Emilia Nathan, Evan Gray, and Tomer Wenderow

Statistical Bioinformatics Project

Professor Slonim

30 November 2024

Statistical Bioinformatics Project Report:

Analysis of Various Clustering Methodologies on Hammond and GLP-1RA Mouse Microglial Data

1. Introduction

The Hammond et al. paper entitled, “Single cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes” rewrites scientific understanding of mouse microglial cells. In the paper, the researchers identify a minimum of nine unique microglial states uncovered through transcriptomic analysis of RNA found in mouse microglial cells. In addition to the unique states, one of the major takeaways from the paper is that the makeup of microglial cells was found to be most heterogeneous in younger samples while inflammation increased with age. Jointly, microglial signatures found after demyelinating injuries were used to identify various reactive microglia subtypes analogous to human multiple sclerosis. These findings were uncovered primarily through the use of single-cell RNA sequencing using mouse specimens and in-situ brain mapping. Overall, the paper’s conclusion is that microglial diversity is prominent primarily in young/developing and injured/aging mice.

Microglia, more broadly, play important roles in both development, the continuous maintenance of brain vitality and function, and their status as immune cells. Such immune cells are activated by changes from microglial homeostasis, causing migration and various actions to

preserve a homeostatic state against various pathological, chemical, or other physiological challenges. Despite the various functions and known importance of microglia, this study was prompted by the limited understanding of microglial classifications beyond "'resting', 'M1' (proinflammatory), or 'M2' (anti-inflammatory)" (Hammond). Understanding that these are perhaps oversimplifications that obscure potentially important findings through the discovery of greater heterogeneity, this study aimed to identify and distinguish microglial groups based on single-cell RNA-seq thanks to data acquired with modern high-throughput techniques. Crucially, the discovery and refinement of these states allows for further profiling of microglial states by age, sex, pathology, and other factors.

While the study is incredibly thorough and explored many different associated factors with microglial heterogeneity, development, and associated injury, we're most interested in pursuing alternate methods to the algorithmic techniques approached in the study, particularly 1) introducing new parameters for Principal Component Analysis (PCA) as opposed to Independent Component Analysis (ICA) undertaken by the Hammond paper 2) incorporating clustering techniques such as K-Means, Hierarchical clustering, and Gaussian Mixture Models (GMM) and 3) introducing new data from analogous mouse microglial data. During our time in Statistical Bioinformatics we have come to understand the importance of adding data and verifying computational findings with additional methods, so we hope to take an approach that extends this paper in those directions. Most concisely, our research question is the following:

Do the clustering and the conclusions in the Hammond et al. paper remain consistent across different samples of mouse microglial data and various clustering methodologies; how do these results translate to analogous or differential findings in an additional mouse microglial RNA sequencing population?

This question is motivated by our desire to further our own understanding of various clustering algorithms and both confirm and expand on the findings in the Hammond paper while incorporating new data into our biostatistical explorations.

2. Methods

Much of our initial work with this project involved preprocessing the Hammond data as well as the GLP such that it could be in a usable format. The Hammond data was already preprocessed to an extent so we simply loaded the .qs file into R. The GLP expression data and metadata were in csv formats. We used the Pandas Python package to filter the data because it appeared more responsive than R for the large dataset. The GLP dataset included cell types other than microglial cells so we isolated the microglial cells. We filtered the cells according to the Hammond paper, so we removed genes that were expressed less than 20 times total (0.025% of all cells in the dataset) and cells that expressed less than 650 total genes. After cleaning the GLP expression data and metadata we combined both into a Seurat object due to the sparsity of the counts and for redundancy of methods with the Hammond data. We then log normalized the data to account for potential skewness and performed feature selection using a variance stabilizing transformation. The variance stabilizing transformation algorithm adjusts the variance of each gene based on mean expression and selects 2000 genes with the most meaningful variance. We then centered and scaled the counts data for each selected gene.

For this portion of the project, we performed dimensionality reduction using principal component analysis (PCA). Although we had planned to use independent component analysis (ICA) as with the Hammond paper, we encountered issues with the implementation of the algorithm in R that made it incompatible with our Seurat objects. We used the amount of

variance explained by each principal component to determine that 50 principal components would capture almost all of the variance in the data. This value was the same for both datasets.

We used K-Means using the kmeans function in the base stats package as our first clustering approach. We used the PCA embeddings along with elbow plots to determine that 11 clusters would be optimal for each of the datasets. (R Core Team for kmeans, Seurat for embeddings).

We then performed hierarchical clustering using the FindClusters function in the Seurat package. We first constructed a shared nearest neighbors graph using the PCA embeddings and a k parameter of 20. Next, we used the Euclidean distance metric with the Smart Local Moving algorithm to detect communities in the data. The result of this analysis is the identification of potential subclusters as outlined in Figures 5a-e. The dendrograms provide a visual aid by which we could identify and merge subclusters. We used the top 5 most expressed genes along with the dendrogram to prune the tree and identify more dominant clusters. (Seurat FindNeighbors, and FindClusters).

We used Gaussian Mixture Models as our third clustering approach. We used the top 35 principal components due to memory constraints and it accounted for roughly 90% of the variance in the features. We used the mclust package to perform this clustering as it provides an uncertainty metric for membership to each cluster (Scrucca). This package builds a series of models using different numbers of clusters and identifies the optimal number using the Bayesian Information Criterion. We then identified the top 5 average expressed genes for each cluster for both datasets.

3. Results

To begin, we completed UMAP on Hammond and GLP-1RA data, K-Means clustering on Hammond and GLP-1RA data, and identified expression by cluster for both Hammond and GLP-1RA. We were initially able to validate the results of our UMAP on the GLP-1RA data by comparing them to the UMAP results offered by the Broad Institute's UMAP on the same data. The Broad's UMAP by cell type is below. MG indicates the microglial samples, which we extracted and performed PCA on.

