Kyla Gabriel
5/3/2024
BMI 710

Understanding scRNA Analysis Using Alzheimer's Disease Data

Introduction

To illustrate the practical applications of single-cell RNA (scRNA) data analysis in scientific research, I utilized established analytical techniques on the dataset generated by Leng et al. in their study titled "Molecular characterization of selectively vulnerable neurons in Alzheimer's disease." [1] This paper investigates Alzheimer's Disease (AD) by examining specific brain regions using single-nucleus RNA sequencing.

This study uses brain tissue from human subjects that were initially frozen before analysis since the tissue was collected postmortem. The investigators performed single-cell isolation by homogenizing the tissue in a lysis buffer, followed by addition of IGEPAL-630 and further homogenization. The solution was filtered and mixed with Optiprep; then it was layered onto a gradient and centrifuged to isolate the nuclei. The experimental approach the investigators used was droplet-based single-nucleus RNA sequencing where they utilized 10x Genomics (10x Chromium v2) to analyze the gene expression in the brain tissue at the cellular level. For future analysis, it is important to note that the sample size is quite small with a total of only 31 participants. Furthermore, it is important to note that some quality control measures were already performed including removing empty droplets that contained no nuclei and nuclei with less than 200 UMIs.

Methodology

Quality Control

To evaluate if quality control methods were needed for this data, the data was first visualized (Figure 1). This figure shows the variations in the amount of expressed genes, RNA count, and the percent of mitochondria UMIs in detected cells. From the plot, it can be seen that the amount of expressed genes (nFeatureRNA) and the percent of mitochondria UMIs (percent_mito) vary drastically. Because cells may die due to the stress caused by the experiment, low quality scRNA may be found as shown in the plot. To fix the potential quality issue, low-quality droplets that don't have as much RNA as expected for that of a normal-functioning cell and droplets that don't express the genes expected of a normal- functioning cell were excised. Specifically, mitochondrial UMIs over 20% were removed as well as genes expressed that were below 500. The percent of mitochondria UMIs versus the detected expressed genes were plotted to confirm that the procedure worked effectively before moving forward (Figure 2).

Data quality issues may also occur from confounding variables that are not necessarily rooted in the experiment itself. To account for batch effects that may be present in the data, the batch correction method, Harmony, was utilized. Because batch effects are most likely to stem from patients and the RNA samples, Harmony was used on these variables in order to group the cells by biology. The before and after effects of using Harmony were plotted to correct for these batch effects (Figure 3)..

Data Processing

To process the data, many techniques were deployed including normalization, feature selection, scaling, and principal component analysis (PCA). Normalization was used in this analysis to make the cells more comparable to each other and feature selection was done to find genes that differed the most across all of the cells. Scaling allowed the genes that were more comparable to each other to be seen and PCA was

done to reduce the dimensionality. UMAP was also used to visually see the dimensionality reduction through a low-dimensional graph . Specifically, this was done using the following functions in R: NormalizeData, FindVariableFeatures, ScaleData, RunPCA, and RunUMAP. To find the most appropriate amount of PCs, I used an elbow plot to find the inflection point and use it as a cutoff point to eliminate data that may be noise (Figure 4). I determined the most appropriate number of PCs to be 15 since this was the point in the elbow plot where the points plateaued.

Clustering and Cell-type Annotation

The major cell types I expected to find in this dataset include these seven neuron or neuron-related cells: excitatory neurons, inhibitory neurons, oligodendrocytes, astrocytes, oligodendrocyte precursor cells, microglia, endothelial cells. These cell types were expected because they are specifically tied to AD as described in the paper that this dataset was sourced from. To find the gene markers for each cluster, I used the differential expression approach of finding the nearest neighbors for each cell in the dataset based on their gene expression profiles then clustering by those profiles (Figure 5). The clusters were annotated using the gene makers to find potentially associated cell types. By cross-referencing which genes in the heatmap had the highest expression in each cluster, I found specific cell types that are associated with those genes using the figures found in the Li et al. and Ciaramella et al. papers and the single-cell analytical tool provided by The Broad Institute (Figure 6).[2,3,4] The cell types in the clusters are most likely dendritic cells, macrophages, and T-cells (Table 1) and perform an immune, protective response for neuronal cells. This fits well especially because this dataset involves cells from AD patients and scientific literature proves that macrophages arise from microglia that are compromised due to the dysfunction caused by AD.[5]

Further Analyses

With this cleaned data, more information about AD can be extrapolated and investigated. One biological question that is important is what differences in T-cell abundance exist across Braak stages. I performed methods, including group_by, summarise, and wilcox.test, to determine differential abundance because it will provide specific information about the cell abundance across a certain demographic. Also, I investigated the T-cell abundance specifically because it was the most abundant cell type in the dataset. After comparing the T-cell abundance across the Braak stages, I found that Stage 0 and Stage 2 had the greatest visual difference, so a Wilcoxon rank sum exact test was performed to evaluate the statistical difference (Figure 7). From this test, it was found to have a *p*-value of 0.057 which we can conclude that there is no significant difference between the median proportion of T-cells between Stage 0 and Stage 2.

