# Characterizing Alzheimer's Disease Pathophysiology with Single-Cell Data

Alice Saparov | asaparov@hms.harvard.edu | 11606784
Jesslyn Ting Yu Felicia Goh | jgoh@hms.harvard.edu | 71606708
Fiona McBride | fiona_mcbride@hms.harvard.edu | 11571778

## Introduction

Alzheimer's Disease (AD) is a progressive disease that causes memory loss and, in the more severe cases, drastically affects individuals' daily life. Currently, it is the most common type of dementia and one of the top ten leading causes of death in America [1]. Alzheimer's is currently thought to be caused by an accumulation of amyloid (protein), which forms plaques around neurons, and tau (protein), which creates tangles within neurons called neurofibrillary tangles. These plaques and neurofibrillary tangles prevent the proper function of neurons and the release of neurotransmitters causing the affected area of the brain to shrink. Many biological and molecular mechanisms have been found to contribute to AD, such as neuroinflammation, oxidative stress, metabolomic pathway defects, and many more are still hypothesized. Thus, while an abundance of research has been conducted to better understand the genetic, molecular, and biological causes of Alzheimer's, and major breakthroughs in understanding have been made, the disease is still not yet fully understood [2]. In this project, we aim to leverage the granularity of single-cell data from AD and control samples in an attempt to identify new cell types or genes that may be associated with Alzheimer's. Since AD is a progressive disease, it is important to study the cell-level alterations, as it is these small changes over time that lead to the full prognosis. We will combine data from multiple brain regions to get a more comprehensive view of how the entire brain is affected. This will provide new potential therapeutic targets and a better understanding of the cellular characterization of the disease.

The analysis of the single-cell Alzheimer's dataset will be split into three modules. The first module, done by Alice, will involve the preliminary steps to obtain and clean the data. This will include merging the single cell count data for the four samples and then performing quality control, batch correction, and clustering on the data. Next, Jesslyn will conduct differential gene expression and gene set enrichment analysis in order to annotate the clusters identified in the previous step by their cell types. This analysis will also elucidate the expression differences between AD and control samples for various cell types, which can then be further investigated and researched. Lastly, Fiona will perform trajectory analysis in the third module. We expect to uncover biologically relevant trajectories within certain cell type clusters as Alzheimer's is a progressive disease. Thus, we expect to observe correlations in the differences in gene expression relative to pseudotime. This trajectory analysis will investigate and compare the different starting points and trajectories generated to gain a better understanding of gene expression changes in Alzheimer's.

**Module 1: Data Preprocessing, QC, Batch Correction, and Clustering**

Alice Saparov | asaparov@hms.harvard.edu | 11606784

**Introduction**

In the first part of the analysis section, I will obtain the raw single-cell data and perform the preliminary cleaning and analysis. This will include creating the initial Seurat object by combining the gene count matrix, barcodes, and features, and adding the relevant metadata. Next, I will perform quality control on the data(using Seurat), batch correction (using Harmony), and clustering (using Seurat) to later identify the various cell types in the brain regions of healthy and Alzheimer's disease patients. I will create various UMAP plots for visualizing the clusters after running PCA on the data and choosing a relevant number of PCs to include in the dimensionality reduction based on an elbow plot.

**Data and Methods**

The input data for the project was obtained from a very recent single-cell study published on Alzheimer's Disease (AD) [3]. Logistically, this dataset was chosen due to its accessibility since the count matrices, feature data, and barcode data were all easy to find online and download [4]. Some single-cell datasets can also be extremely large, and running analysis on these larger datasets did not seem feasible for this project as the data would be more difficult to share and process locally. Thus, it was also advantageous that the total data, when zipped, was 284.2 MB which was reasonable to share and perform analysis on. Ultimately, my group was also interested in attempting to understand more of the mechanisms causing and advancing AD as a lot is still not understood about the biology of the disease.

The data was downloaded from the Gene Expression Omnibus (GEO) which is a functional genomics data repository accessible to the public. It was downloaded as a zip file which contained separate single-cell sequencing count matrices (.mtx), barcodes (.tsv), and features (.tsv) for sequencing data from four individuals [4]. Two of the individuals had AD (Braak stage III-IV) and the two others were controls that were age and gender-matched to the AD individuals. This data was generated using single nuclei samples that were obtained from frozen post-mortem brain samples from the UK brain bank and processed using standard protocols for 10X Genomics sequencing [3]. Analysis of this data was performed in RStudio (2022.07.1+554) primarily using Seurat (4.3.0) [5] for single-cell data manipulation and Harmony (0.1.1) [4] for batch correction. Additional R packages used included ggplot2 (3.4.2), Matrix (1.5-4), and dplyr (1.1.1).

Data

The four samples, which each contained zipped matrix, feature, and barcode files, were read into R matrices and then converted to a Seurat object using the generated count matrices. Using the metadata information provided in the data description [4], sample (a1, a2, a3, and a4) and type (control vs AD) were added to the metadata of each sample's Seurat object. The four individual samples were then merged into a singular Seurat object ('ad') to be used for all downstream steps.

Quality Control and Preprocessing

Prior to beginning any quality control (QC), there were 28,769 cells in the data and four relevant metadata variables: nCount_RNA (number of molecules in each cell), nFeature_RNA (number of genes in each cell), sample (four individual brain samples), and type (disease state - control v AD). The first QC step included visualizing the number of genes in each cell (Fig. 1A) and filtering out potentially dead, dying, or empty cells, which is indicated by a low nFeature_RNA value. This was done by subsetting the data to only include cells that had at least 500 genes present in each cell. Next, the data was normalized using log normalization with a scale factor of 10,000 so that the cells are more comparable. Then, to find genes that were highly variable across the cells, I selected for

2,000 variable features using a variance stabilizing transformation (vst) selection method. The top five variable genes were found to be FLT1, LINC01090, RELN, CNR1, and NPY (Fig. 2), which can later potentially be investigated in the clustering results or subsequent analysis. After this, the data were centered and scaled so the genes would be more comparable with each other, and PCA was performed to reduce the dimensionality. Using an elbow plot (Fig. 1B), I was able to determine the ideal number of principal components (PC) to use based on where the standard deviation began to plateau. This occurred between about 20-25 PCs, so I chose to proceed with using 20 PCs.

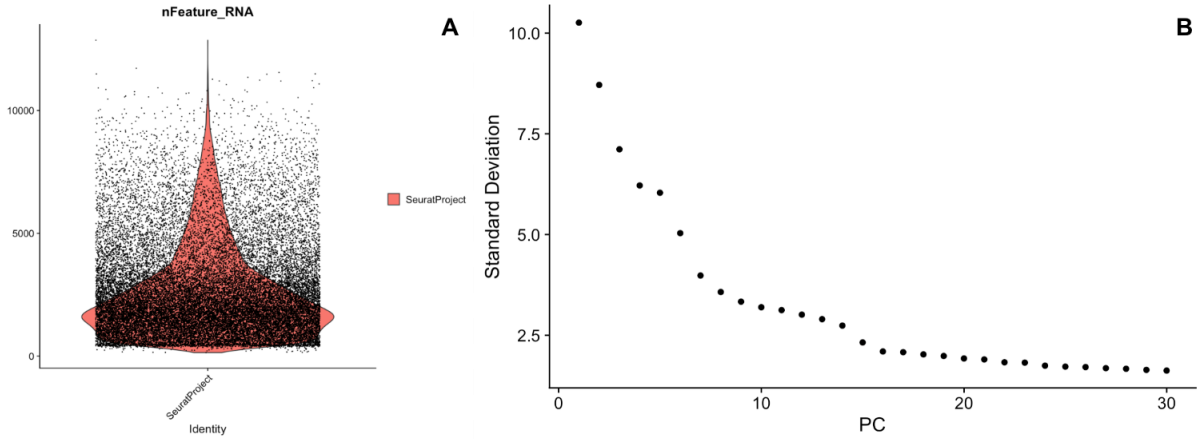



**Figure 1. Quality control plots to assist in decision-making.** (A) Violin plot showing the number of genes found in each cell, low counts may indicate dying, dead, or empty cells. (B) Elbow plot showing the standard deviation (explained) decrease over the top 30 principal components (PCs) after running PCA. The standard deviation plateaus around 20-30 PCs.

