores with dashes because underscores are not allowed in Seurat
unique_gene_ids <- gsub("_", "-", unique_gene_ids, fixed = T, )

# Function to remove rows that could not be matched and rename
# the RNA assay slot of the Seurat object
RenameGenesSeurat <- function(obj, newnames) {
  RNA <- obj@assays$RNA

  if (nrow(RNA) == length(newnames)) {

```

```r
  # Rename features
  if (length(RNA@counts)) RNA@counts@Dimnames[[1]]            <- newnames
  if (length(RNA@data)) RNA@data@Dimnames[[1]]                <- newnames

} else {
  stop("Unequal gene sets: nrow(RNA) != nrow(newnames)")
}
obj@assays$RNA <- RNA

  # Fix the row.names in meta.features
  row.names(obj[["RNA"]]@meta.features) <- row.names(obj[["RNA"]])
  return(obj)
}

# Prepare the renamed object
sr_renamed <- RenameGenesSeurat(sr, unique_gene_ids)

# Save
saveRDS(sr_renamed, paste0("./seurat_objects/", TIMEPOINT, "_sr_renamed.rds"))

# Cleanup
sr <- sr_renamed
rm(ah, ahDb, ahEdb, gene_ids, empty_genes, ensembl.genes, unique_gene_ids, RenameGenesSeurat, sr_renamed)

#### QC ####

# Plot number of features and counts
VlnPlot(sr, features = c("nFeature_RNA", "nCount_RNA"), pt.size = 0)

# Calculate mitochondrial genes percentage
sr[["percent.mt"]] <- PercentageFeatureSet(sr, pattern = "^mt-")

# Filter cells with fewer than 200 expressed genes or more than 2500
# Filter cells that have >1% mitochondrial counts
sr_qc <- subset(sr, subset = nFeature_RNA > 250 & nFeature_RNA < 2500 & percent.mt < 1)

# Add cell cycle genes information
# Basic function to convert human to mouse gene names
convert_genes_human_to_mouse <- function(x){
  human = useMart("ensembl", dataset = "hsapiens_gene_ensembl")
  mouse = useMart("ensembl", dataset = "mmusculus_gene_ensembl")
  genesV2 = getLDS(attributes = c("hgnc_symbol"), filters = "hgnc_symbol", values = x , mart = human, attributesL = c("mgi_symbol"), martL = mouse, uniqueRows=T)
  mouse_genes <- unique(genesV2[, 2])
  # Print the first 6 genes found to the screen
  print(head(mouse_genes))

  mouse_genes
}

s.genes <- convert_genes_human_to_mouse(cc.genes.updated.2019$s.genes)
g2m.genes <- convert_genes_human_to_mouse(cc.genes.updated.2019$g2m.genes)

# Check if cells separate by cell cycle phase
dir.create("plots/cell_cycle", recursive = T, showWarnings = F)

sr_phase <- sr_qc
sr_phase <- CellCycleScoring(sr_phase, s.features = s.genes, g2m.features = g2m.genes, set.ident = TRUE)
sr_phase <- NormalizeData(sr_phase)
sr_phase <- FindVariableFeatures(sr_phase)
sr_phase <- ScaleData(sr_phase, features = rownames(sr_phase))
sr_phase <- RunPCA(sr_phase, features = c(s.genes, g2m.genes))
DimPlot(sr_phase, shuffle = TRUE) +
  ggtitle(paste0(TIMEPOINT, " cell cycle PCA"))
ggsave(paste0("plots/cell_cycle/", TIMEPOINT, "_PCA.png"), width = 8, height = 4)

# Assign cell cycle scores
sr_qc <- CellCycleScoring(sr_qc, s.features = s.genes, g2m.features = g2m.genes)

# Plot number of features and counts
VlnPlot(sr_qc, features = c("nFeature_RNA", "nCount_RNA"), pt.size = 0)

# Save
saveRDS(sr_qc, paste0("./seurat_objects/", TIMEPOINT, "_sr_renamed_qc.rds"))

# Cleanup
sr <- sr_qc
rm(sr_qc, s.genes, g2m.genes, sr_phase, convert_genes_human_to_mouse)


#### Integrate datasets ####

# Split the dataset into a list of two seurat objects based on inocula
sr.list <- SplitObject(sr, split.by = "inocula")

# Remove the PBS group
#sr.list <- sr.list[-1]

# Normalize datasets individually by SCTransform()
sr.list <- lapply(X = sr.list, FUN = SCTransform, method = "glmGamPoi")

# Select the integration features
features <- SelectIntegrationFeatures(object.list = sr.list, nfeatures = 3000)

# Run the PrepSCTIntegration() function prior to identifying anchors
sr.list <- PrepSCTIntegration(object.list = sr.list, anchor.features = features)

# When running FindIntegrationAnchors(), and IntegrateData(),
# set the normalization.method parameter to the value SCT.
int.anchors <- FindIntegrationAnchors(
  object.list = sr.list,
  normalization.method = "SCT",
  anchor.features = features)

sr_integrated <- IntegrateData(
  anchorset = int.anchors,
  normalization.method = "SCT")

# Save
saveRDS(sr_integrated, paste0("./seurat_objects/", TIMEPOINT, "_sr_renamed_qc_integrated.rds"))

# Cleanup
sr <- sr_integrated
rm(sr.list, features, int.anchors, sr_integrated)

#### Annotation (reference integration) ####

# Load pre-processed reference
sr_ref <- readRDS("../splitseq_paper_reference/sr_integrated_reference.rds")

query <- sr
rm(sr)

# Find the transfer anchors between the two datasets
query.anchors <- FindTransferAnchors(
  reference = sr_ref,
  query = query,
  dims = 1:50,
  reference.reduction = "pca",
  normalization.method = "SCT"
)

# Make a vector with the metadata to transfer from the reference to the query
labels_to_transfer <- list(
  cluster_number = "cluster_number"
)