UMAP on GLP-1RA Data (The Broad Institute)

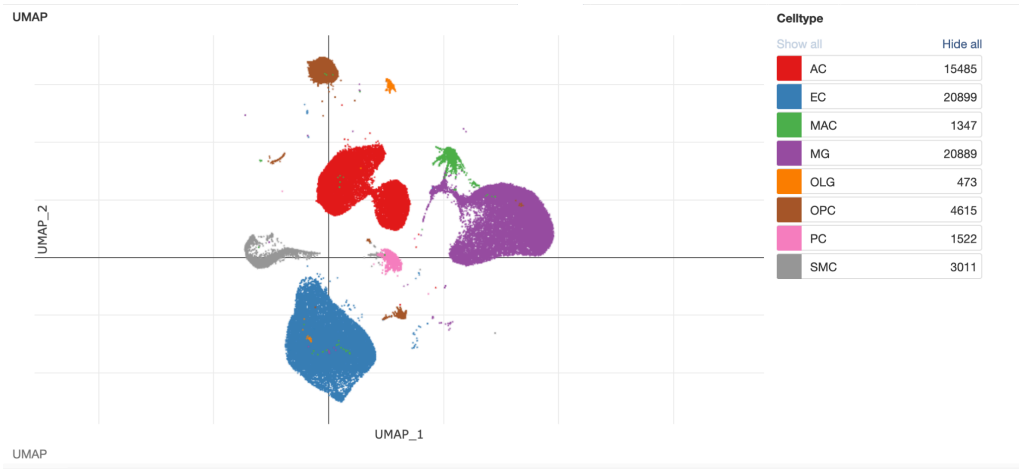

Figure 1a. UMAP (Performed by The Broad Institute) on GLP-1RA Data, identifying cell types by clustering