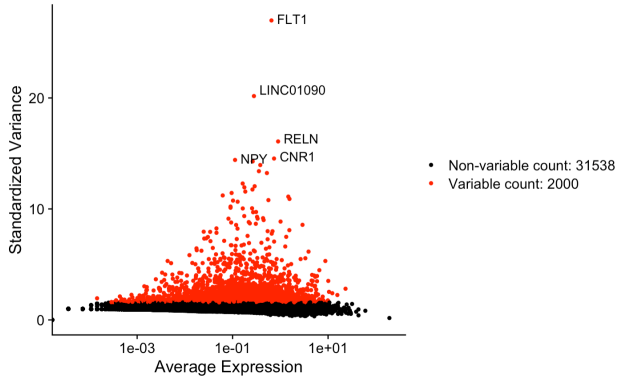



**Figure 2. Top five variable genes.** Standardized variance over the average expression of genes with the five most variable genes labeled. Genes were filtered to extract the top 2,000 variable genes (red), remaining 31,538 genes are considered non-variable (black).

## Batch Correction

In order to determine if batch effects are present in the data, I created a UMAP using 20 PCs (as previously identified) to be able to visualize the data in a 2D space. Batch effects arise when differences or variation observed between groups in the data are attributed to technical effects as opposed to biological reasons [6]. Since the purpose of single-cell studies is to extract molecular and biological explanations for variations in the data, it is necessary to correct for any potential batch effects. I generated UMAPs grouped by sample (Fig. 3A) and by type (Fig. 3B), and there appeared to be poor mixing of the groups on the UMAPs for both variables which indicates batch effects are present. In this project we are interested in observing the biological differences between the type (control v AD), so I want to avoid correcting type as this can limit potential biological effects seen in downstream analyses. However, I corrected the batch effects by sample since the variation arising is likely technical due to the forced combination of different samples that were processed separately. Harmony was used to perform the batch correction, and there

was better mixing between the groups for both sample (Fig. 3C) and type (Fig. 3D), indicating that the batch correction effectively resolved some of the technical variation observed.

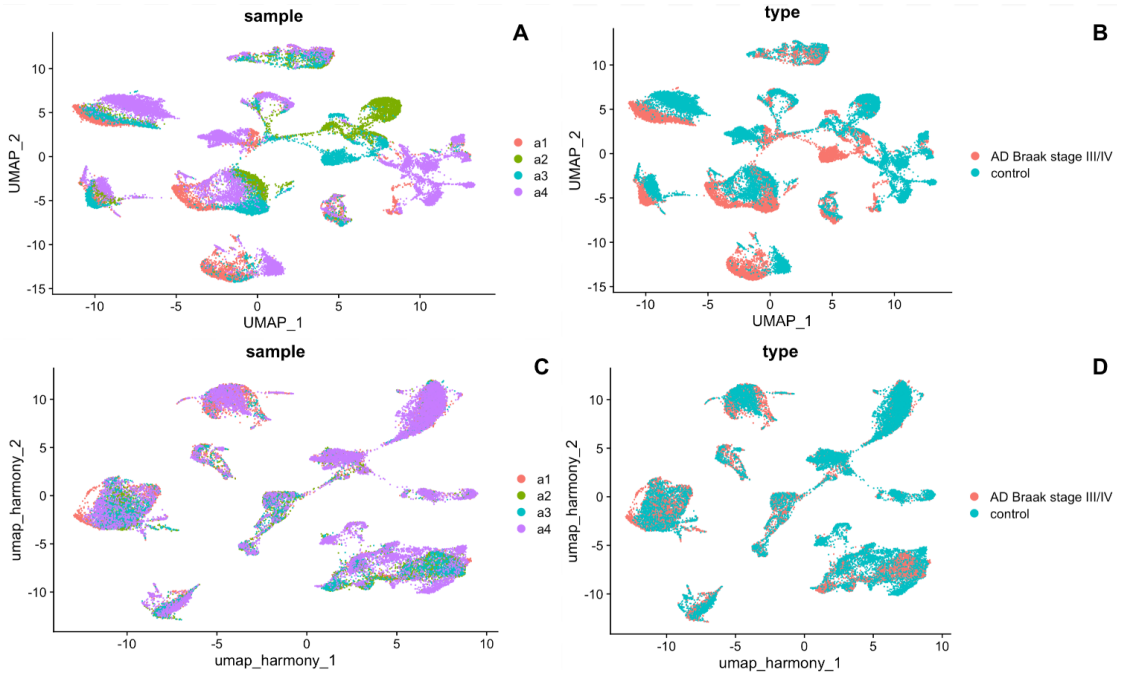



**Figure 3. UMAP plots for evaluation of batch effects and correction.** (A) UMAP plot grouped by sample prior to batch correction showing poor mixing between sample groups indicative of batch effects. (B) UMAP plot grouped by type prior to batch correction showing poor mixing between type groups indicative of batch effects. (C) UMAP plot grouped by sample post Harmony batch correction with better mixing between sample groups. (D) UMAP plot grouped by type post Harmony batch correction with better mixing between type groups.

## Clustering

Now that the data was batch corrected, neighbors and clusters were calculated using the original Louvain algorithm with 20 dimensions. After trying a few different resolution values, I used a resolution of 0.1 to form the clusters since it formed the most similar number of clusters to the original research article [3]. I validated the clustering with a UMAP grouped by cluster and was able to identify 13 different cell-type clusters in the data (Fig. 4). These clusters will be annotated in the next module.

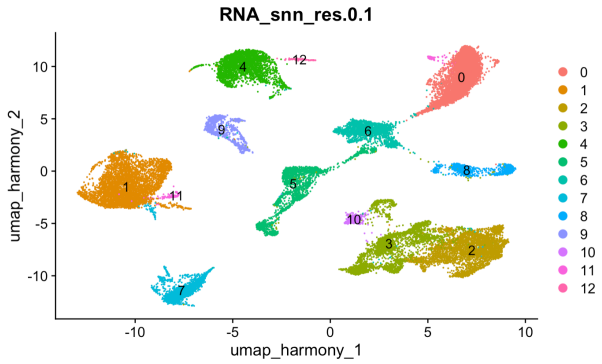



**Figure 4. UMAP plot for validation of clustering.** UMAP plot grouped by cluster resulting in 13 potential cell type clusters identified.

## Results and Discussion

After the initial clustering was completed, I was interested in investigating which clusters the top five variable genes previously found were being expressed in. I created a feature plot of the five genes identified (Fig. 5) and found that FLT1 and RELN appeared to have the most localized expression within specific clusters. These may

suggest the cell types of these clusters, but the cell type annotation of clusters will be performed by Jesslyn in the next module.

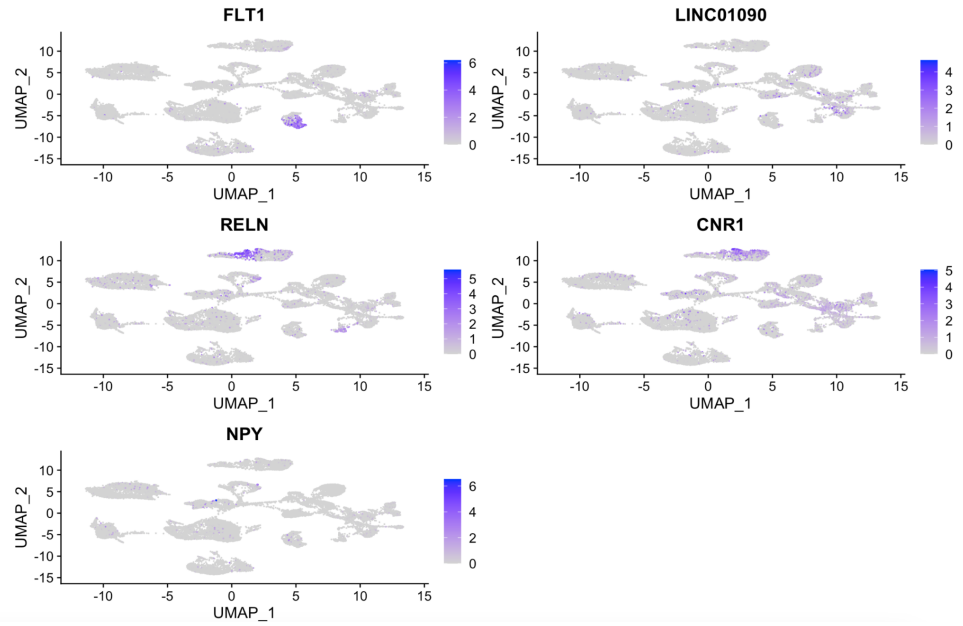



**Figure 5. Feature plots of top five variable genes.** Feature plots showing gene expression of the top five variable genes identified (FLT1, LINC01090, RELN, CNR1, and NPY) relative to cell type clusters generated.