# Do the transfer
query <- TransferData(
  reference = sr_ref,
  query = query,
  anchorset = query.anchors,
  refdata = labels_to_transfer,
  dims = 1:50
)

# Add extra info to the coldata object
extra.info <- read.xlsx("../splitseq_paper_reference/splitseq_clusters.xlsx")

df <- data.frame(cluster_number = as.integer(query$predicted.cluster_number)) %>%
  left_join(extra.info, by = "cluster_number")

df$cluster_number <- NULL
df$keep <- NULL

row.names(df) <- colnames(query)

query <- AddMetaData(query, df)

# Calculate mapping score and add to metadata
query <- AddMetaData(
  object = query,
  metadata = MappingScore(anchors = query.anchors),
  col.name = "mapping.score"
)

# Cleanup
rm(df, extra.info, labels_to_transfer, query.anchors)

# Assess the score of the predictions
ggplot() + aes(query$predicted.cluster_number.score) + geom_histogram()
ggplot() + aes(query$mapping.score) + geom_histogram()

# Filter the query object based on label transfer quality
query.filt <- subset(query, region != "Olfactory Bulb")

# Number of cells in each cluster
cells_per_cluster <- query.filt@meta.data %>%
  group_by(cluster_full_name) %>%
  count() %>%
  arrange(n)

# Select the clusters with fewer than 100 cells
clusters_to_keep <- cells_per_cluster[cells_per_cluster$n > 100,]
clusters_to_keep <- clusters_to_keep$cluster_full_name

# Filter query to remove clusters with fewer than 100 cells
query.filt <- subset(query.filt, subset = cluster_full_name %in% clusters_to_keep)

# Re-cluster the query
query.filt <- query.filt %>%
  SCTransform(method = "glmGamPoi") %>%
```

```r
    RunPCA() %>%
    RunUMAP(dims = 1:30)

  # Save
  saveRDS(query, paste0("./seurat_objects/", TIMEPOINT, "_sr_renamed_qc_integrated_annotated.rds"))
  saveRDS(query.filt, paste0("./seurat_objects/", TIMEPOINT, "_sr_renamed_qc_integrated_annotated_filtered.rds"))


  # Cleanup
  sr <- query.filt
  rm(query, query.filt, clusters_to_keep)

  # Sanity check after annotation
  # Check that the annotated clusters use the gene markers
  dir.create("plots/marker_genes", showWarnings = F, recursive = T)

  # Relevel clusters by cluster id so that the plots are nicer
  sr$cluster_full_name <- factor(sr$cluster_full_name,
                                 levels = unique(sr$cluster_full_name)[order(as.integer(str_extract(unique(sr$cluster_full_name), "^\\d+")),
                                                                      decreasing = T)])

  VlnPlot(sr, features = c("Gria1", "Snhg11", "Mbp", "Plp1", "Vcan", "Dock8", "Flt1", "Slc1a2", "Plpp3", "Dnah11"),
          pt.size = 0, stack = T, group.by = "cluster_full_name")

  ggsave(paste0("plots/marker_genes/", TIMEPOINT, "_marker_genes.png"), width = 12, height = 8)

  # Astro: Aqp4, Slc1a2, Plpp3, Gja1
  # Oligodendrocytes: Mbp, Plp1
  # Oligodendrocyte Precursor Cells: Vcan & Mbp, Pdgfra
  # Endothelial/smooth muscle Cells: Rgs5, Flt1, Ly6c1, Pltp
  # Microglia/macrophages: Dock2, Dock8, Csf1r, P2ry12
  # Ependymal cells: Dnah11
  # Neurons: Gria1, Snhg11?

  dir.create("./cluster_metrics", showWarnings = F)

  # Number of cells in each cluster
  cells_per_cluster <- sr@meta.data %>%
    group_by(cluster_full_name) %>%
    count() %>%
    arrange(n)
  write.table(cells_per_cluster, paste0("./cluster_metrics/", TIMEPOINT, "_cells_per_cluster.tsv"))

  # Number of cells in each group
  cells_per_group <- data.frame(table(sr$group, sr$inocula, sr$animal))
  names(cells_per_group) <- c("group", "inocula", "animal", "n_cells")
  write.table(cells_per_group, paste0("./cluster_metrics/", TIMEPOINT, "_cells_per_group.tsv"))

  # Calculate mean and sd of number of features per cluster
  nFeatures_per_cluster <- sr@meta.data %>%
    group_by(cluster_full_name) %>%
    summarise_at(vars(nFeature_RNA ),list(mean = ~round(mean(.),0), median = median, sd = ~round(sd(.),0))) %>%
    arrange(median)

  # Calculate number of counts per cluster
```

```
nCounts_per_cluster <- sr@meta.data %>%
  group_by(cluster_full_name) %>%
  summarise_at(vars(nCount_RNA ),list(mean = ~round(mean(.),0), median = median, sd = ~round(sd(.),0))) %>%
  arrange(median)

extra_metrics <- nCounts_per_cluster %>%
  left_join(nFeatures_per_cluster, by = "cluster_full_name")
colnames(extra_metrics) <- c("Cluster name", "Counts mean", "Counts median", "Counts SD",
                             "Features mean", "Features median", "Features SD")
write.xlsx(extra_metrics, paste0("./cluster_metrics/", TIMEPOINT, "_extra_metrics.xlsx"), overwrite = T)

# Plots
dir.create("./plots/reduced_dimensions", showWarnings = F, recursive = T)

DimPlot(sr, group.by = "cluster_full_name", label = T, repel = T) +
  ggtitle(TIMEPOINT)
ggsave(paste0("./plots/reduced_dimensions/", TIMEPOINT, "_UMAP.png"), width = 16, height = 10)

DimPlot(sr, group.by = "cluster_full_name", split.by = "inocula") +
  ggtitle(TIMEPOINT)
ggsave(paste0("./plots/reduced_dimensions/", TIMEPOINT, "_split_inocula_UMAP.png"), width = 16, height = 10)