UMAP on GLP-1RA Data

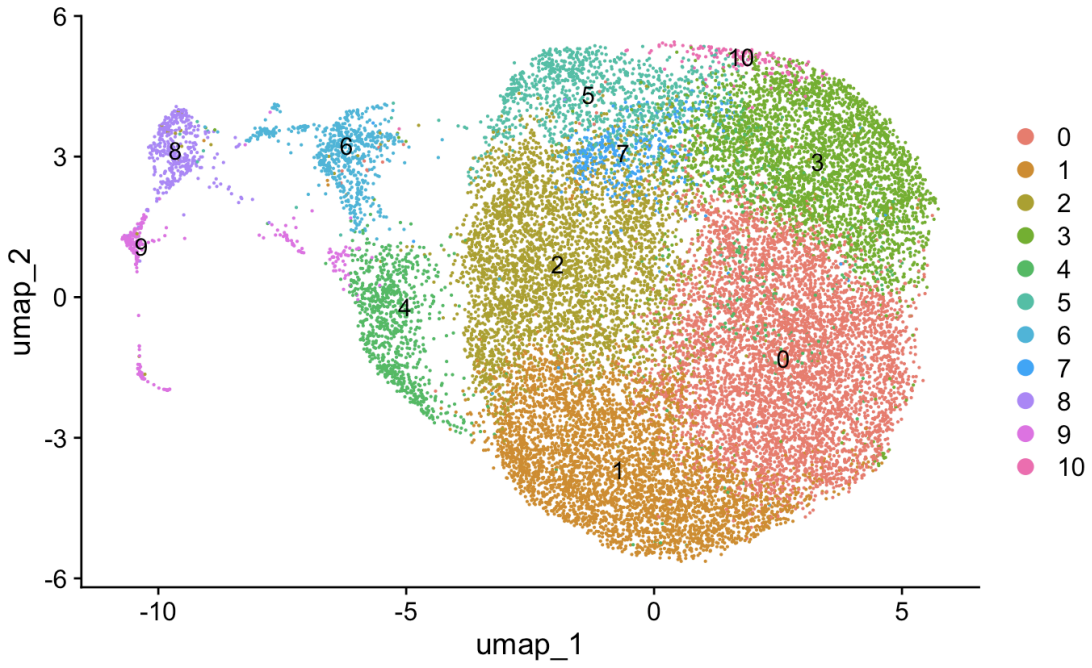

Figure 1b. UMAP Performed on GLP-1RA Samples

UMAP on Hammond Data

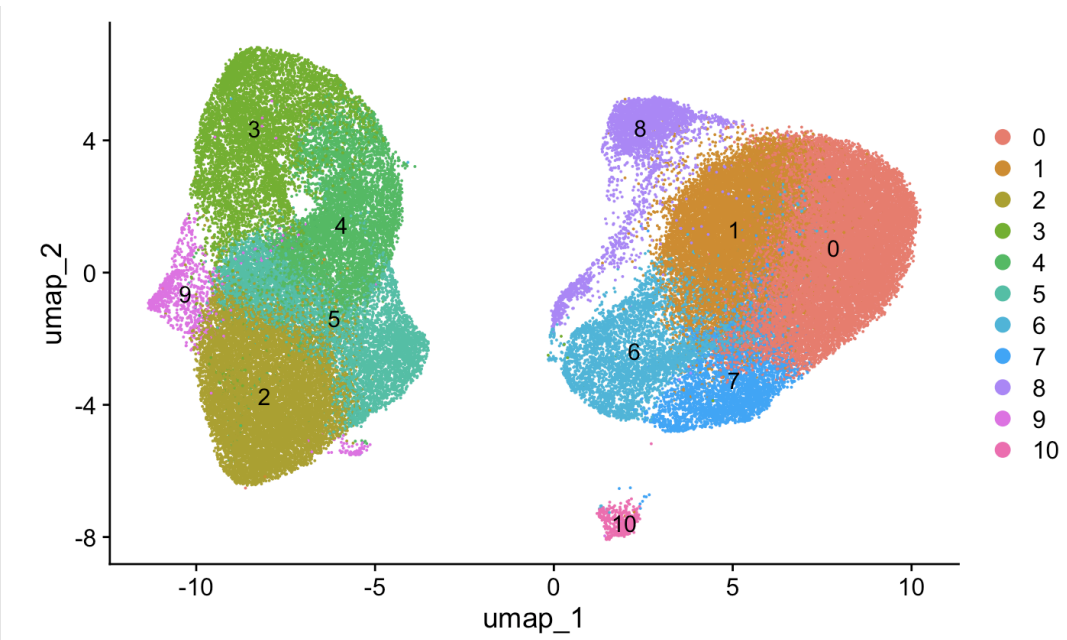

Figure 1c. UMAP Performed on Hammond Data

UMAP on Hammond Data by Age

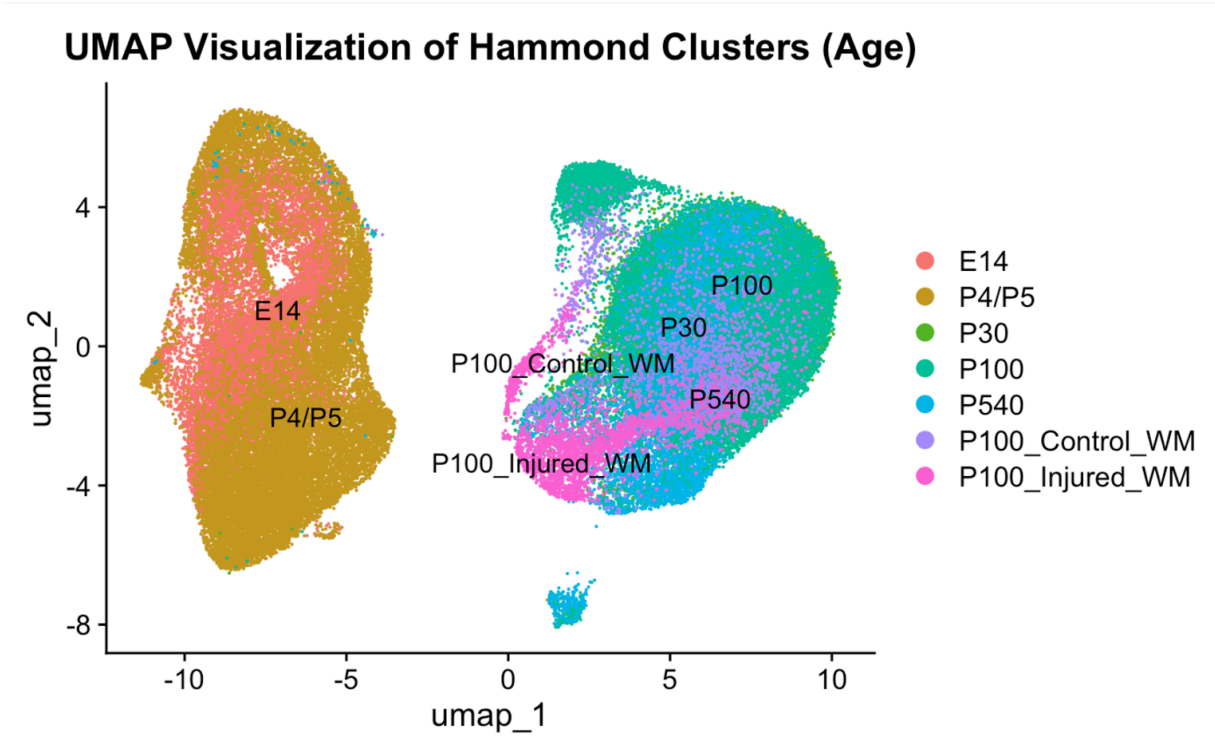

Figure 1f. Visualization of Hammond dataset using UMAP clustered by Age

UMAP on GLP-1RA Data by Age

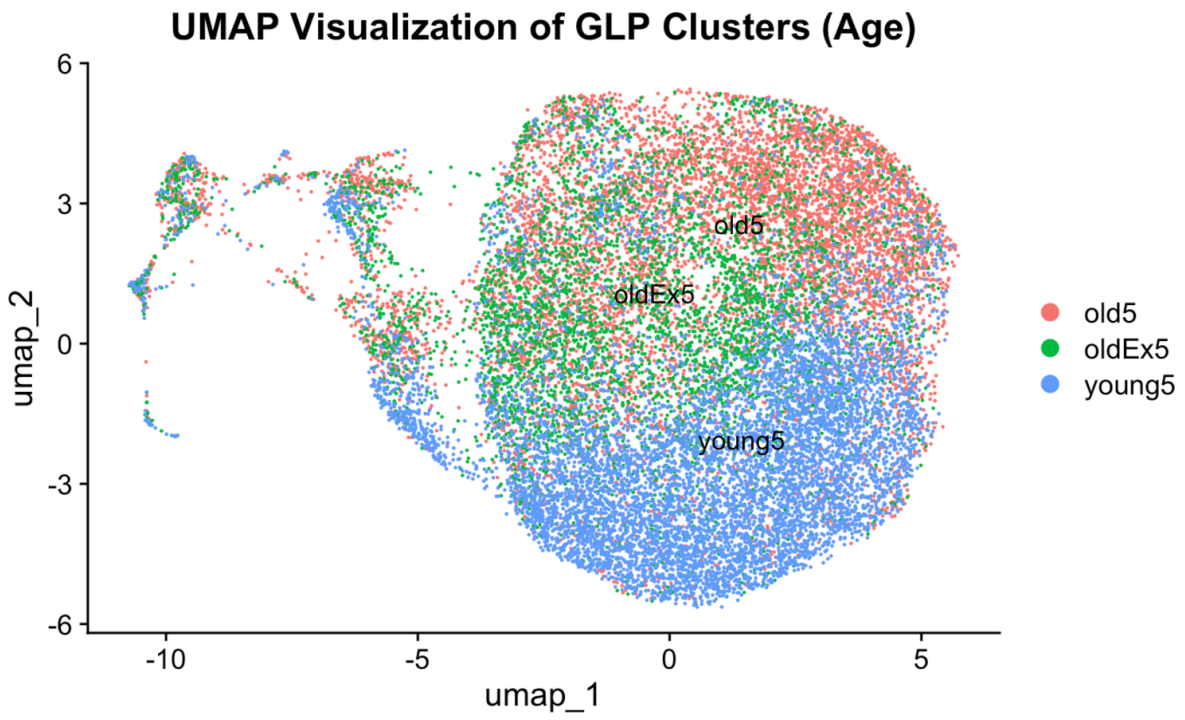

Figure 1g. Visualization of GLP-1RA dataset using UMAP clustered by Age. Notice the characteristic “hook structure” with highly mixed age ranges.