FLT1 was found to be associated with angiogenesis and vasculogenesis and is a member of the VEGFR family [7]. There is also supporting evidence that elevated expression of FLT1 is correlated with cognitive decline and AD [8]. While it appears that the mechanisms of action are not fully understood, FLT1 could serve as a starting point in future AD research. RELN has also been known to be associated with AD [9,10] and could serve as a potential future genetic target. Further analysis into the differential abundance, differential expression, and trajectory of certain genes will be conducted by Jesslyn and Fiona to hopefully uncover some more of the relevant genes and biology to AD. Now, this preprocessed, batch corrected, and clustered data set will serve as the starting point of analysis that Jesslyn will be conducting in Module 2.

**Module 2: Differential Abundance, Differential Expression, and Gene Set Enrichment Analyses**

Jesslyn Ting Yu Felicia Goh | jgoh@hms.harvard.edu | 71606708

In this part of the analysis, I characterize biological differences between brain samples from AD patients and healthy controls via differential abundance, difference expression, and gene set enrichment analyses. Better understanding of cell types or genes that differ most between AD and controls may guide future explorations of novel therapeutic targets. The data used in this module is a direct continuation of the Seurat scRNA-seq data object that has been quality controlled, normalized, batch-corrected, and clustered from the preceding module by Alice.

Before conducting the proposed downstream analyses, I annotated the clusters obtained from the previous module by the corresponding cell type. Since the authors from which our data was acquired did not provide a list of marker genes they used to annotate their cell types, I referenced cell type markers of the human adult brain published in *Nature Biotechnology* by Lake et al. (2017) [11]. I annotated each cluster by the corresponding cell type in two steps:

(1) I generated a gene expression DotPlot (Figure 6A) for the top three markers of each brain cell type provided by Lake et al. [11]. I investigated the top markers of 35 cell types in the brain and assigned each of our clusters to the cell type whose markers it exhibits the most expression. Clusters 2, 11, and 12 did not exhibit high expression for any of the markers, so their annotations are unresolved for now.

(2) I reconfirmed the cell type assignments from Step 1 and resolved any unannotated clusters by matching the cell type that the top 5 marker genes for each cluster corresponded to based on the cell type markers provided by Lake et al. [11].

Following cluster cell type assignments, I replotted the UMAP but colored by the annotated cell type (Figure 6B). Neurons (Inhibitory and Excitatory) are closer to each other towards the right side of the UMAP, whereas glial cells tend to be on the other side (Figure 6B). The position of more similar cell types being closer to each other further confirms the validity of the annotations.

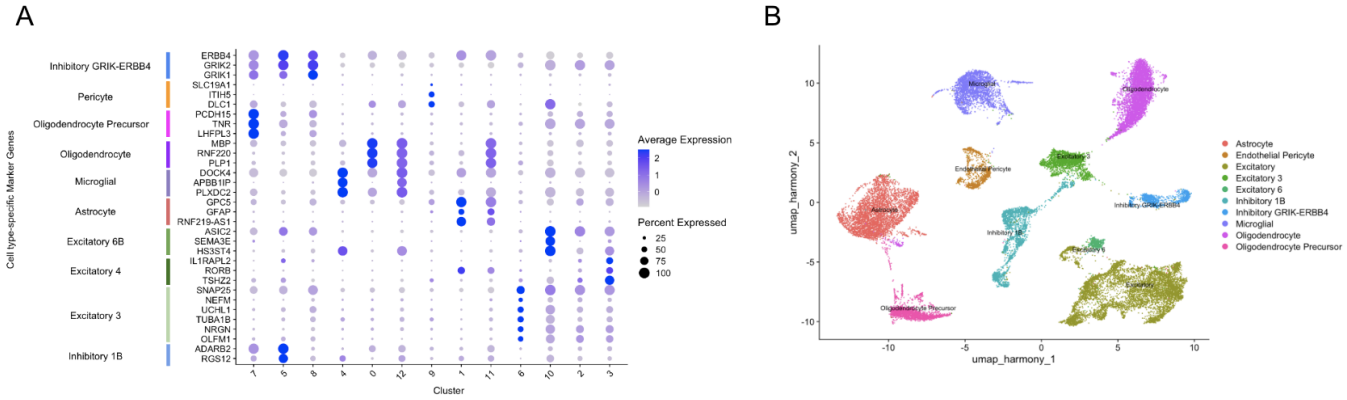



**Figure 6. Cluster cell type annotations based on cell type specific marker genes.** (A) Dotplot for each identified cluster across important cell-type-determining genes provided by Lake et al. (2018). Cell type specific markers for 35 different cell types were investigated, including 13 subtypes of excitatory neurons, 11 subtypes of inhibitory neurons, 2 subtypes of purkinje neurons, granule cells, endothelial pericytes, astrocytes, microglia, oligodendrocytes, and oligodendrocyte precursor cells. Only cell types whose markers are strongly expressed in one of our 13 clusters are shown. (B) UMAP of cells colored by annotated cell type.

Following confident cell type annotations, I conducted differential abundance analysis to investigate how cell type composition differs between AD and control brains. I generated stacked and grouped bar charts to visualize and compare the proportion of each cell type in AD versus control brains (Figure 7). The percent stacked bar plot enables comparison of proportions between cell types within an AD or control brain, whereas the grouped bar chart facilitates comparisons between sample types for a particular cell type. Both panels show a general shift from

higher neuronal to higher glial proportions from control to AD brains. This observation is consistent with current knowledge of AD, where critical processes to neurons are disrupted and result in neuronal death [13]. This explains the consistently higher proportions of both excitatory and inhibitory neurons in the control brain samples compared to the AD samples (Figure 7B). A chi-squared statistical test confirms the statistical significance of the observed differences with $p << 0.001$ except for Inhibitory 1B neurons. On the other hand, glial cells such as microglia, oligodendrocytes, and astrocytes, are higher in AD brains (chi-squared test, $p << 0.001$). Normal functioning astrocytes and microglia are known to serve neuroprotective functions as they respond to inflammatory substances and are responsible for clearing unwanted accumulates in the brain [14-16]. The hallmarks of chronic inflammation and beta-amyloid plaques are thought to be consequences of abnormalities in glial cells that fail to fulfill their duty and are thus continuously signaled [13-16], which explains their higher proportions in AD brains (Figure 7B).

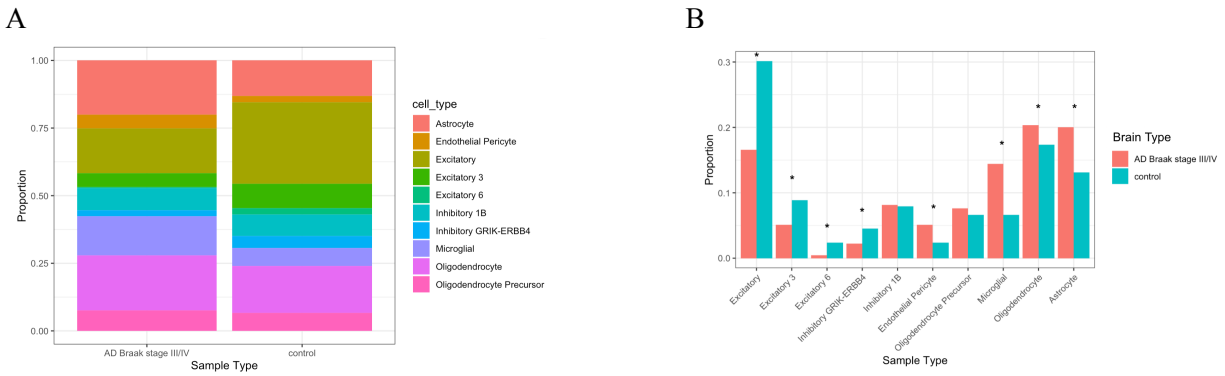



**Figure 7. Differential abundance analysis for annotated cell types between AD and control brains.** (A) Percent stacked bar plot for each sample type (AD or control) colored by each cell type. (B) Grouped bar chart that compares the proportion of each cell type in AD versus controlled brains. * indicates statistically significant difference via the chi-squared test, $p << 0.001$.