#### Cell type proportions ####

pt <- table(sr$group, sr$inocula)
pt <- as.data.frame(pt)
colnames(pt) <- c("Cell type", "Experimental group", "Frequency")
pt$`Cell type` <- as.character(pt$`Cell type`)

dir.create("./plots/celltype_proportions", showWarnings = F, recursive = T)

myColors <- brewer.pal(10, "Set3")
names(myColors) <- c("Migrating Interneurons",
                     "Cortical Neurons",
                     "Medium Spiny Neurons",
                     "Astrocytes",
                     "OPC",
                     "Oligodendrocytes",
                     "VLMC",
                     "Ependymal",
                     "Immune",
                     "Vascular")

ggplot(pt, aes(x = `Experimental group`, y = Frequency, fill = `Cell type`)) +
  geom_col(position = "fill", width = 0.5) +
  scale_fill_manual(name = "Cell type", values = myColors) +
  ylab("Proportion") +
  ggtitle(TIMEPOINT)
ggsave(paste0("./plots/celltype_proportions/", TIMEPOINT, "_cell_proportions.png"), width = 8, height = 4)

## Plots using the permutation test
prop_test <- sc_utils(sr)
prop_test <- permutation_test(
  prop_test,
  cluster_identity = "group",
  sample_1 = "CD1",
  sample_2 = "RML",
  sample_identity = "inocula"
)
```

```
permutation_plot(prop_test, log2FD_threshold = log2(1.2)) +
  ylab("Log2-fold difference in cell numbers") +
  xlab("Cell type") +
  ggtitle(TIMEPOINT)
ggsave(paste0("./plots/celltype_proportions/", TIMEPOINT, "_cell_proportions_scPropTest.png"), width = 8, height = 4)

# Repeat for CD1 vs PBS
prop_test <- sc_utils(sr)
prop_test <- permutation_test(
  prop_test,
  cluster_identity = "group",
  sample_1 = "PBS",
  sample_2 = "CD1",
  sample_identity = "inocula"
)
permutation_plot(prop_test, log2FD_threshold = log2(1.2)) +
  ylab("Log2-fold difference in cell numbers") +
  xlab("Cell type") +
  ggtitle(TIMEPOINT)
ggsave(paste0("./plots/celltype_proportions/", TIMEPOINT, "_PBS_vs_CD1_cell_proportions_scPropTest.png"), width = 8, height = 4)

# Test only groups of neurons
sr_neurons <- subset(sr, subset = group %in% c("Medium Spiny Neurons", "Cortical Neurons", "Migrating Interneurons"))

prop_test <- sc_utils(sr_neurons)
prop_test <- permutation_test(
  prop_test,
  cluster_identity = "cluster_full_name",
  sample_1 = "CD1",
  sample_2 = "RML",
  sample_identity = "inocula"
)
# Reorder the data to have the clusters in order for the plot
prop_test@results$permutation$clusters <-
  factor(prop_test@results$permutation$clusters,
         levels = prop_test@results$permutation$clusters[order(as.integer(str_extract(prop_test@results$permutation$clusters, "^\\d+")), decreasing = T)])

permutation_plot(prop_test, log2FD_threshold = log2(1.2), order_clusters = F) +
  ylab("Log2-fold difference in cell numbers") +
  xlab("Cell cluster") +
  ggtitle(TIMEPOINT)
ggsave(paste0("./plots/celltype_proportions/", TIMEPOINT, "_neurons_cell_proportions_scPropTest.png"), width = 8, height = 4)

rm(prop_test, pt, myColors, sr_neurons)

#### DGE ####
# Function to perform DGE between clusters and two conditions
get_DEGs <- function(cluster, condition1, condition2, seurat_obj){
  genes <- tryCatch(
    {
      FindMarkers(seurat_obj,
                       ident.1 = paste0(cluster, "_", condition1),
                       ident.2 = paste0(cluster, "_", condition2)) %>%
      rownames_to_column(var = "gene")
    }, error = function(cond) return (NULL))

  if(!is.null(genes) && nrow(genes) > 0) {
    cbind(cluster = cluster, genes)
  }
```

```r
  }

  # Prepare a vector of all all cluster names
  all_clusters <- unique(as.character(sr$cluster_full_name))

  # Add new Idents to Seurat object
  Idents(sr) <- paste0(sr$cluster_full_name, "_", sr$inocula)

  # Run DGE on all clusters
  degs <- map_dfr(all_clusters, get_DEGs, condition1 = "RML", condition2 = "CD1", seurat_obj = sr)

  # Plot the distribution of the p-values
  ggplot(degs, aes(p_val_adj)) + geom_histogram()

  # Keep DEGs with adjusted p-values < 0.05
  degs.filtered <- subset(degs, subset = p_val_adj < 0.05)

  # Add info if DEG is unique in each cluster
  add_unique_info <- function(row, degs.filtered) {
    current_cluster <- row["cluster"]
    genes <- subset(degs.filtered, subset = cluster != current_cluster)[, "gene"]
    gene <- row["gene"]

    return(!gene %in% genes)
  }
  degs.filtered$gene_unique <- apply(degs.filtered, 1, add_unique_info, degs.filtered = degs.filtered)

  # Number of DEGs in each cluster
  table(degs.filtered$cluster)

  # Bar chart to visualize the number of DEGs in each cluster
  dir.create("./plots/DGE", showWarnings = F, recursive = T)
  ggplot(as.data.frame(table(degs.filtered$cluster)), aes(Var1, Freq)) + geom_col() + coord_flip() +
    xlab("Clusters") + ylab("Number of DEGs") + ggtitle(paste0(TIMEPOINT, " number of DEGs (adj_p_val < 0.05)"))
  ggsave(paste0("./plots/DGE/", TIMEPOINT, "_number_of_DEGs_by_cluster.png"), width = 12, height = 8)