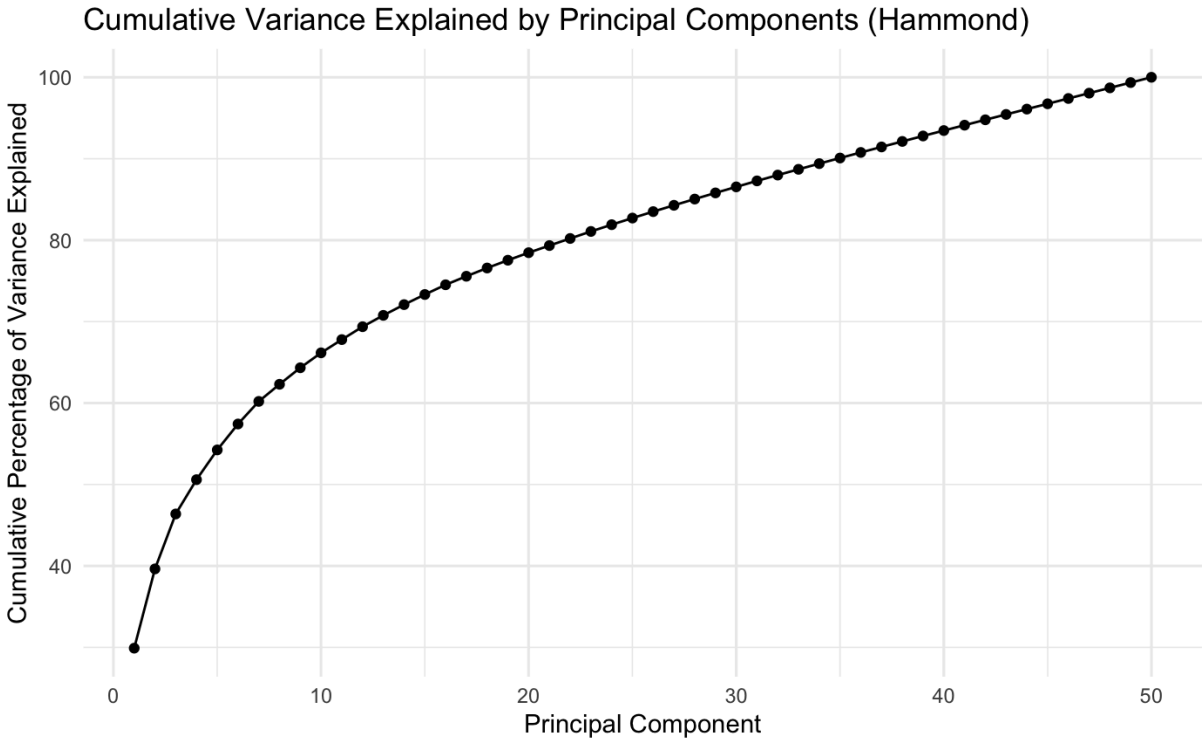

Figure 2a: Cumulative Explained Variance by Principal Component for Hammond Data

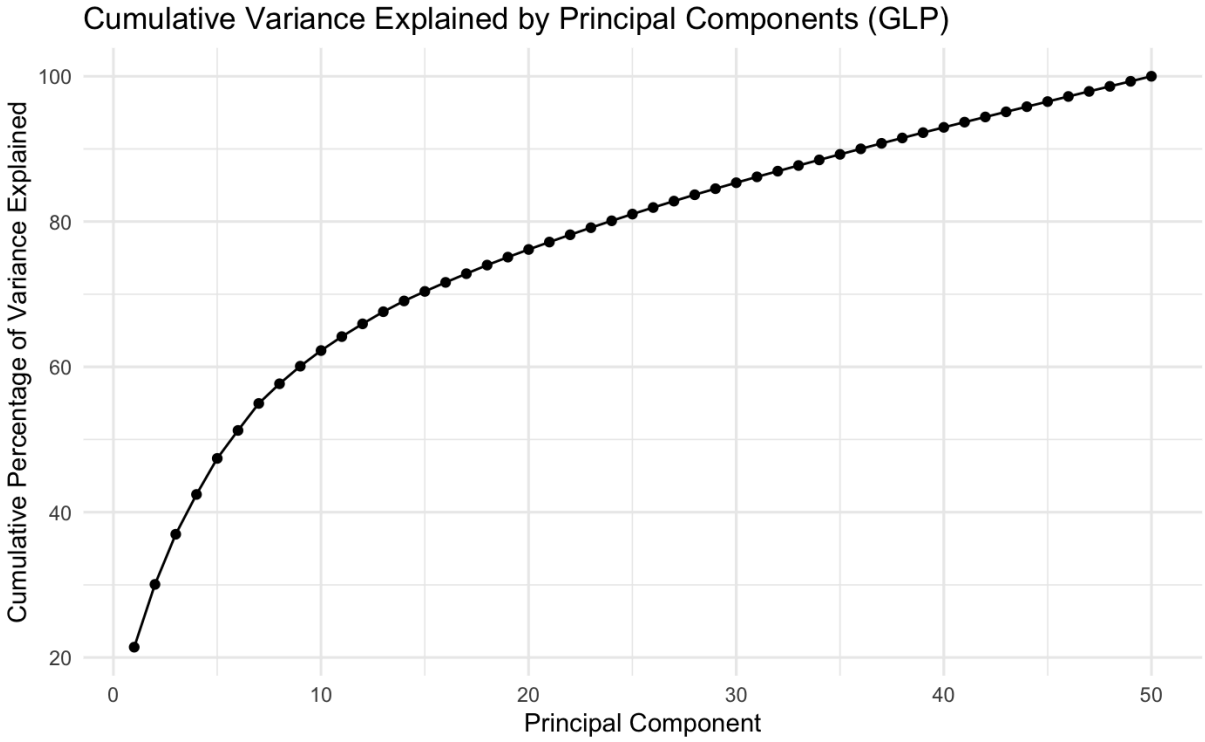

Figure 2b: Cumulative Explained Variance by Principal Component for GLP1-RA Data

UMAP Visualization of GMM Clusters on Hammond Data

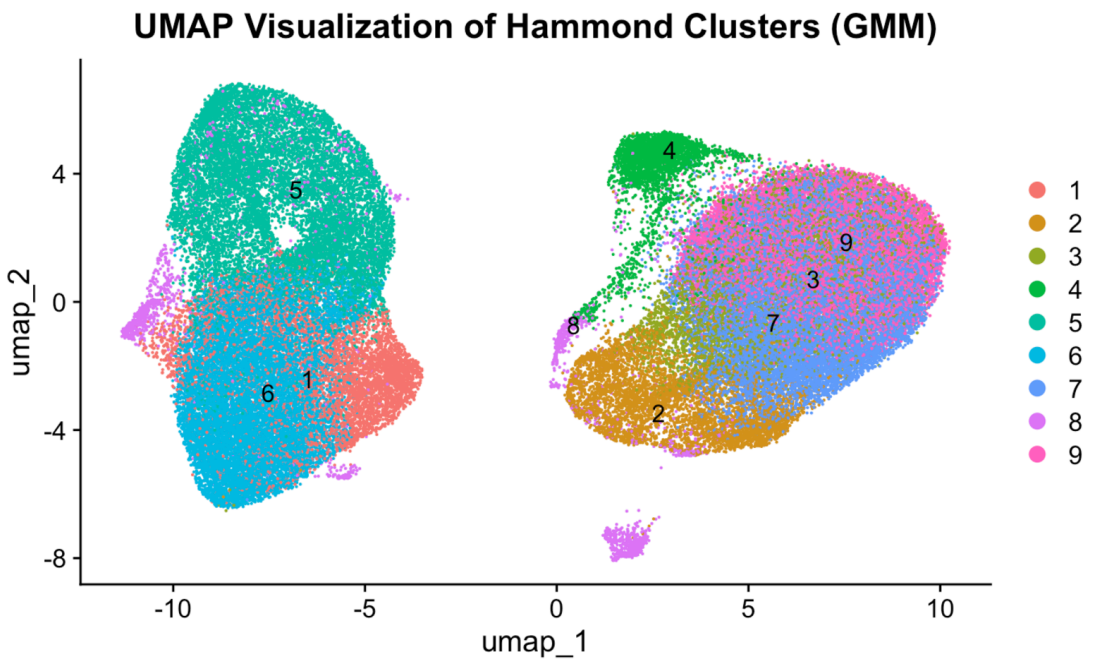

Figure 3a. Gaussian Mixture Model Clustering on Hammond Data

UMAP Visualization of GMM Clusters on GLP-1RA Data

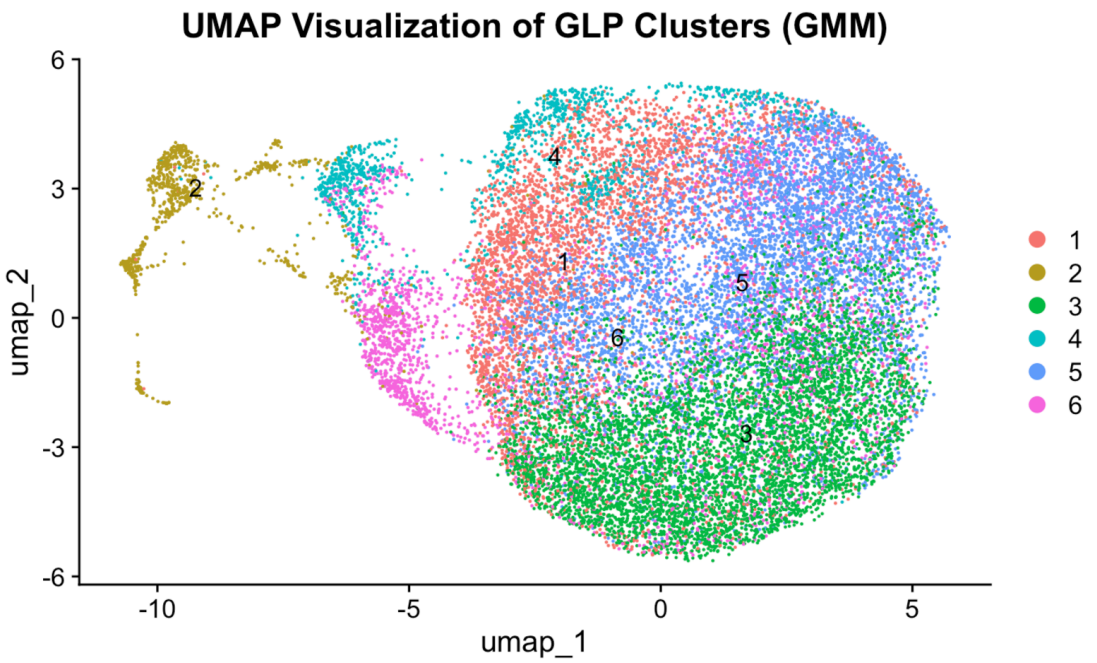

Figure 3b. Gaussian Mixture Model Clustering on GLP-1RA Data

UMAP Visualization of K-Means Clusters on Hammond Data

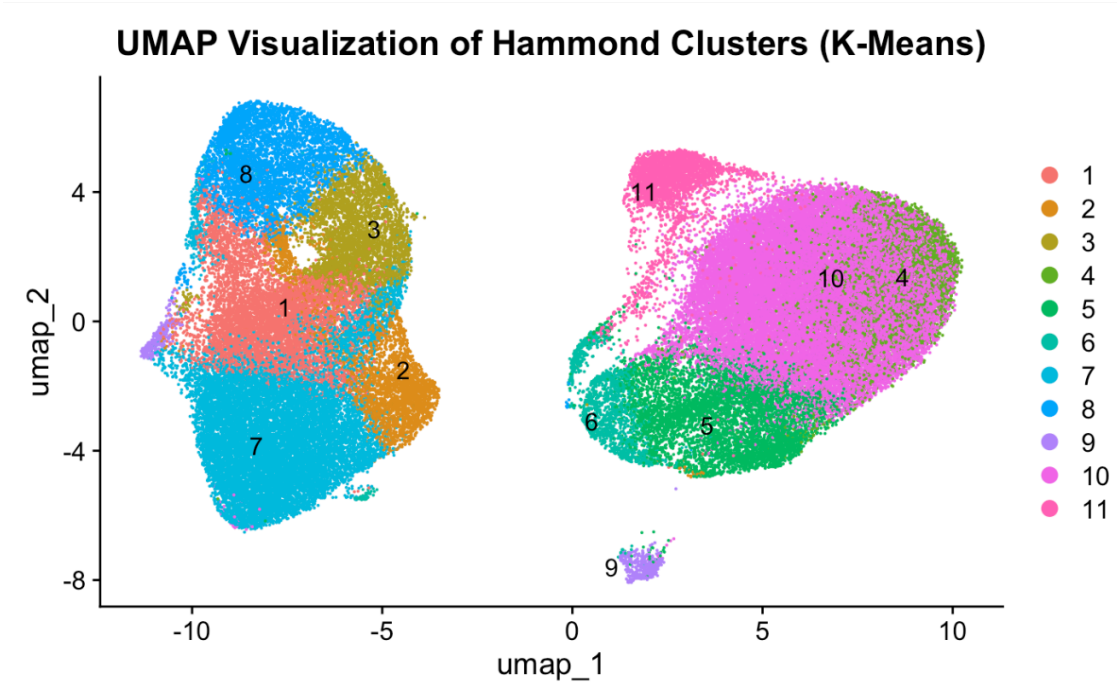

Figure 4a. K-Means Clustering on Hammond Data

K-Means Clustering on Hammond Data