To better understand the underlying biological mechanisms that drive the development from healthy to AD brains, I conducted differential expression (DE) analysis for selected cell types. I determined which cell types to focus on by quantifying the number of DE genes between AD and healthy brain samples for each cell type (Figure 8A). The negative and positive x-axis indicate the number of downregulated and upregulated genes, respectively. I decided to focus on five cell types with the highest number of DE genes, namely, Microglial, Oligodendrocyte, Endothelial Pericyte, Astrocyte, and Excitatory 3 cells (Figure 8A). I generated Heatmaps of the expression of top DE genes for each cell type to investigate which genes might be of interest (Figure 8B-F). Notably, *NEAT1* and *FKBP5* are consistently upregulated in AD samples, whereas ROBO2 seems to be consistently downregulated.

*NEAT1* is a long-non-coding RNA that promotes inflammation through the activation of inflammasomes. The knockdown of *NEAT1* has been reported to exhibit protective effects in AD mouse models [20].

A publication by Zannas et al. (2019) highlights the upregulation of the *FKBP5* gene through epigenetic mechanisms associated with aging or stress [17]. Importantly, higher levels of *FKBP5* expression promotes inflammation as it activates NF-kB, a critical immune response mediator that regulates the expression of many downstream pro-inflammatory genes [8]. The role of the NF-kB inflammatory pathway in neurodegeneration and pathogenesis of AD have been previously reported [23]. While the *FKBP5* protein was mainly characterized in stress physiology, with discussions about targeting *FKBP5* for treating stress-related disorders [19], it has not been extended to discussions about AD. The upregulation of the *FKBP5* gene in all five cell types in AD samples is consistent with hallmarks of chronic inflammation in AD brains and suggest the potential of leveraging drug-repurposing from psychiatric development of FKBP5-related drugs for treatment of AD.

*ROBO2* is a crucial protein for axon guidance and cell migration. In their publication, Kaneko et al. (2010) describe the role of ROBO receptors in helping young neurons migrate rapidly through long directional ranges in the adult brain [21]. In a follow-up study, the authors observe impaired neuroblast migration in ROBO2-knockdown models, demonstrating the importance of the protein in cell localization and possibly even in repair and regeneration of the brain [22]. There is currently limited work in associating *ROBO2* expression with AD, and the importance of *ROBO2* for neuroblast migration may provide insights into the impaired ability of neuronal repair in AD brains.

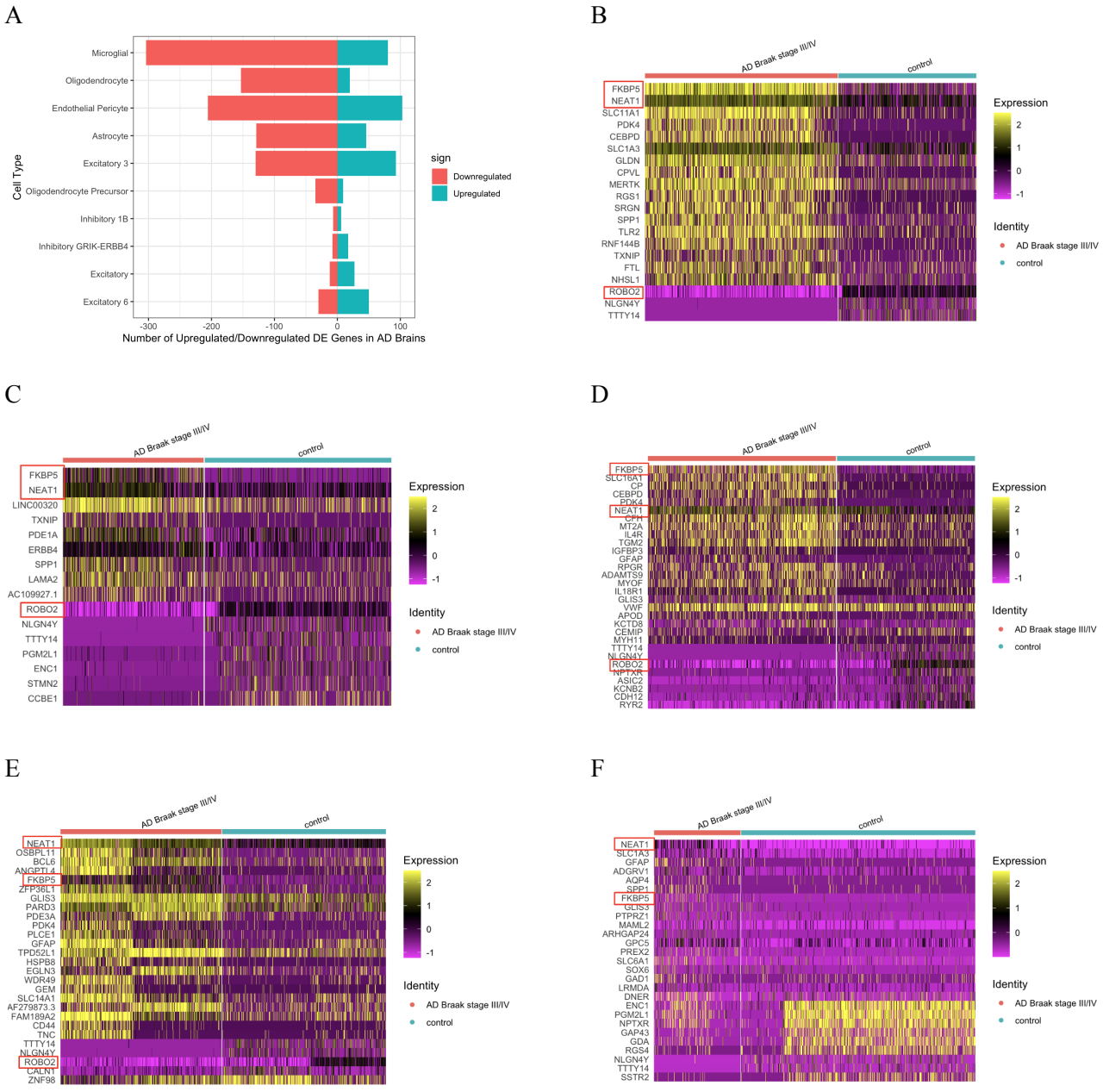


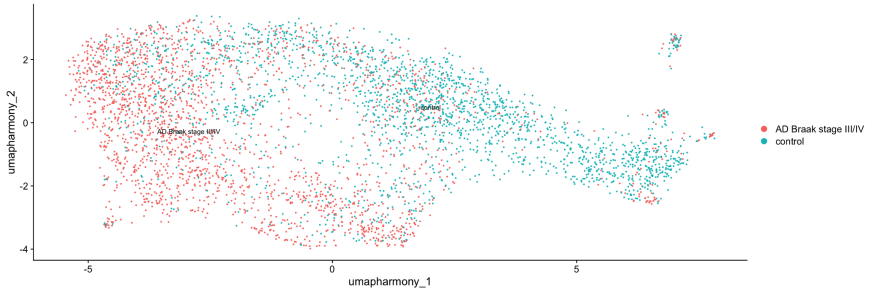

**Figure 8. Differential abundance analysis for annotated cell types between AD and control brains.** (A) Number of down-regulated and up-regulated genes between AD and control samples for each cell type. Heatmap of top DE gene expression for Microglia (B), Oligodendrocyte (C), Endothelial Pericyte (D), Astrocyte (E) and Excitatory 3 (F) cells.

Another observation is that, unlike the other non-neuronal cells, the heatmap for Excitatory 3 neurons seem to show downregulation of genes in AD brains compared to control brains (Figure 8F). This observation may be explained by the understanding that many neurons stop functioning in AD and eventually die.

I also observe an interesting phenomenon in the Heatmap for Astrocytes (Figure 8E), where astrocytes in AD brain samples seem to be divided into two subpopulations. To further investigate this, I subsetted the Seurat object to only contain Astrocytes and redid the normalization, batch-correction, and dimensionality reduction steps. Then, I constructed a UMAP colored by brain type as well as the expression of select genes that are highly expressed in only half of the astrocytes based on the Heatmap in Figure 8F (Figure 9). Based on the UMAP colored by brain type, there is a visible separation of Astrocytes from AD and control samples, with some mixing in the middle (Figure 9A). The expression of selected genes also seem to change gradually from control to AD astrocytes, suggesting some continuum as astrocytes transition from healthy to AD (Figure 9B). Furthermore, there seems to be two types of AD astrocytes, where one type highly expresses EGLN3 and PDE3A genes (Figure 9B right column), and another type highly expresses ANGPTL4, GFAP, and TNC genes (Figure 9B left column).