Furthermore, I wanted to investigate the scientific question of what differentially expressed genes are present in patients with Stage 0 AD vs Stage 6 AD. I am using differential expression methods, such as DESeq2 in this experiment, because it will allow me to see which differentially expressed genes are present between different conditions. After using DESeq2 to understand which genes are differentially expressed, I found that there are genes that are actually different between these two groups. After performing a Wilcoxon rank sum test, I found that these two groups do not have median proportions that are statistically different with the *p*-value being 0.1. From this, we can conclude that there are not differentially expressed genes in AD patients with Stage 0 versus AD patients with Stage 6.

Citations

1. Leng, Kun, et al. "Molecular characterization of selectively vulnerable neurons in Alzheimer's disease." Nature neuroscience 24.2 (2021): 276-287.
2. Li, Songlin, et al. "Activated bone marrow-derived macrophages eradicate Alzheimer's-related Aβ42 oligomers and protect synapses." Frontiers in immunology 11 (2020): 49.
3. Ciaramella, Antonio, et al. "Myeloid dendritic cells are decreased in peripheral blood of Alzheimer's disease patients in association with disease progression and severity of depressive symptoms." Journal of Neuroinflammation 13 (2016): 1-11.
4. *Single Cell Portal*, singlecell.broadinstitute.org/single_cell. Accessed 3 May 2024.
5. Katsumoto A, Takeuchi H, Takahashi K, Tanaka F. Microglia in Alzheimer's Disease: Risk Factors and Inflammation. Front Neurol. 2018 Nov 15;9:978. doi: 10.3389/fneur.2018.00978. PMID: 30498474; PMCID: PMC6249341.

Appendix

Figures

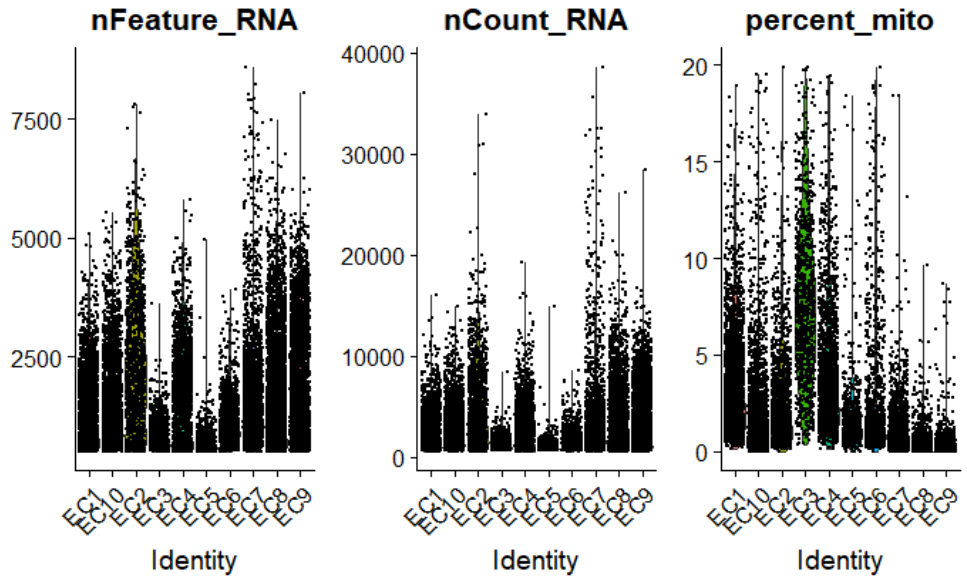

Figure 1: Violin plot displaying variations in amount of expressed genes, amount of RNA, and percent of mitochondria in detected cells