  # Plot number of DEGs vs number of cells in cluster
  tb <- cells_per_cluster %>% right_join(as.data.frame(table(degs.filtered$cluster)), by = c("cluster_full_name" = "Var1"))

  ggplot(tb, aes(n, Freq, label=cluster_full_name)) +
    geom_point() +
    xlab("number of cells") +
    ylab("number of DEGs") +
    geom_text_repel(max.overlaps = 20)
  ggsave(paste0("./plots/DGE/", TIMEPOINT, "_number_of_DEGs_vs_cells.png"), width = 12, height = 12)

  # Save gene list
  dir.create("./DGE_gene_lists", showWarnings = F, recursive = T)
  write.xlsx(degs.filtered, paste0("./DGE_gene_lists/", TIMEPOINT, "_DEGs_by_cluster.xlsx"), overwrite = T)

  # Compare CD1 vs PBS
  degs_contr <- map_dfr(all_clusters, get_DEGs, condition1 = "CD1", condition2 = "PBS", seurat_obj = sr)
  degs.filtered_contr <- subset(degs_contr, subset = p_val_adj < 0.05)
```

```
 write.xlsx(degs.filtered_contr, paste0("./DGE_gene_lists/", TIMEPOINT, "_CD1_vs_PBS_DEGs_by_cluster.xlsx"), overwrite = T)
```

### 7.7.4 Pseudobulk differential gene expression

```
# Load libraries
library("Seurat")
library("ggplot2")
library("tidyverse")
library("openxlsx")
library("SingleCellExperiment")
library("DESeq2")

#### Pseudobulk analysis ####
timepoints <- c("20dpi", "40dpi", "80dpi", "120dpi", "end")

degs_list <- list()

for (i in seq_along(timepoints)) {

  # Set the timepoint variable
  TIMEPOINT <- timepoints[i]

  # Load the Seurat file for the timepoint
  sr <- readRDS(paste0("../seurat_objects/", TIMEPOINT, "_sr_renamed_qc_integrated_annotated_filtered.rds"))

  # Keep only the CD1 and RML samples
  sr_RML_CD1 <- subset(sr, subset = inocula %in% c("RML", "CD1"))

  # Cleanup
  rm(sr)

  # Function to run DESeq2 for each cluster and generate
  # relevant plots
  run_DESeq <- function(current_cluster, seurat_obj) {

    print(paste0("Working on cluster: ", current_cluster))

    # Subset again to select the cluster of interest
    sr_cluster <- subset(seurat_obj, subset = cluster_full_name == current_cluster)

    # Convert Seurat object to SingleCellExperiment
    sce <- as.SingleCellExperiment(sr_cluster)

    # Convert characters to factors
    sce$animal <- factor(sce$animal)
    sce$inocula <- factor(sce$inocula, levels = c("CD1", "RML"))

    # Count aggregation to sample level
    sce_agg <- Matrix.utils::aggregate.Matrix(t(counts(sce)),
                           groupings = sce$animal,
                           fun = "sum")

    # Transpose the matrix
    sce_agg <- t(sce_agg)

    # Prepare the metadata
    sce_metadata <- data.frame(animal = as.numeric(colnames(sce_agg))) %>%
      left_join(read.xlsx("../samples.xlsx"), by = "animal") %>%
      column_to_rownames("animal")

```

```r
    # Build the DESeq2 object
    dds <- DESeqDataSetFromMatrix(sce_agg,
                                  colData = sce_metadata,
                                  design = ~ inocula)


    # Transform counts for data visualization
    rld <- rlog(dds, blind=TRUE)

    # Plot PCA
    dir.create("PCA_plots", showWarnings = F)

    pca_plot <- DESeq2::plotPCA(rld, intgroup = "inocula")
    ggsave(paste0("./PCA_plots/", TIMEPOINT, "_", gsub("/", "-", current_cluster, fixed = T), ".png"), pca_plot, width = 10, height = 10)

    # Run the DESeq2 pipeline
    dds <- DESeq(dds)

    # Get the results
    res <- results(dds,
                   contrast = c("inocula", "RML", "CD1"),
                   alpha = 0.05)

    # Shrink lfc
    res <- lfcShrink(dds,
                     coef = "inocula_RML_vs_CD1",
                     res = res)

    # Significant DE genes
    res_sig <- data.frame(res) %>%
      filter(padj < 0.05) %>%
      arrange(padj) %>%
      rownames_to_column("gene")


    # Heatmap of the significant genes
    if (nrow(res_sig) >= 2) {
      save_pheatmap_png <- function(x, filename, width=1200, height=1000, res = 150) {
        png(filename, width = width, height = height, res = res)
        grid::grid.newpage()
        grid::grid.draw(x$gtable)
        dev.off()
      }

        dir.create("DEGs_heatmaps", showWarnings = F)

        # Extract normalized counts for only the significant genes
        sig_norm <- data.frame(counts(dds, normalized = TRUE)) %>%
          rownames_to_column(var = "gene") %>%
          dplyr::filter(gene %in% res_sig$gene) %>%
          select(-gene)

        hm_anno <- sce_metadata[,"inocula", drop = F]
        row.names(hm_anno) <- colnames(sig_norm)

        # Run pheatmap using the metadata data frame for the annotation
        hm <- pheatmap::pheatmap(sig_norm,
                 color = RColorBrewer::brewer.pal(6, "YlOrRd"),
                 border_color = NA,
                 cluster_rows = T,
                 show_rownames = F,
                 annotation = hm_anno,
                 scale = "row")

      save_pheatmap_png(hm, paste0("./DEGs_heatmaps/", TIMEPOINT, "_", gsub("/", "-", current_cluster, fixed = T), ".png"))
    }

    # Return the results
    if(!is.null(res_sig) && nrow(res_sig) > 0) {
      cbind(cluster = current_cluster, res_sig)
    }
  }