*EGLN3* has been reported to be a negative regulator of the NF-kB signaling pathway [24]. *PDE* is a candidate therapeutic target for AD due to its critical role in cAMP regulation, where aberrant cAMP signaling has been associated with AD [25]. On the other hand, *ANGPTL4* has been shown to be upregulated in AD astrocytes, with implications for its role in depositing beta-amyloid aggregates in the vasculature [26]. *GFAP* is a known biomarker for AD pathology that is upregulated and released by astrocytes as a consequence of astrogliosis due to the buildup of beta-amyloid plaques [27]. Finally, *TNC* is a protein that is upregulated in response to inflammation and has been thought to contribute to chronic inflammation in AD [18]. Altogether, all of the aforementioned genes have well-studied associations with AD, and the two populations of astrocytes that highly express different sets of genes may suggest (1) different trajectories in the transition from healthy to AD astrocytes, or (2) one population may be the cause and the other may be the effect of AD pathology.

A




B

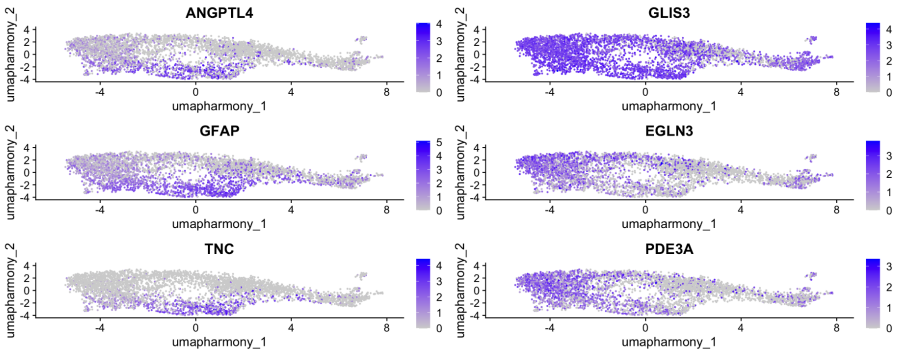



**Figure 9. UMAP of Astrocytes from control and AD samples.** (A) UMAP of astrocytes colored by sample type. (B) UMAP of astrocytes colored by gene expression of the labeled gene.

To further elucidate the subtypes of astrocytes in AD, I conducted cluster analysis using the Louvain algorithm with 0.1 and 0.15 clustering resolutions (Figure 10). The 0.1 clustering resolution (Figure 10A) identified three distinct clusters that align with healthy versus AD labels in Figure 9A. The 0.15 clustering resolution (Figure 10B) identified five clusters, where two of the clusters align with subpopulations of AD astrocytes that highly express

different genes as described above. This clustering resolution also separates healthy astrocytes into distinct clusters, where cluster 1 may be astrocytes that are gradually transitioning from a healthy to AD state (Figure 10B). These hypotheses will be studied later in our Trajectory Analysis module by Fiona.

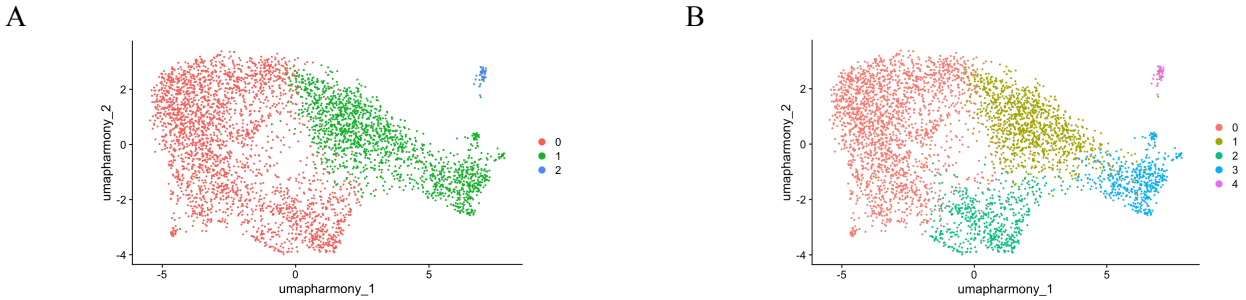



**Figure 10. UMAP of Astrocytes from AD and control samples colored by the assigned clusters.** Clustering of Astrocytes using the Louvain algorithm with 0.1 (A) and 0.15 (B) clustering resolutions.

In order to understand the pathway that may be driving differences between the subclusters of astrocytes in the 0.15 resolution, I identified marker genes and conducted gene set enrichment analysis (GSEA) for each cluster. The gene sets used were the Hallmark and the Gene Ontology (GO) gene sets provided by MSigDB. Cluster 1 had a statistically significant enrichment of the metabolic-related pathways from the GO gene set including AEROBIC_RESPIRATION (NES = 2.38, FDR < 0.001) and OXIDATIVE_PHOSPHORYLATION (NES = 2.38, FDR < 0.001). This highlights an important finding. Neuronal-Astrocyte metabolic interactions in healthy brains have been well-studied, where astrocytes provide neurons with vital metabolic support due to high energy requirements for neuronal activation. However, while neurons rely on aerobic oxidative respiration, astrocytes typically resort to anaerobic glycolysis [31]. The enrichment of aerobic respiration in cluster 1 astrocytes, therefore, exhibit signs of unusual astrocyte metabolism. A study by Monterey et al. (2021) reports increased mitochondrial oxidative metabolism in astrocytes as a consequence of the NF-kB pathway, which could result in further inflammation due to oxidative stress [14]. Altogether, the enrichment of aerobic respiration in cluster 1, which is situated between AD and control astrocytes, suggests that these cells may be in a transitory state.

In addition to cluster 1, cluster 3 had significant negative enrichment of the CYTOSKELETON_ORGANIZATION pathway from the GO gene set (NES = -2.98, FDR < 0.05). A study by Schiweck et al. (2018) reports the morphological complexity of mature astrocytes, where they can reorganize their cytoskeleton to undergo morphological changes in response to inflammation and neurodegeneration [30]. The fact that cluster 3 (astrocytes from control samples) displays negative enrichment of cytoskeleton organization indicates that results from this GSEA analysis is consistent with known knowledge about astrocytes.

I also conducted GSEA for four remaining cell types. Most notably, AD Microglial and Endothelial Pericyte cells both exhibit significant enrichment of INFLAMMATORY_RESPONSE (NES = 2.38 and 2.39, FDR < 0.001) and TNFA_SIGNALING_VIA_NFKB (NES = 2.29 and 2.25, FDR < 0.001) pathways in the Hallmark gene set. The enrichment of these pathways, especially the involvement of TNF-alpha and NF-kB, are consistent with AD pathophysiology [23,29]. Similar pathways involved in the immune system are also significantly enriched in AD Microglial and Endothelial Pericytes, such as REGULATION_OF_IMMUNE_SYSTEM_PROCESS (NES = 3.43 and 2.38, FDR < 0.001) and RESPONSE_TO_CYTOKINE (NES = 2.53 and 2.69, FDR < 0.001). While the role of inflammation and immune response in AD is not novel, it has typically been associated with microglial and astrocytes. It is interesting that these pathways are shown to be upregulated in AD microglial and endothelial pericytes instead. While there is the possibility of misannotation, the cell types have been assigned via two steps of verification as previously described (Figure 6). The results from our study may therefore suggest endothelial pericytes as a new cell type of interest involved in AD pathogenesis and may serve as new therapeutic targets.

Altogether, the analyses from this module highlight a few key findings: (1) there is a general decrease in neuronal cells and an increase in glial cells in AD samples, (2) several genes are consistently upregulated or downregulated across multiple cell types in AD, indicating the potential of new therapeutic targets, (3) two subpopulations of astrocytes in AD samples exhibit different expression patterns and enriched pathways that may help elucidate AD pathogenesis, and (4) several inflammatory pathways are enriched in known (microglial) and unexpected (endothelial pericyte) cells, which may open new avenues for studying AD pathogenesis. In the next module, Fiona dives deeper into learning about the transition from healthy to AD cell states via trajectory analysis. This will help us better understand the different clusters of astrocytes, as well as the trajectories of other cell types.