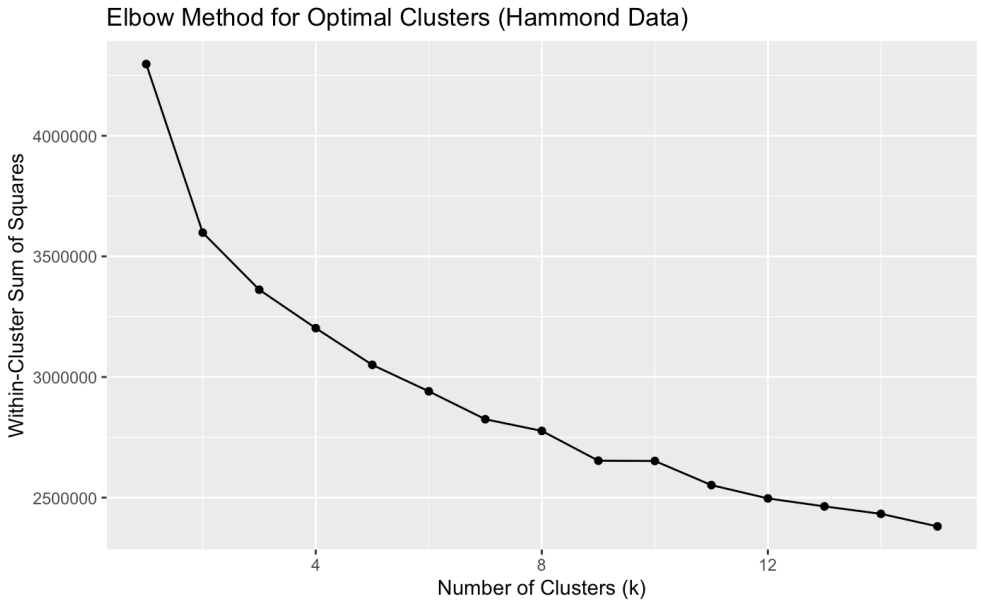

Figure 4b. K-Means Clustering for Optimal Clusters on Hammond Data

UMAP Visualization of K-Means Clusters on GLP-1RA Data

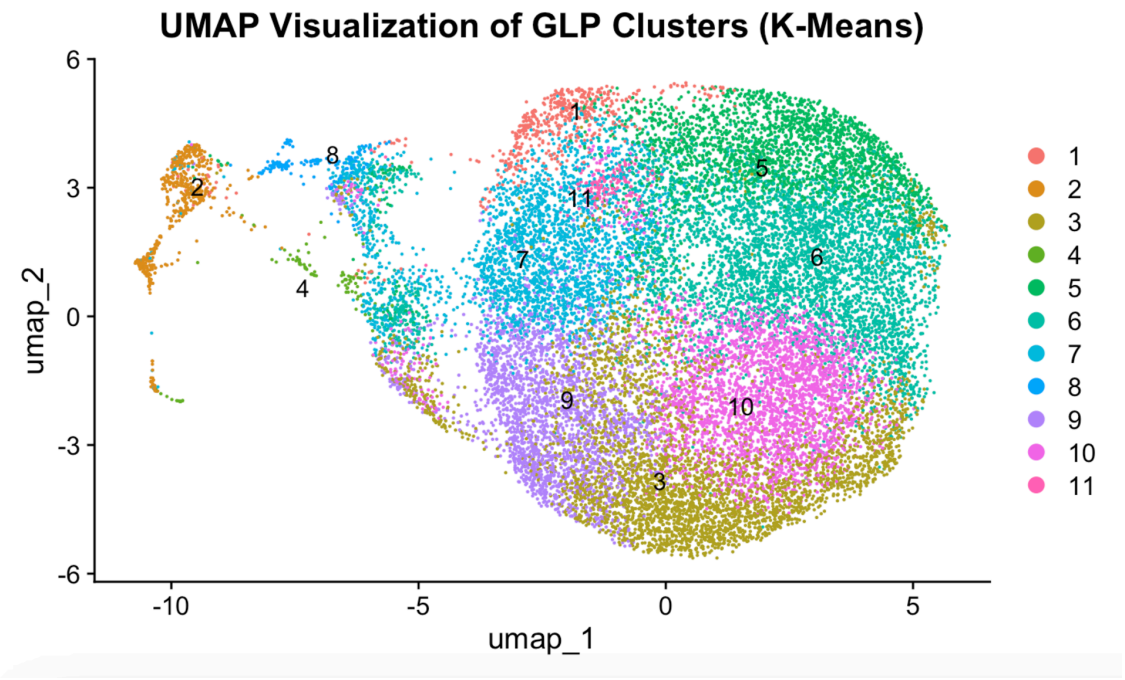

Figure 4c. K-Means Clustering on GLP-1RA Data

K-Means Clustering on GLP-1RA Data

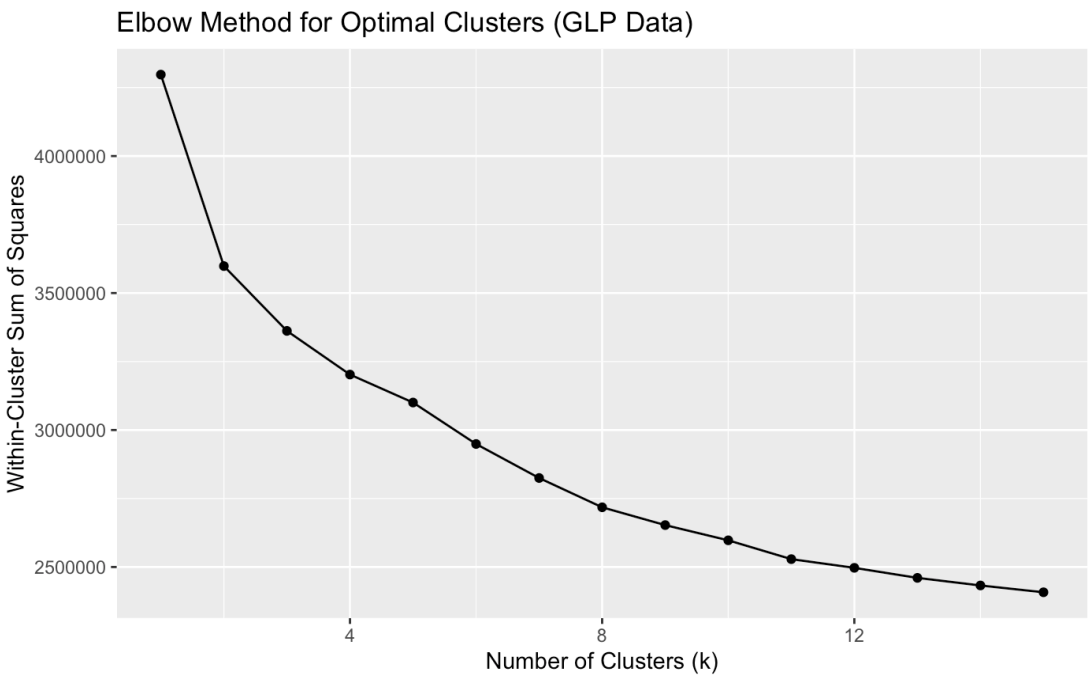

Figure 4d. K-Means Clustering for Optimal Clusters on GLP-1RA Data

UMAP Visualization of Hierarchical Clusters on Hammond Data

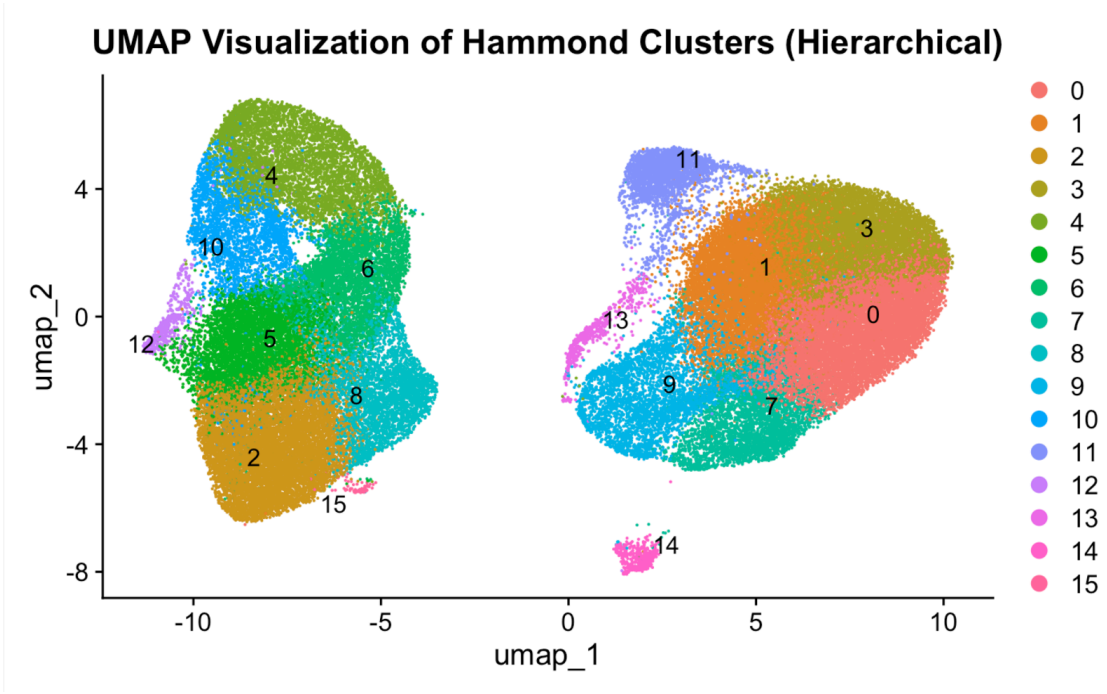

Figure 5a. UMAP Visualization of Hierarchical Clusters on Hammond Data

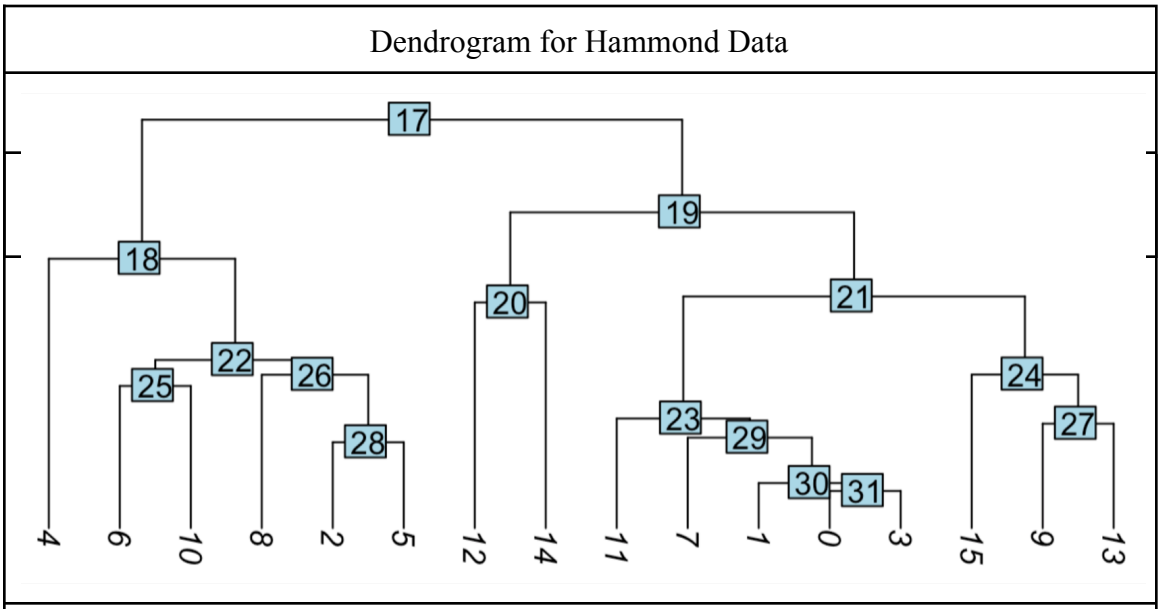

Figure 5b. Dendrogram for Hammond Data formed based on hierarchical clustering

UMAP Visualization of Hierarchical Clusters on GLP-1RA Data