  # Run DGE on all clusters
  all_clusters <- unique(sr_RML_CD1$cluster_full_name)
  degs <- map_dfr(all_clusters, run_DESeq, seurat_obj = sr_RML_CD1)

  degs_list[[TIMEPOINT]] <- degs
}

# Cleanup
rm(i, degs, sr_RML_CD1)

# Filter out CD1 vs PBS genes
genes_to_exclude <- c("Calm1", "Cdk8", "Cmss1", "Malat1", "mt-Rnr1", "mt-Rnr2", "Rn18s")
list_subt <- lapply(degs_list, function(x) x[!x$gene %in% genes_to_exclude,])

names(list_subt) <- timepoints

# Save as xlsx
write.xlsx(degs_list, "DEGs_DESeq2.xlsx", overwrite = TRUE)
write.xlsx(list_subt, "DEGs_DESeq2_subtracted_v2.xlsx", overwrite = TRUE)

#### PCA plots accross all time points ####

dir.create("PCA_plots/all_timepoints", showWarnings = F, recursive = T)

timepoints <- c("20dpi", "40dpi", "80dpi", "120dpi", "end")

load_sr_objects <- function(TIMEPOINT) {

  # Load the Seurat file for the timepoint
  sr <- readRDS(paste0("../seurat_objects/", TIMEPOINT, "_sr_renamed_qc_integrated_annotated_filtered.rds"))

  # Keep only the CD1 and RML samples
  sr_RML_CD1 <- subset(sr, subset = inocula %in% c("RML", "CD1"))

  # Cleanup
  rm(sr)

  return(sr_RML_CD1)
}

# Load all time points in memory
srs <- sapply(timepoints, load_sr_objects)

# Merge to create a combined object
sr_merged <- merge(srs[[1]], y = srs[-1], project = "mouse_sc")

# Remove assays and save the merged
DefaultAssay(sr_merged) <- "RNA"
sr_merged[["SCT"]] <- NULL
sr_merged[["integrated"]] <- NULL
sr_merged[["prediction.score.cluster_number"]] <- NULL

# Remove spurious sample 828719
sr_merged <- subset(sr_merged, subset = animal != "828719")
```

```r
# Remove 40 and 80 dpi because there are no interesting transcriptomic changes
sr_merged <- subset(sr_merged, subset = timepoint %in% c("20dpi", "120dpi", "end"))
saveRDS(sr_merged, "sr_merged_for_PCA_plots.rds")

# Cleanup
rm(srs, load_sr_objects)

# Load the file for future use
#sr_merged <- readRDS("sr_merged_for_PCA_plots.rds")

# Prepare a vector with all clusters
all_clusters <- unique(sr_merged$cluster_full_name)

# Modified function from the DESeq2 visualisation functions
# to allow specification of different labeling for the timepoints
# adapted from https://github.com/mikelove/DESeq2/blob/master/R/plots.R
plotPCA_custom <- function(object, ntop=500, returnData=FALSE) {
  # calculate the variance for each gene
  rv <- rowVars(assay(object))

  # select the ntop genes by variance
  select <- order(rv, decreasing=TRUE)[seq_len(min(ntop, length(rv)))]

  # perform a PCA on the data in assay(x) for the selected genes
  pca <- prcomp(t(assay(object)[select,]))

  # the contribution to the total variance for each component
  percentVar <- pca$sdev^2 / sum( pca$sdev^2 )

  intgroup.df <- as.data.frame(colData(object)[, c("inocula", "timepoint"), drop=FALSE])
  intgroup.df$animal <- row.names(colData(object))
  intgroup.df$inocula <- factor(intgroup.df$inocula, levels = c("CD1", "RML"))
  intgroup.df$timepoint <- factor(intgroup.df$timepoint, levels = timepoints)

  # assembly the data for the plot
  d <- data.frame(PC1=pca$x[,1], PC2=pca$x[,2], intgroup.df, name=colnames(object))

  if (returnData) {
    attr(d, "percentVar") <- percentVar[1:2]
    return(d)
  }

  # Set custom shapes
  custom_shapes <- c("20dpi" = 15, "120dpi" = 16, "end" = 17)

  ggplot(data=d, aes_string(x="PC1", y="PC2", color="inocula", shape = "timepoint", label="animal")) +
    geom_point(size=3) +
    #geom_text_repel(max.overlaps = 20) +
    scale_shape_manual(values = custom_shapes) +
    xlab(paste0("PC1: ",round(percentVar[1] * 100),"% variance")) +
    ylab(paste0("PC2: ",round(percentVar[2] * 100),"% variance")) +
    coord_fixed()
}

# Function to run DESeq2 for each cluster and generate
# relevant plots
generate_PCA_plot <- function(current_cluster, seurat_obj) {

  print(paste0("Working on cluster: ", current_cluster))

  # Subset to select the cluster of interest
  sr_cluster <- subset(seurat_obj, subset = cluster_full_name == current_cluster)

  # Convert Seurat object to SingleCellExperiment
  sce <- as.SingleCellExperiment(sr_cluster)

  # Convert characters to factors
  sce$animal <- factor(sce$animal)

  # Count aggregation to sample level
  sce_agg <- Matrix.utils::aggregate.Matrix(t(counts(sce)),
                                            groupings = sce$animal,
                                            fun = "sum")

  # Transpose the matrix
  sce_agg <- t(sce_agg)

  # Prepare the metadata
  sce_metadata <- data.frame(animal = as.numeric(colnames(sce_agg))) %>%
    left_join(read.xlsx("../samples.xlsx"), by = "animal") %>%
    column_to_rownames("animal")

  # Build the DESeq2 object
  dds <- DESeqDataSetFromMatrix(sce_agg,
                                colData = sce_metadata,
                                design = ~ inocula)

  # Transform counts for data visualization
  rld <- tryCatch({
    vst(dds)
  }, error = function(cond) varianceStabilizingTransformation(dds))