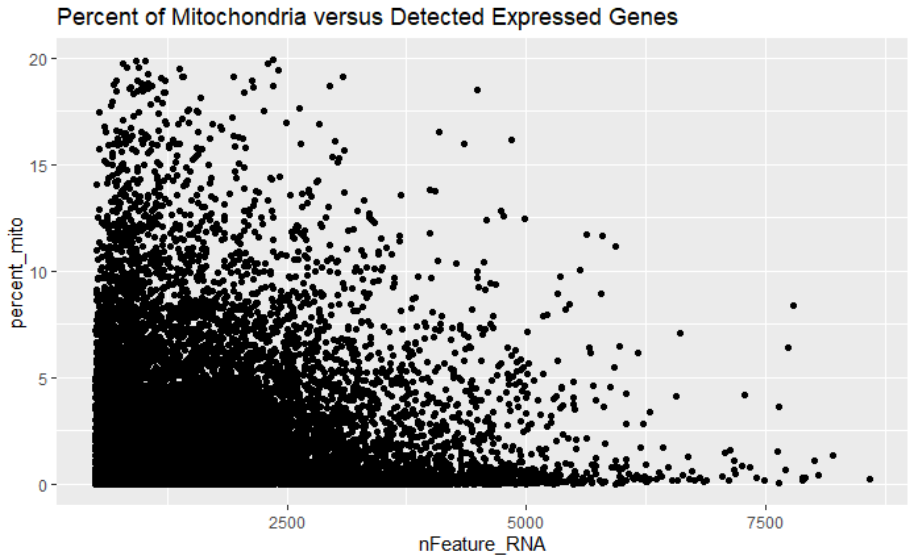

Figure 2: Percent of mitochondria for the detected expressed genes after applying quality control

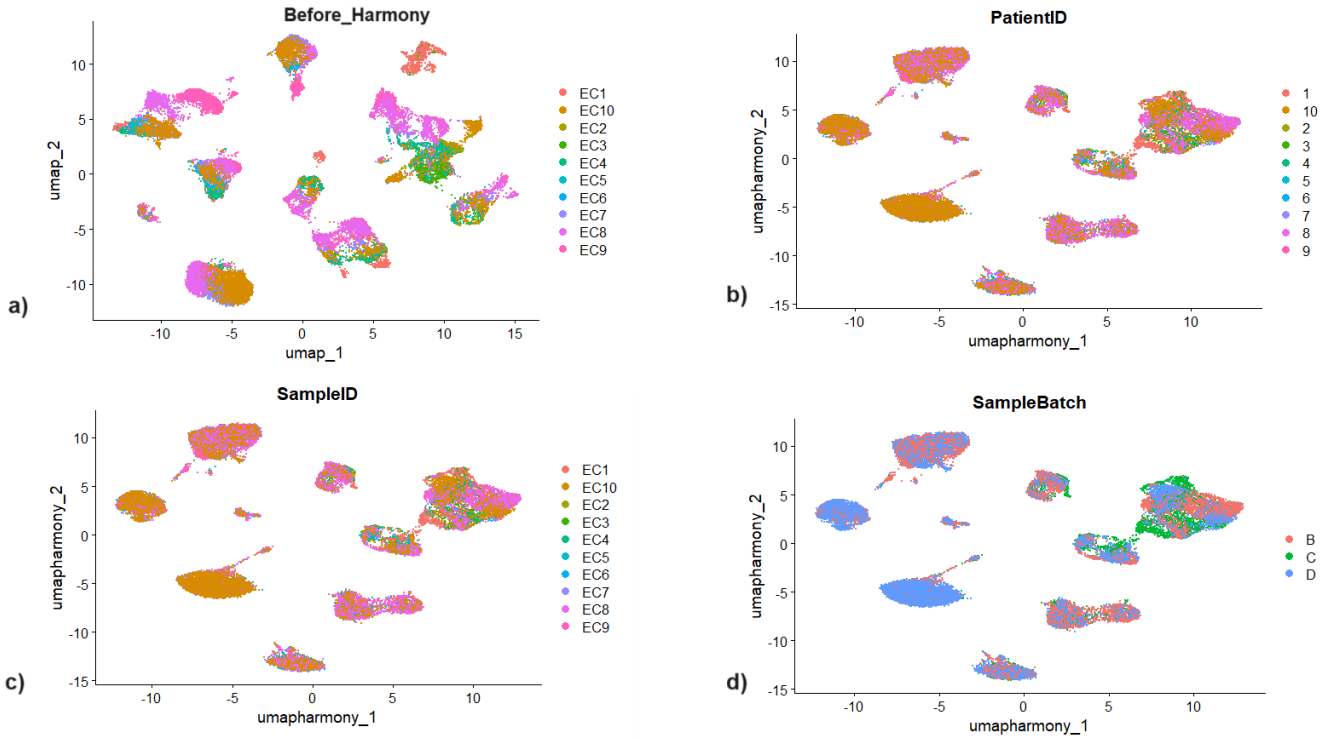

Figure 3: Harmony batch correction; (a) before applying Harmony, (b) correction for patients, (c) correction for sample identifiers, (d) correction for sample batch