**Module 3: Trajectory Analysis for Key Cell Types**

Fiona McBride | fiona_mcbride@hms.harvard.edu | 11571778

In this module, I investigate cell trajectories to better understand how certain cell types transition from healthy to AD states. I then find genes correlated with the trajectory; these genes may be key drivers of cells becoming diseased. Jesslyn identified five interesting cell types in the previous module that appeared to have a significant number of up- and down-regulated genes between AD and control patients. While I conducted investigatory cell trajectory analysis for all five cell types as well as for the entire cellular profile of the patients, the only cell types that showed interesting trajectories were astrocytes, microglia, and oligodendrocytes. These are the cell types I will focus on for the remainder of this module.

The data used throughout this section is the Seurat scRNA-seq data object with annotated cell clusters, produced in module 2. I subsetted the whole Seurat object to get three smaller Seurat objects, each containing cells from one of the clusters of interest. To be able to identify transitions in cell states within each cell type, I re-processed the smaller Seurat objects using the built-in Seurat functions with the same parameters as described in module 1 in order to keep processing consistent throughout the entire pipeline. I then created a UMAP plot for each cell type colored by disease status to be able to visualize how the diseases separate along the first two dimensions of the UMAP. Separation along these axes indicates that there is variation between disease states that can be picked up in the RNA expression profiles, making them good candidates for trajectory analysis.

I used the R package Monocle3 and the corresponding built-in functions to infer the single-cell trajectories [32-34]. The first step was to convert the Seurat object into a cell data set (cds) object, which is the data format used for downstream processing in Monocle3. Then, I performed feature selection based on the variable genes identified in each subsetted Seurat object, UMAP dimensionality reduction, clustering, and projected the data into the lower dimensional space.

The next step was to order the cells based on their pseudotime, which is an artificial measure of time that represents different stages of cellular differentiation [35]. This was achieved using the order_cells function, which requires a manual selection of the root node. Choosing a root node is a subjective process—I chose nodes that were within the control cluster and further away from the AD cluster because I wanted to explore the genes that were correlated with the healthy to disease trajectory. Once the root node was selected, pseudotime was calculated using the built-in Monocle3 function, and I found the top five genes that were positively and negatively correlated with pseudotime.

Astrocytes had very distinct differentiation between disease states in the Seurat UMAP plots (Figure 11A, top), indicating that there are a lot of disease-specific gene expression patterns. This was replicated in the trajectory plot (Figure 11A, middle). I selected a root node roughly at the edge of the control sample cells. Figure 11C (bottom) shows the pseudotime gradient coloring the cells of the trajectory-based UMAP.

Microglia also show distinct separation on the Seurat UMAP when colored by disease state (Figure 11B, top), which is preserved in the trajectory UMAP (Figure 11B, middle). Interestingly, there is a small cluster of cells that projects much further from the others in both the Seurat UMAP and trajectory UMAP. Further investigation could determine

which genes are contributing to this separation, but that was not considered for this current project. I selected a root node towards the top left corner of the plot, again in the middle of the control cluster (Figure 11B, bottom). Again, I observe a pseudotime gradient going from healthy to disease.

Oligodendrocytes appeared to have the least amount of separation in the Seurat UMAP colored by disease state (Figure 11C, top). However, the trajectory UMAP does show clear separation between control and AD cells (Figure 11C, middle). I selected a root node toward the edge of the control cluster and observed a clear pseudotime gradient from healthy to disease (Figure 11C, bottom).

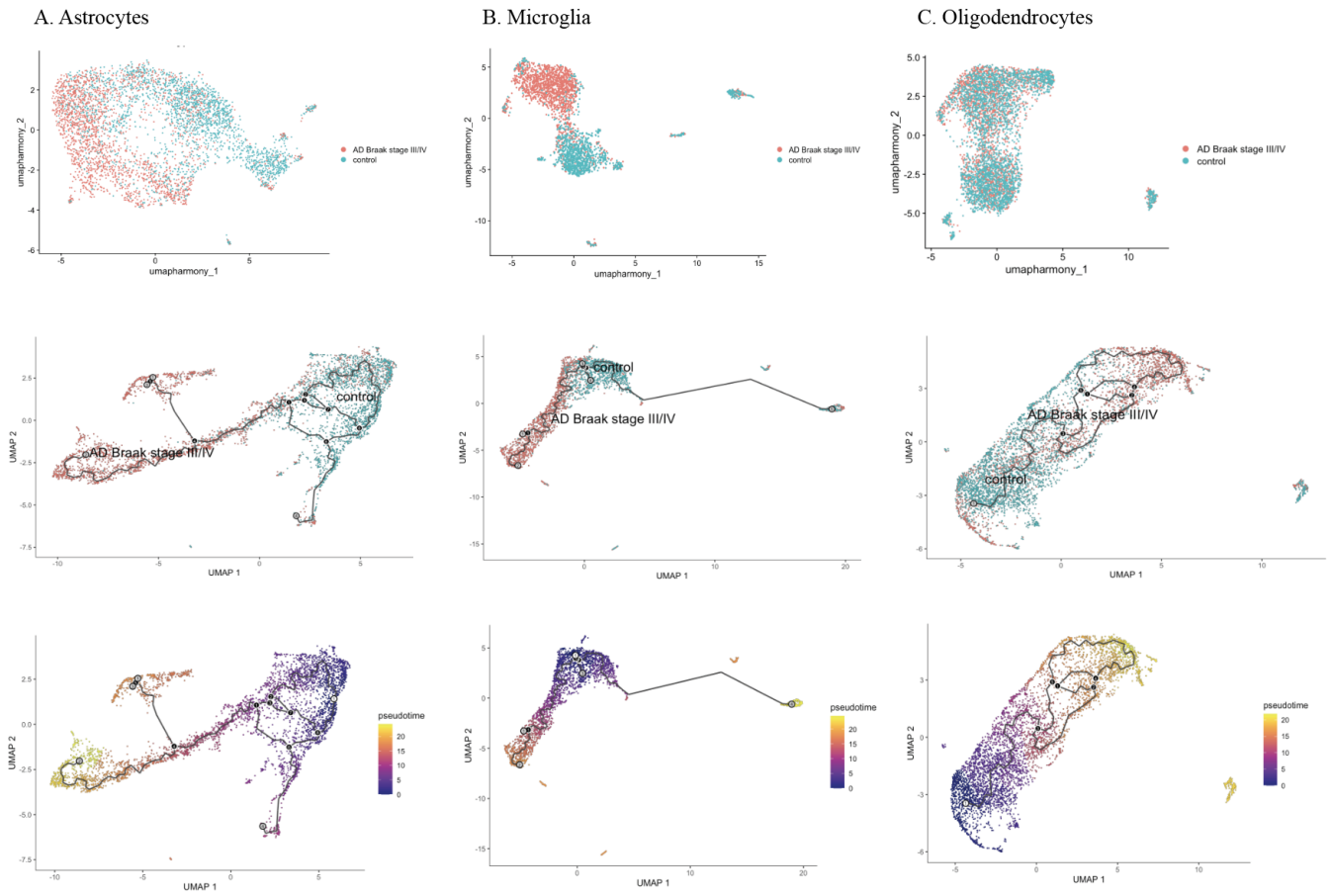



**Figure 11. Trajectory analysis for astrocytes, microglia, and oligodendrocytes.** (A) UMAP of astrocytes in Seurat object (top), UMAP of astrocyte trajectory from Monocle3 (middle), UMAP of astrocyte trajectory, colored by pseudotime (bottom) (B) UMAP of microglia in Seurat object (top), UMAP of microglia trajectory from Monocle3 (middle), UMAP of microglia trajectory, colored by pseudotime (bottom) (C) UMAP of oligodendrocytes in Seurat object (top), UMAP of oligodendrocytes trajectory from Monocle3 (middle), UMAP of oligodendrocytes trajectory, colored by pseudotime (bottom)

Next, I found the top 5 genes in each cell type that were positively correlated with pseudotime and the top 5 genes that were negatively correlated with pseudotime. These genes (Table 1) are likely to be important drivers of cells transitioning from healthy to disease states.