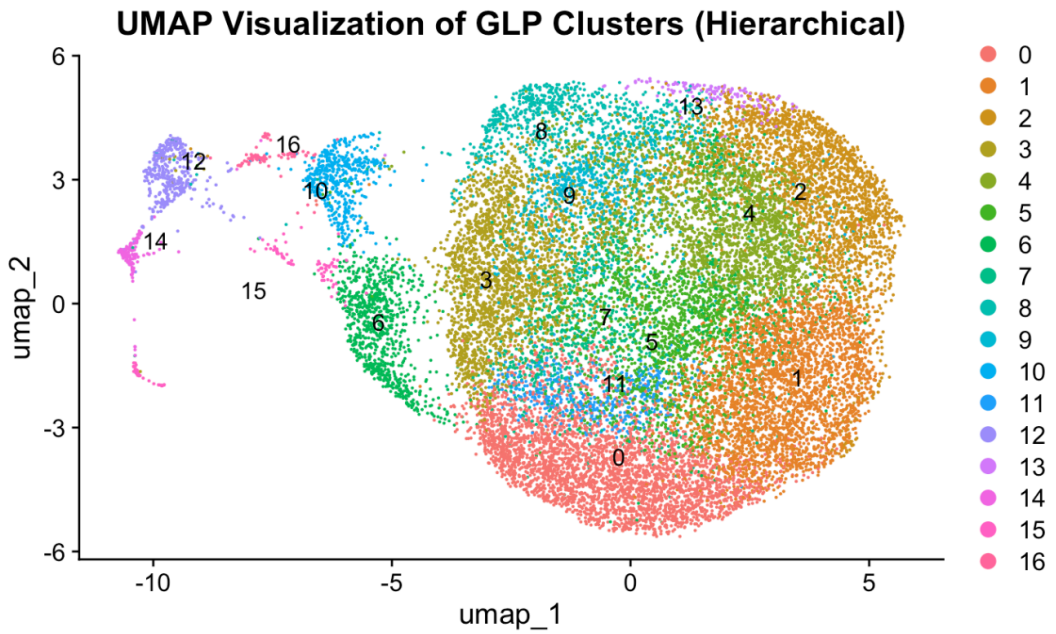

Figure 5c. UMAP Visualization of Hierarchical Clusters on GLP-1RA Data

Dendrogram for GLP-1RA Data

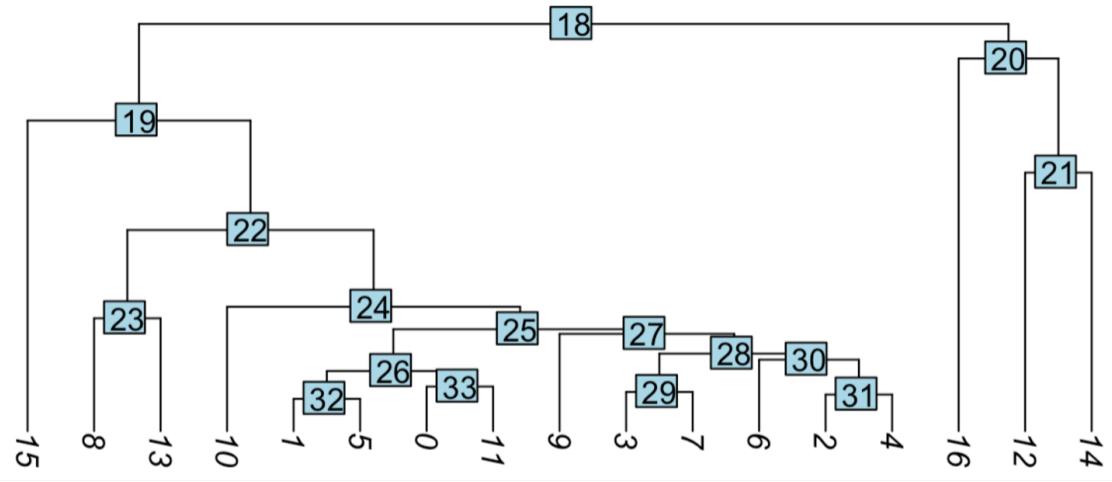

Figure 5d. Dendrogram for GLP-1RA Data formed based on hierarchical clustering

UMAP Visualization of Hierarchical Clusters with Pruning on GLP-1RA Data

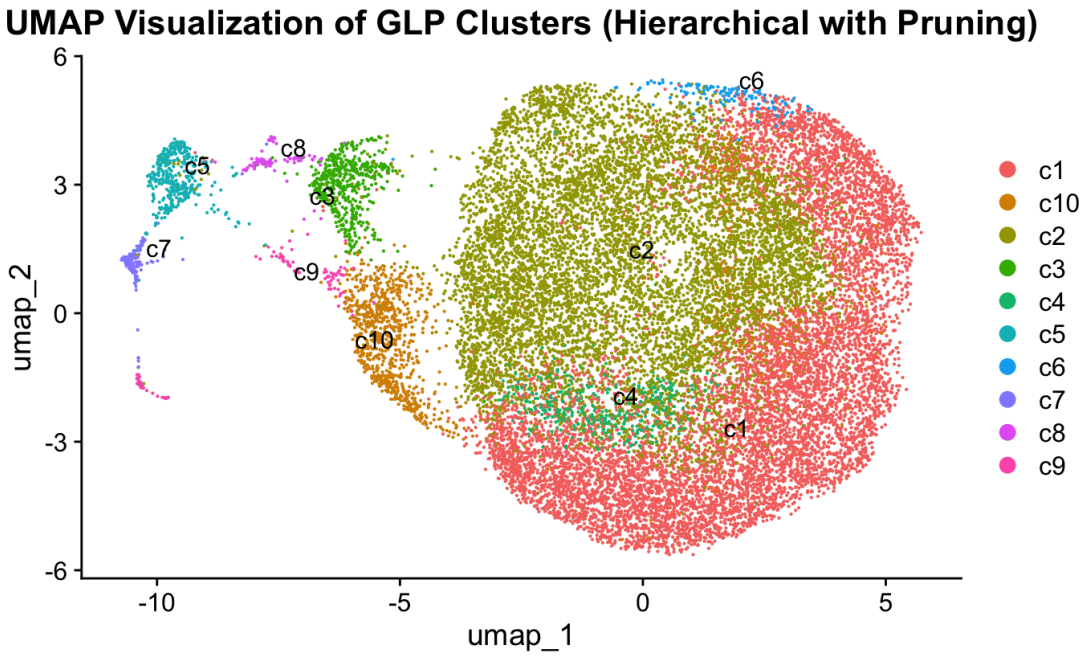

Figure 5e. UMAP Visualization of Hierarchical Clusters with Pruning on GLP-1RA

K-Means Clustering (Hammond Data)
See Hammond_K-Means_Gene_Clusters.csv for more detailed expression numerics

Cluster	Top 5 Genes Expressed	Cluster	Top 5 Genes Expressed
g1	1. Apoe 2. Tmsb4x 3. Ftl1 4. Cst3 5. C1qc	g7	1. Ftl1 2. Tmsb4x 3. Apoe 4. Ctsb 5. Lgmn