  # Plot PCA
  pca_plot <- plotPCA_custom(rld) + ggtitle(current_cluster)
  ggsave(paste0("PCA_plots/all_timepoints/", gsub("/", "-", current_cluster, fixed = T), ".png"), width = 6, height = 4)
}

sapply(all_clusters, generate_PCA_plot, seurat_obj = sr_merged)
```

### 7.7.5 Gene Set Enrichment Analysis and Gene Ontology Over-representation Analysis

```r
# Load packages
library("Seurat")
library("openxlsx")
library("ggplot2")
library("ggrepel")
library("RColorBrewer")
library("tidyverse")
library("ensembldb")
library("AnnotationHub")
library("clusterProfiler")

# Load the Seurat object
TIMEPOINT <- "20dpi"
sr <- readRDS(paste0("./seurat_objects/", TIMEPOINT, "_sr_renamed_qc_integrated_annotated_filtered.rds"))

# Load the annotation resource.
ah <- AnnotationHub()

# fetch one of the databases
ahOrgDb <- ah[["AH92582"]]

# Prepare a vector of all all cluster names
all_clusters <- unique(as.character(sr$cluster_full_name))
```

```

## ORA - Over-Representation Analysis

# Create directory
dir.create("cluster_profiler/ORA", showWarnings = F, recursive = T)

run_ORA <- function(cluster_, degs.filtered, seurat_obj, ontology){
  genes <- subset(degs.filtered, subset = cluster == cluster_)$gene
  if (length(genes) == 0) {
    return()
  }
  print(paste0("working on cluster: ", cluster_))
  ego <- enrichGO(gene          = genes,
                  universe      = row.names(seurat_obj),
                  OrgDb         = ahOrgDb,
                  keyType       = "SYMBOL",
                  ont           = ontology,
                  pAdjustMethod = "BH")
  ego <- head(ego)

  if(!is.null(ego) && nrow(ego) > 0) {
    cbind(cluster = cluster_, ego)
  }
}
run_ORA_list <- function(cluster_, degs.filtered, seurat_obj, ontology, count_cutoff){
  genes <- subset(degs.filtered, subset = cluster == cluster_)$gene
  if (length(genes) == 0) {
    return()
  }
  print(paste0("working on cluster: ", cluster_))
  ego <- enrichGO(gene          = genes,
                  universe      = row.names(seurat_obj),
                  OrgDb         = ahOrgDb,
                  keyType       = "SYMBOL",
                  ont           = ontology,
                  pAdjustMethod = "BH")
  if(!is.null(ego)) {
    ego@result <- ego@result[ego@result$Count >= count_cutoff,]
    return(ego)
  }
}

ora.BP_list <- lapply(all_clusters,
                 run_ORA_list,
                degs.filtered = degs.filtered,
                seurat_obj = sr,
                ontology = "BP",
                count_cutoff = 3)
names(ora.BP_list) <- all_clusters
dotplot(merge_result(ora.BP_list), font.size = 12, title = paste0("ORA - BP - ", TIMEPOINT)) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
ggsave(paste0("cluster_profiler/ORA/", "ORA_BP_", TIMEPOINT, ".png"), width = 8, height = 6)

ora.CC_list <- lapply(all_clusters,
                      run_ORA_list,
                      degs.filtered = degs.filtered,
                      seurat_obj = sr,
                      ontology = "CC",
                      count_cutoff = 3)
names(ora.CC_list) <- all_clusters
dotplot(merge_result(ora.CC_list), font.size = 12, title = paste0("ORA - CC - ", TIMEPOINT)) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
ggsave(paste0("cluster_profiler/ORA/", "ORA_CC_", TIMEPOINT, ".png"), width = 8, height = 6)
```

```r

ora.MF_list <- lapply(all_clusters,
                      run_ORA_list,
                      degs.filtered = degs.filtered,
                      seurat_obj = sr,
                      ontology = "MF",
                      count_cutoff = 3)
names(ora.MF_list) <- all_clusters
dotplot(merge_result(ora.MF_list), font.size = 12, title = paste0("ORA - MF - ", TIMEPOINT)) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
ggsave(paste0("cluster_profiler/ORA/", "ORA_MF_", TIMEPOINT, ".png"), width = 8, height = 6)

ora.BP <- map_dfr(all_clusters,
                      run_ORA,
                      degs.filtered = degs.filtered,
                      seurat_obj = sr,
                      ontology = "BP")
  ora.CC <- map_dfr(all_clusters,
                      run_ORA,
                      degs.filtered = degs.filtered,
                      seurat_obj = sr,
                      ontology = "CC")
  ora.MF <- map_dfr(all_clusters,
                      run_ORA,
                      degs.filtered = degs.filtered,
                      seurat_obj = sr,
                      ontology = "MF")

  worksheets <- list(BP = ora.BP, CC = ora.CC, MF = ora.MF)
  write.xlsx(worksheets, paste0("cluster_profiler/ORA/", "ORA_", TIMEPOINT, ".xlsx"), overwrite = T)

  ## GSEA - Gene Set Enrichment Analysis

  # Create directory
  dir.create("cluster_profiler/GSEA", showWarnings = F, recursive = T)

  # Run GSEA for each cluster separately

  run_GSEA_list <- function(cluster_, seurat_obj, ontology){
    print(paste0("working on cluster: ", cluster_))

    # Subset the Seurat object to keep cluster of interest and only RML and CD1 groups
    srTmp <- subset(seurat_obj, subset = cluster_full_name == cluster_ & inocula %in% c("RML", "CD1"))