Two genes, *NEAT1* and *XIST*, were found to be positively associated with pseudotime in two of the three cell types (Table 1, green cells). As described in module 2, *NEAT1* promotes inflammation, and knockdown of this RNA has been associated with protective effects in AD mouse models [20]. The positive correlation between *XIST* and pseudotime is somewhat surprising; *XIST* produces a long non-coding RNA that inactivates one copy of the X chromosome in females [36]. However, all of our samples in this study are male, making *XIST* upregulation

unusual. The connection between *XIST* and AD has not been well characterized, but a recent preprint article by Abdulai-Saiku et al. (2022) shows evidence that the expression of the maternal X chromosome is linked to increased cognitive impairment in mice [37]. Since males only inherit maternal X chromosomes, inactivating them would potentially reduce the rate of cognitive impairment, which is contrary to our results. This would be an interesting phenomenon to study further.

Three genes, *CSMD1*, *SYT1*, and *RBFOX1*, were found to be negatively correlated with pseudotime in two of the three cell types (Table 1, light red cells), and one gene, *KCNIP4*, was negatively correlated with pseudotime in all cell types (Table 1, dark red cells). Genetic markers in the gene *CSMD1* have been associated with AD, though the effects depend on the variant expressed [38]. The gene product of *SYT1* is important for regulating neuronal function at synapses and associates with presenilin 1 to perform synaptic upkeep by regulating synaptic vesicle cycling. Downregulation of *SYT1* contributes to the dysregulation of synapses and has been shown to be characteristic of AD patients [39]. The *RBFOX1* gene produces a neuronal RNA-binding protein that tends to be localized around plaques; reduced *RBFOX1* expression is correlated with worse cognition and increased amyloid-β plaque [40]. *KCNIP4* was downregulated in all cell types, but the effect of this gene is determined by the alternative splicing variant expressed, which we don't know in our data. A decrease of *KCNIP4* Var I and an increase of *KCNIP4* Var 4 has been found to increase inflammation and progress neurodegeneration [41, 42]. If the decreased expression of *KCNIP4* found in this study corresponds to a decrease in Var I, our findings are consistent with previous literature.

| **Astrocyte** | | **Microglia** | | **Oligodendrocyte** | |
|---|---|---|---|---|---|
| Positive | Negative | Positive | Negative | Positive | Negative |
| NEAT1 | KCNIP4 | XIST | KCNIP4 | XIST | RBFOX1 |
| OSBPL11 | CSMD1 | DPYD | SYT1 | TMEM144 | KCNIP4 |
| ERBIN | SYT1 | PIP4K2A | RBFOX1 | KCNMB4 | NRXN1 |
| RANBP3L | MT-CO1 | FA2H | CSMD1 | LINC00320 | FAM155A |
| PDE3A | LRRTM4 | PDE8A | ROBO2 | NEAT1 | MT-ATP6 |

**Table 1. Genes correlated to pseudotime for astrocytes, microglia, and oligodendrocytes.** Genes are sorted in descending order (genes with the highest correlation are in the first row). Genes that are positively correlated in two cell types are highlighted in light green. Genes that are negatively correlated in two cell types are highlighted in light red. Genes that are negatively correlated in all three cell types are highlighted in dark red.

## Conclusion and Future Directions

In this project, we used single-cell RNA sequencing data of two control patients and two AD patients to investigate the cell-level mechanisms that characterize Alzheimer's Disease. Differential gene expression showed that glial cells are among the top cell types with the most DE genes, which is consistent with the existing understanding of their role in AD pathogenesis. Furthermore, we identified several known and novel genes that are consistently up or down-regulated across brain cell types that may warrant further analysis for their therapeutic potential. GSEA identified known pathways that are enriched but in an unexpected cell type (endothelial pericytes), which may call for more attention in future studies. Trajectory analysis showed that there are clear patterns of transition from control to AD for astrocytes, microglia, and oligodendrocytes. Some of the genes identified in the differential gene expression analysis were also identified as being key genes in the trajectory analysis; these genes may serve as good targets for therapeutic intervention.

Some takeaways from completing this project include the following:

1. The importance of documenting code and keeping all analyses.
    a. It was occasionally difficult to find the analyses done by the original authors, which limited our ability to compare our results if there was not a direct reference to them in the original text.
    b. All of the code for our project was very dependent on the previous step, so there needed to be clear references to what was done in case anything went wrong downstream.
2. Many of the genes that were identified through differential expression analysis or trajectory analysis were found to have variable effects depending on which gene variant was expressed, which was not information included in our dataset. To be able to fully explore the biological connections between genes and disease, we would've needed multiple data modalities.

If we were to continue this project, there are multiple avenues of further research to consider. Since the AD samples provided were only from stages III/IV, it may be beneficial to also obtain samples from stages I/II as well. This may help elucidate a smoother transition from healthy to AD cell states, as well as the genes and pathways involved. There also appear to be two separate clusters of astrocytes with slightly different trajectories. It would be interesting to look at the genes that are correlated with each of those clusters to see if there are different ways that astrocytes become diseased in AD. There are lots of biological areas of study, as many of the genes we identified to be differentially expressed or correlated with trajectory have yet to be fully studied in Alzheimer's models. One limitation of this study is the very small sample size; it is difficult to assess whether some of the differences we see are truly different biological bases of AD, or if the variation is arising because our small subset of patients had a variation that we did not catch with batch correction. Redoing these analyses with a larger sample size may confirm some of the trends we found or eliminate trends from further study.

**References**


[1] What is Alzheimer's Disease? | CDC. Published January 7, 2023. Accessed May 8, 2023. https://www.cdc.gov/aging/aginginfo/alzheimers.htm

[2] Calabrò M, Rinaldi C, Santoro G, Crisafulli C. The biological pathways of Alzheimer disease: a review. *AIMS Neurosci*. 2020;8(1):86-132. doi:10.3934/Neuroscience.2021005

[3] Soreq L, Bird H, Mohamed W, Hardy J. Single-cell RNA sequencing analysis of human Alzheimer's disease brain samples reveals neuronal and glial specific cells differential expression. *PLOS ONE*. 2023;18(2):e0277630. doi:10.1371/journal.pone.0277630

[4] GEO Accession viewer. Accessed May 7, 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175814

[5] Tools for Single Cell Genomics. Accessed May 7, 2023. https://satijalab.org/seurat/

[6] Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289-1296. doi:10.1038/s41592-019-0619-0

[7] FLT1 fms related receptor tyrosine kinase 1 [Homo sapiens (human)] - Gene - NCBI. Accessed May 7, 2023. https://www.ncbi.nlm.nih.gov/gene/2321

[8] Mahoney ER, Dumitrescu L, Moore AM, et al. Brain expression of the vascular endothelial growth factor gene family in cognitive aging and alzheimer's disease. *Mol Psychiatry*. 2021;26(3):888-896. doi:10.1038/s41380-019-0458-5

[9] Seripa D, Matera MG, Franceschi M, et al. The RELN locus in Alzheimer's disease. *J Alzheimers Dis*. 2008;14(3):335-344. doi:10.3233/jad-2008-14308

[10] Herring A, Donath A, Steiner KM, et al. Reelin depletion is an early phenomenon of Alzheimer's pathology. *J Alzheimers Dis*. 2012;30(4):963-979. doi:10.3233/JAD-2012-112069

[11] Lake, B., Chen, S., Sos, B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol 36, 70–80 (2018). https://doi.org/10.1038/nbt.4038

[12] Hao and Hao et al. Integrated analysis of multimodal single-cell data. Cell, (2021). https://doi.org/10.1016/j.cell.2021.04.048 [Seurat V4]

[13] National Institute on Aging. (n.d.). What happens to the brain in Alzheimer's disease? Retrieved May 5, 2023, from https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease

[14] Monterey, M. D., Wei, H., Wu, X., & Wu, J. Q. (2021). The Many Faces of Astrocytes in Alzheimer's Disease. Frontiers in neurology, 12, 619626. https://doi.org/10.3389/fneur.2021.619626

[15] Kim, Y. S., Jung, H. M., & Yoon, B. E. (2018). Exploring glia to better understand Alzheimer's disease. Animal cells and systems, 22(4), 213–218. https://doi.org/10.1080/19768354.2018.1508498

[16] Uddin, M. S., & Lim, L. W. (2022). Glial cells in Alzheimer's disease: From neuropathological changes to therapeutic implications. Ageing research reviews, 78, 101622. https://doi.org/10.1016/j.arr.2022.101622

[17] Zannas, A. S., Jia, M., Hafner, K., Baumert, J., Wiechmann, T., Pape, J. C., Arloth, J., Ködel, M., Martinelli, S., Roitman, M., Röh, S., Haehle, A., Emeny, R. T., Iurato, S., Carrillo-Roa, T., Lahti, J., Räikkönen, K., Eriksson, J. G., Drake, A. J., Waldenberger, M., … Binder, E. B. (2019). Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk. Proceedings of the National Academy of Sciences of the United States of America, 116(23), 11370–11379. https://doi.org/10.1073/pnas.1816847116