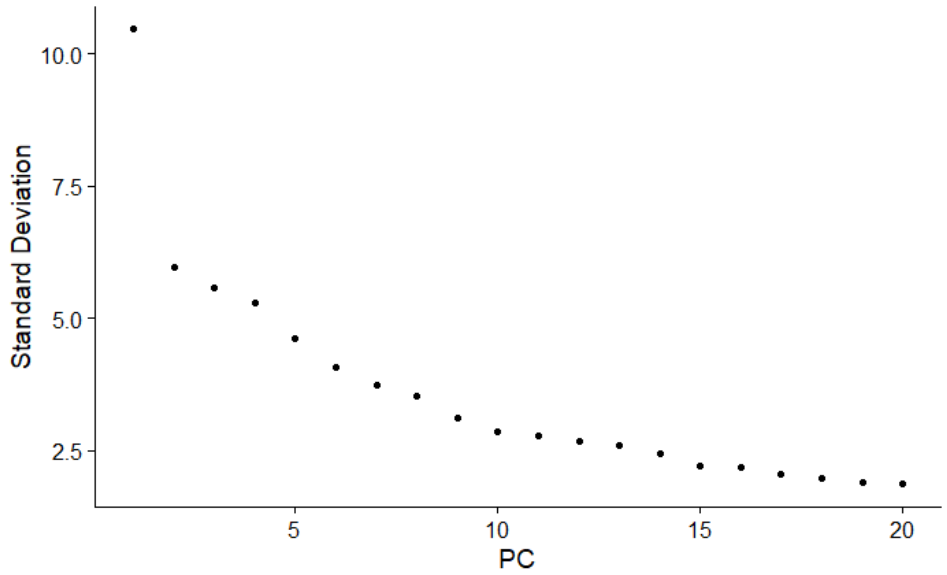

Figure 4: Elbow plot displaying PC analysis

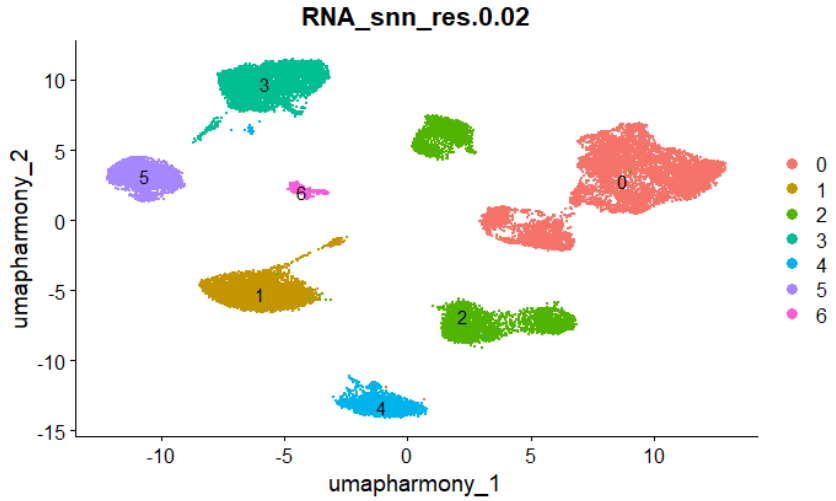

Figure 5: Pre-annotation cell type clustering

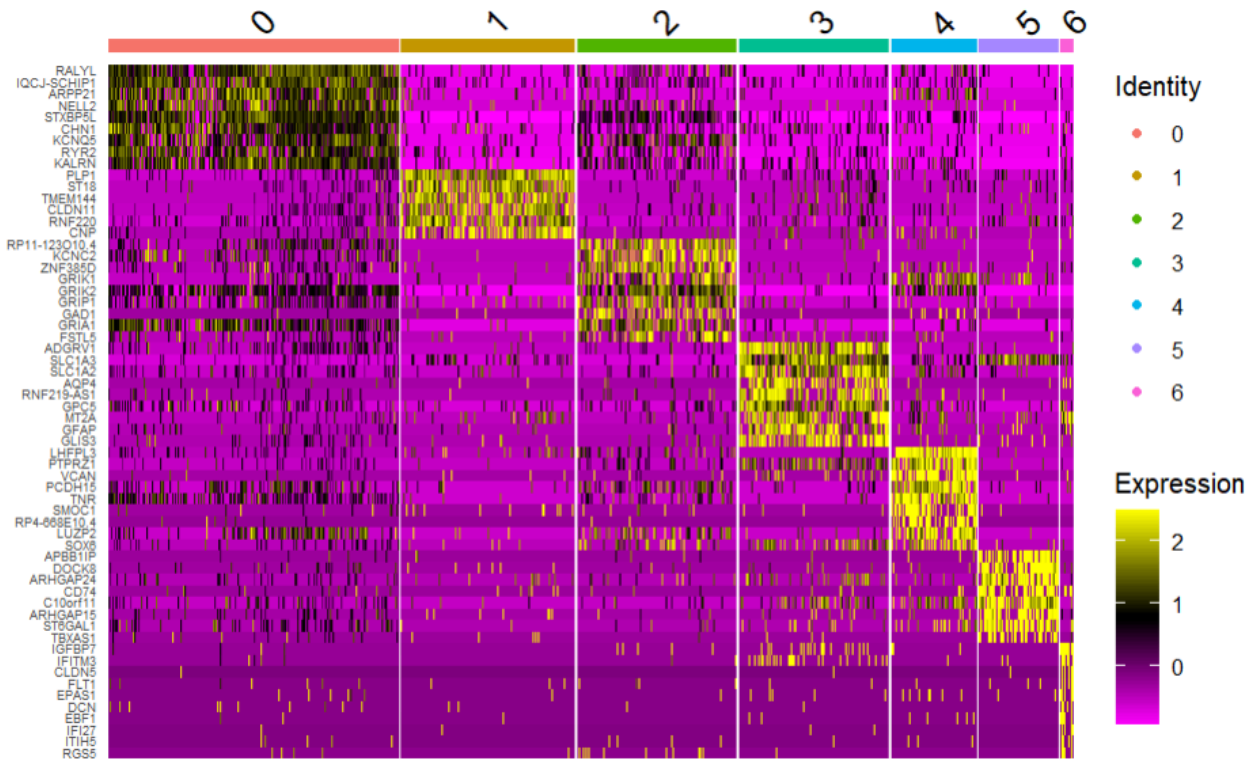

Figure 6: Heat map of gene expression for each cluster

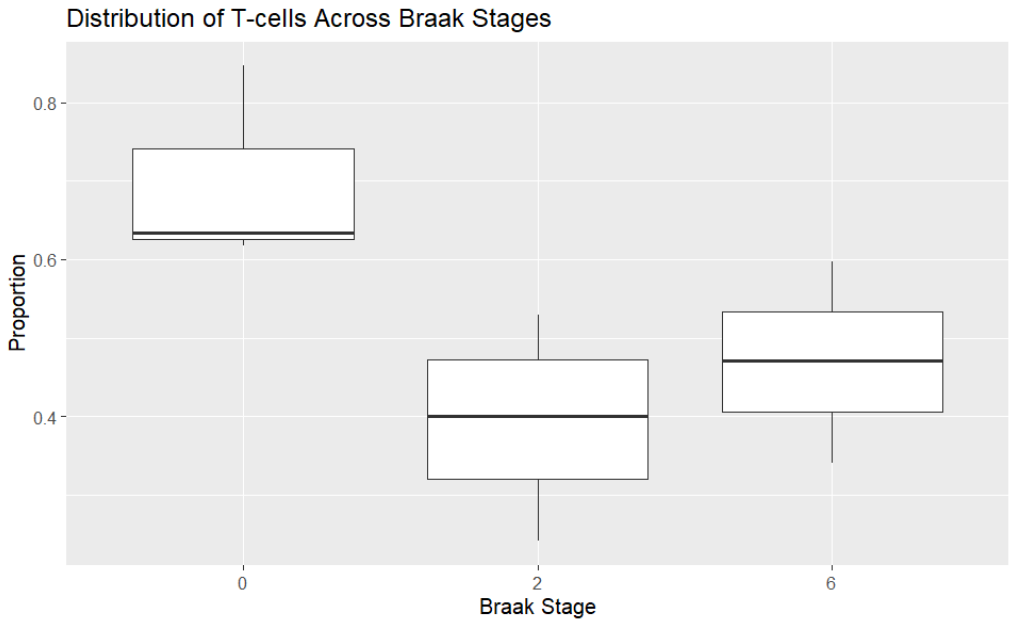

Figure 7: Proportion of T-cells across Braak stages for Alzheimer's Disease patients

Tables

Table 1: Determined cell types based on the most expressed genes in each cluster

Cluster	Genes of Interest	Cell Type(s)

0	CHN1 and MT2A	T-cells
1	IGFBP7 and MT2A	T-cells
2	IGFBP7 and TBXAS1	Macrophages
3	CD74 and ARHGAP24	Dendritic Cells
4	CD74 and IFITM3	Dendritic Cells
5	IFI27 and VCAN	Macrophages
6	CHN1 and MT2A	Macrophages