    # Perform a fast Wilcoxon rank sum test using presto
    gsea.genes <- presto::wilcoxauc(srTmp, group_by = 'inocula')
    gsea.genes <- gsea.genes[which(gsea.genes$group == "RML"),]

    geneList <- gsea.genes$logFC
    names(geneList) <- gsea.genes$feature
    geneList <- sort(geneList, decreasing = TRUE)

    ego <- tryCatch({
      gseGO(geneList     = geneList,
            OrgDb        = ahOrgDb,
            ont          = ontology,
            keyType       = "SYMBOL",
            minGSSize    = 10,
            maxGSSize    = 500,
            pvalueCutoff = 0.05)
    },
    error = function(cond) NULL)

    if(!is.null(ego) && nrow(ego) > 0) {
      return(ego@result)
    }
  }

  gsea.BP <- lapply(all_clusters,
                    run_GSEA_list,
                    seurat_obj = sr,
                    ontology = "BP")
  names(gsea.BP) <- all_clusters

  gsea.MF <- lapply(all_clusters,
                    run_GSEA_list,
                    seurat_obj = sr,
                    ontology = "MF")
  names(gsea.MF) <- all_clusters

  gsea.CC <- lapply(all_clusters,
                    run_GSEA_list,
                    seurat_obj = sr,
                    ontology = "CC")
  names(gsea.CC) <- all_clusters

  # Save the results
  worksheets <- list(BP = bind_rows(gsea.BP, .id = "cluster"),
                     CC = bind_rows(gsea.CC, .id = "cluster"),
                     MF = bind_rows(gsea.MF, .id = "cluster"))
  write.xlsx(worksheets, paste0("cluster_profiler/GSEA/", "GSEA_", TIMEPOINT, "_per_cluster.xlsx"), overwrite = T)

  # Run GSEA for all cells of all clusters
  gsea.genes <- presto::wilcoxauc(subset(sr, subset = inocula %in% c("RML", "CD1")),
                                group_by = 'inocula')
  gsea.genes <- gsea.genes[which(gsea.genes$group == "RML"),]

  geneList <- gsea.genes$logFC
  names(geneList) <- gsea.genes$feature
  geneList <- sort(geneList, decreasing = TRUE)

  run_GSEA_all_clusters <- function(ontology) {
    gsea <- gseGO(geneList = geneList,
                     OrgDb = ahOrgDb,
                     ont = ontology,
                     keyType = "SYMBOL",
                     minGSSize = 10,
                     maxGSSize = 500,
                     pvalueCutoff = 0.05)
    godata <- GOSemSim::godata('org.Mm.eg.db', ont = ontology)
    gsea <- enrichplot::pairwise_termsim(gsea, method="Wang", semData = godata)
    return(gsea)
  }

  gsea.BP <- run_GSEA_all_clusters(ontology = "BP")
  gsea.CC <- run_GSEA_all_clusters(ontology = "CC")
  gsea.MF <- run_GSEA_all_clusters(ontology = "MF")

  # Save results
  worksheets <- list(BP = gsea.BP, CC = gsea.CC, MF = gsea.MF)
  write.xlsx(worksheets, paste0("cluster_profiler/GSEA/", "GSEA_", TIMEPOINT, "_all_clusters.xlsx"), overwrite = T)

  # Save gseaResult objects for the generation of plots
  saveRDS(gsea.BP, paste0("cluster_profiler/GSEA/", "gseaResult_BP_", TIMEPOINT, ".rds"))
  saveRDS(gsea.CC, paste0("cluster_profiler/GSEA/", "gseaResult_CC_", TIMEPOINT, ".rds"))
  saveRDS(gsea.MF, paste0("cluster_profiler/GSEA/", "gseaResult_MF_", TIMEPOINT, ".rds"))

  # Plot
  ridgeplot(gsea.BP) +
    labs(x = "enrichment distribution", y = "GO terms") +
    ggtitle(paste0("GSEA - BP - ", TIMEPOINT))
  ggsave(paste0("cluster_profiler/GSEA/GSEA_BP_", TIMEPOINT, ".png"), width = 10, height = 10)

  ridgeplot(gsea.CC) +
    labs(x = "enrichment distribution", y = "GO terms") +
    ggtitle(paste0("GSEA - CC - ", TIMEPOINT))
  ggsave(paste0("cluster_profiler/GSEA/GSEA_CC_", TIMEPOINT, ".png"), width = 10, height = 14)

  ridgeplot(gsea.MF) +
    labs(x = "enrichment distribution", y = "GO terms") +
    ggtitle(paste0("GSEA - MF - ", TIMEPOINT))
  ggsave(paste0("cluster_profiler/GSEA/GSEA_MF_", TIMEPOINT, ".png"), width = 10, height = 10)


  # Cleanup
  rm(ah, gse.BP, gse.CC, gse.MF, ora.BP, ora.CC, ora.MF,
     worksheets, ahOrgDb, run_GSEA, run_ORA, all_clusters,
     godata_MF, godata_CC, godata_BP)
```

## 7.8 R session information

```
> sessionInfo()
R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.3 LTS

Matrix products: default
BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.8.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8    LC_MESSAGES=C
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats4    stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] UpSetR_1.4.0                DESeq2_1.32.0               SingleCellExperiment_1.14.1
 [4] SummarizedExperiment_1.24.0 MatrixGenerics_1.6.0        matrixStats_0.61.0
 [7] R.matlab_3.6.2              cowplot_1.1.1               clusterProfiler_4.0.5
[10] scProportionTest_0.0.0.9000 AnnotationHub_3.0.2         BiocFileCache_2.0.0
[13] dbplyr_2.1.1                ensembldb_2.16.4            AnnotationFilter_1.16.0
[16] GenomicFeatures_1.44.2      AnnotationDbi_1.56.2        Biobase_2.54.0
[19] GenomicRanges_1.46.1        GenomeInfoDb_1.30.0         IRanges_2.28.0
[22] S4Vectors_0.32.3            BiocGenerics_0.40.0         forcats_0.5.1
[25] stringr_1.4.0               dplyr_1.0.7                 purrr_0.3.4
[28] readr_2.1.1                 tidyr_1.1.4                 tibble_3.1.6
[31] tidyverse_1.3.1             RColorBrewer_1.1-2          ggrepel_0.9.1
[34] ggplot2_3.3.5               openxlsx_4.2.4              SeuratObject_4.0.4
[37] Seurat_4.0.5