[18] Liu, T., Zhang, L., Joo, D., & Sun, S. C. (2017). NF-κB signaling in inflammation. Signal transduction and targeted therapy, 2, 17023–. https://doi.org/10.1038/sigtrans.2017.23

[19] Codagnone, M. G., Kara, N., Ratsika, A., Levone, B. R., van de Wouw, M., Tan, L. A., Cunningham, J. I., Sanchez, C., Cryan, J. F., & O'Leary, O. F. (2022). Inhibition of FKBP51 induces stress resilience and alters hippocampal neurogenesis. Molecular psychiatry, 27(12), 4928–4938. https://doi.org/10.1038/s41380-022-01755-9

[20] Zhao, M. Y., Wang, G. Q., Wang, N. N., Yu, Q. Y., Liu, R. L., & Shi, W. Q. (2019). The long-non-coding RNA NEAT1 is a novel target for Alzheimer's disease progression via miR-124/BACE1 axis. Neurological research, 41(6), 489–497. https://doi.org/10.1080/01616412.2018.1548747

[21] Kaneko, N., Marín, O., Koike, M., Hirota, Y., Uchiyama, Y., Wu, J. Y., Lu, Q., Tessier-Lavigne, M., Alvarez-Buylla, A., Okano, H., Rubenstein, J. L., & Sawamoto, K. (2010). New neurons clear the path of astrocytic processes for their rapid migration in the adult brain. Neuron, 67(2), 213–223. https://doi.org/10.1016/j.neuron.2010.06.018

[22] Kaneko, N., Herranz-Pérez, V., Otsuka, T., Sano, H., Ohno, N., Omata, T., Nguyen, H. B., Thai, T. Q., Nambu, A., Kawaguchi, Y., García-Verdugo, J. M., & Sawamoto, K. (2018). New neurons use Slit-Robo signaling to migrate through the glial meshwork and approach a lesion for functional regeneration. Science advances, 4(12), eaav0618. https://doi.org/10.1126/sciadv.aav0618

[23] Sun, E., Motolani, A., Campos, L., & Lu, T. (2022). The Pivotal Role of NF-kB in the Pathogenesis and Therapeutics of Alzheimer's Disease. International journal of molecular sciences, 23(16), 8972. https://doi.org/10.3390/ijms23168972

[24] Fu, J., & Taubman, M. B. (2013). EGLN3 inhibition of NF-κB is mediated by prolyl hydroxylase-independent inhibition of IκB kinase γ ubiquitination. Molecular and cellular biology, 33(15), 3050–3061. https://doi.org/10.1128/MCB.00273-13

[25] Sheng, J., Zhang, S., Wu, L., Kumar, G., Liao, Y., Gk, P., & Fan, H. (2022). Inhibition of phosphodiesterase: A novel therapeutic target for the treatment of mild cognitive impairment and Alzheimer's disease. Frontiers in aging neuroscience, 14, 1019187. https://doi.org/10.3389/fnagi.2022.1019187

[26] Chakraborty, A., Kamermans, A., van Het Hof, B., Castricum, K., Aanhane, E., van Horssen, J., Thijssen, V. L., Scheltens, P., Teunissen, C. E., Fontijn, R. D., van der Flier, W. M., & de Vries, H. E. (2018). Angiopoietin like-4 as a novel vascular mediator in capillary cerebral amyloid angiopathy. Brain : a journal of neurology, 141(12), 3377–3388. https://doi.org/10.1093/brain/awy274

[27] Cicognola, C., Janelidze, S., Hertze, J. et al. (2021). Plasma glial fibrillary acidic protein detects Alzheimer pathology and predicts future conversion to Alzheimer dementia in patients with mild cognitive impairment. Alz Res Therapy 13, 68. https://doi.org/10.1186/s13195-021-00804-9

[28] Xie, K., Liu, Y., Hao, W., Walter, S., Penke, B., Hartmann, T., Schachner, M., & Fassbender, K. (2013). Tenascin-C deficiency ameliorates Alzheimer's disease-related pathology in mice. Neurobiology of aging, 34(10), 2389–2398. https://doi.org/10.1016/j.neurobiolaging.2013.04.013

[29] Chang, R., Yee, K. L., & Sumbria, R. K. (2017). Tumor necrosis factor α Inhibition for Alzheimer's Disease. Journal of central nervous system disease, 9, 1179573517709278. https://doi.org/10.1177/1179573517709278

[30] Schiweck, J., Eickholt, B. J., & Murk, K. (2018). Important Shapeshifter: Mechanisms Allowing Astrocytes to Respond to the Changing Nervous System During Development, Injury and Disease. Frontiers in cellular neuroscience, 12, 261. https://doi.org/10.3389/fncel.2018.00261

[31] Turner, D. A., & Adamson, D. C. (2011). Neuronal-astrocyte metabolic interactions: understanding the transition into abnormal astrocytoma metabolism. Journal of neuropathology and experimental neurology, 70(3), 167–176. https://doi.org/10.1097/NEN.0b013e31820e1152

[32] Trapnell, C., Cacchiarelli, D., Grimsby, J. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32, 381–386 (2014). https://doi.org/10.1038/nbt.2859

[33] Qiu, X., Hill, A., Packer, J. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* 14, 309–315 (2017). https://doi.org/10.1038/nmeth.4150

[34] Qiu, X., Mao, Q., Tang, Y. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 14, 979–982 (2017). https://doi.org/10.1038/nmeth.4402

[35] Campbell, K.R., Yau, C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nat Commun* 9, 2442 (2018). https://doi.org/10.1038/s41467-018-04696-6

[36] Loda, Agnese, and Edith Heard. “Xist RNA in action: Past, present, and future.” *PLoS genetics* vol. 15,9 e1008333. 19 Sep. 2019, https://doi.org/10.1371/journal.pgen.1008333

[37] Samira Abdulai-Saiku, Shweta Gupta, Dan Wang, Arturo J. Moreno, Yu Huang, Deepak Srivastava, Barbara Panning, & Dena B. Dubal. (2022). The maternal X chromosome impairs cognition and accelerates brain aging through epigenetic modulation in female mice. *BioRxiv*, 2022.03.09.483691. https://doi.org/10.1101/2022.03.09.483691

[38] Stepanov, Vadim et al. “Genetic Variants in CSMD1 Gene Are Associated with Cognitive Performance in Normal Elderly Population.” *Genetics research international* vol. 2017 (2017): 6293826. https://doi.org/10.1155/2017/6293826

[39] Keller, Laura J et al. “Presenilin 1 increases association with synaptotagmin 1 during normal aging.” *Neurobiology of aging* vol. 86 (2020): 156-161. https://doi.org/10.1016/j.neurobiolaging.2019.10.006

[40] Raghavan, N. S., Dumitrescu, L., Mormino, E., Mahoney, E. R., Lee, A. J., Gao, Y., Bilgel, M., Goldstein, D., Harrison, T., Engelman, C. D., Saykin, A. J., Whelan, C. D., Liu, J. Z., Jagust, W., Albert, M., Johnson, S. C., Yang, H. S., Johnson, K., Aisen, P., Resnick, S. M., … Alzheimer's Disease Neuroimaging Initiative (2020). Association Between Common Variants in RBFOX1, an RNA-Binding Protein, and Brain Amyloidosis in Early and Preclinical Alzheimer Disease. *JAMA neurology*, *77*(10), 1288–1298. https://doi-org.ezp-prod1.hul.harvard.edu/10.1001/jamaneurol.2020.1760

[41] Massone, S., Vassallo, I., Castelnuovo, M., Fiorino, G., Gatta, E., Robello, M., Borghi, R., Tabaton, M., Russo, C., Dieci, G., Cancedda, R., & Pagano, A. (2011). RNA polymerase III drives alternative splicing of the potassium channel-interacting protein contributing to brain complexity and neurodegeneration. *The Journal of cell biology*, *193*(5), 851–866. https://doi-org.ezp-prod1.hul.harvard.edu/10.1083/jcb.201011053

[42] Rockefeller University Press. (2011, June 20). Noncoding RNA may promote Alzheimer's disease. *ScienceDaily*. Retrieved May 8, 2023 from www.sciencedaily.com/releases/2011/05/110530152336.htm