loaded via a namespace (and not attached):
  [1] rappdirs_0.3.3                rtracklayer_1.52.1
  [3] scattermore_0.7               R.methodsS3_1.8.1
  [5] bit64_4.0.5                   R.utils_2.11.0
  [7] irlba_2.3.5                   DelayedArray_0.20.0
  [9] data.table_1.14.2             rpart_4.1-15
 [11] KEGGREST_1.34.0               RCurl_1.98-1.5
 [13] generics_0.1.1                RSQLite_2.2.9
 [15] shadowtext_0.0.9              RANN_2.6.1
 [17] future_1.23.0                 bit_4.0.4
 [19] tzdb_0.2.0                    enrichplot_1.12.3
 [21] spatstat.data_2.1-0           xml2_1.3.3
 [23] lubridate_1.8.0               httpuv_1.6.3
 [25] assertthat_0.2.1              viridis_0.6.2
 [27] hms_1.1.1                     promises_1.2.0.1
 [29] fansi_0.5.0                   restfulr_0.0.13
 [31] progress_1.2.2                readxl_1.3.1
 [33] igraph_1.2.9                  DBI_1.1.1
 [35] geneplotter_1.70.0            htmlwidgets_1.5.4
 [37] spatstat.geom_2.3-0           ellipsis_0.3.2
 [39] backports_1.4.0               annotate_1.72.0
 [41] biomaRt_2.48.3                deldir_1.0-6
 [43] vctrs_0.3.8                   ROCR_1.0-11
 [45] abind_1.4-5                   cachem_1.0.6
 [47] withr_2.4.3                   ggforce_0.3.3
 [49] grr_0.9.5                     sctransform_0.3.2
 [51] treeio_1.16.2                 GenomicAlignments_1.28.0
 [53] prettyunits_1.1.1             goftest_1.2-3
 [55] cluster_2.1.2                 DOSE_3.18.3
 [57] ape_5.5                       lazyeval_0.2.2
 [59] crayon_1.4.2                  genefilter_1.74.1
 [61] pkgconfig_2.0.3               tweenr_1.0.2
 [63] nlme_3.1-153                  ProtGenerics_1.24.0
 [65] rlang_0.4.12                  globals_0.14.0
 [67] lifecycle_1.0.1               miniUI_0.1.1.1
 [69] downloader_0.4                filelock_1.0.2
 [71] modelr_0.1.8                  cellranger_1.1.0
 [73] polyclip_1.10-0               lmtest_0.9-39
 [75] Matrix_1.4-0                  aplot_0.1.1
 [77] zoo_1.8-9                     Matrix.utils_0.9.8
 [79] reprex_2.0.1                  ggridges_0.5.3
 [81] pheatmap_1.0.12               png_0.1-7
 [83] viridisLite_0.4.0             rjson_0.2.20
 [85] bitops_1.0-7                  R.oo_1.24.0
 [87] KernSmooth_2.23-20            Biostrings_2.62.0
 [89] blob_1.2.2                    qvalue_2.24.0
 [91] parallelly_1.29.0             gridGraphics_0.5-1
 [93] scales_1.1.1                  memoise_2.0.1
 [95] magrittr_2.0.1                plyr_1.8.6
 [97] ica_1.0-2                     zlibbioc_1.40.0
 [99] scatterpie_0.1.7              compiler_4.1.1
[101] BiocIO_1.2.0                  fitdistrplus_1.1-6
[103] Rsamtools_2.8.0               cli_3.1.0
[105] XVector_0.34.0                listenv_0.8.0
[107] patchwork_1.1.1               pbapply_1.5-0
[109] MASS_7.3-54                   mgcv_1.8-38
[111] tidyselect_1.1.1              stringi_1.7.6
[113] yaml_2.2.1                    GOSemSim_2.18.1
[115] locfit_1.5-9.4                grid_4.1.1
[117] fastmatch_1.1-3               tools_4.1.1
[119] future.apply_1.8.1            parallel_4.1.1
[121] rstudioapi_0.13               gridExtra_2.3
[123] farver_2.1.0                  Rtsne_0.15
[125] ggraph_2.0.5                  digest_0.6.29
[127] BiocManager_1.30.16           shiny_1.7.1
[129] Rcpp_1.0.7                    broom_0.7.10
[131] BiocVersion_3.13.1            later_1.3.0
[133] RcppAnnoy_0.0.19              httr_1.4.2
[135] colorspace_2.0-2              rvest_1.0.2
[137] XML_3.99-0.8                  fs_1.5.2
[139] tensor_1.5                    reticulate_1.22
[141] splines_4.1.1                 yulab.utils_0.0.4
[143] uwot_0.1.11                   tidytree_0.3.6
[145] spatstat.utils_2.2-0          graphlayouts_0.7.2
[147] ggplotify_0.1.0               plotly_4.10.0
[149] xtable_1.8-4                  jsonlite_1.7.2
[151] ggtree_3.0.4                  tidygraph_1.2.0
[153] ggfun_0.0.4                   R6_2.5.1
[155] pillar_1.6.4                  htmltools_0.5.2
[157] mime_0.12                     glue_1.5.1
[159] fastmap_1.1.0                 BiocParallel_1.26.2
[161] interactiveDisplayBase_1.30.0 codetools_0.2-18
[163] fgsea_1.18.0                  utf8_1.2.2
[165] lattice_0.20-45               spatstat.sparse_2.0-0
[167] curl_4.3.2                    leiden_0.3.9
[169] gtools_3.9.2                  zip_2.2.0
[171] GO.db_3.13.0                  survival_3.2-13
[173] munsell_0.5.0                 DO.db_2.9
[175] GenomeInfoDbData_1.2.7        haven_2.4.3
[177] reshape2_1.4.4                gtable_0.3.0
[179] spatstat.core_2.3-2
```