g2	1. Tmsb4x 2. Ftl1 3. Apoe 4. Ctsb 5. Lgmn	g8	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa
g3	1. Tmsb4x 2. Apoe 3. Ftl1 4. Ctsb 5. Lgmn	g9	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa
g4	1. Apoe 2. Malat1 3. Cd74 4. Tmsb4x 5. Ftl1	g10	1. Cst3 2. Malat1 3. Apoe 4. Tmsb4x 5. Ctsd
g5	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa	g11	1. Apoe 2. Ftl1 3. Tmsb4x 4. Cst3 5. Itm2b
g6	1. Spp1 2. Apoe 3. Ftl1 4. Ctsb 5. Tmsb4x		

K-Means Clustering (GLP-1RA Data)
See GLP_K-Means_Gene_Clusters.csv for more detailed expression numerics

Cluster	Top 5 Genes Expressed	Cluster	Top 5 Genes Expressed
g1	1. Malat1 2. Apoe 3. Cst3 4. mt-Atp6 5. Tmsb4x	g7	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa
g2	1. Cd74 2. Malat1 3. Tmsb4x 4. mt-Atp6 5. Fth1	g8	1. Cst3 2. Malat1 3. Tmsb4x 4. mt-Atp6 5. C1qa
g3	1. Malat1 2. Tmsb4x 3. mt-Atp6 4. Bsg 5. mt-Co2	g9	1. Cst3 2. Tmsb4x 3. Malat1 4. C1qa 5. C1qb
g4	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. C1qb	g10	1. Cst3 2. Tmsb4x 3. Fth1 4. Apoe 5. mt-Atp6
g5	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa	g11	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. Hexb
g6	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. mt-Atp6		

Hierarchical Clustering (Hammond Data)

See Hammond_Hierarchical_Gene_Clusters.csv for more detailed expression numerics

Cluster	Top 5 Genes Expressed	Cluster	Top 5 Genes Expressed	Cluster	Top 5 Genes Expressed
g0	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa	g6	1. Tmsb4x 2. Apoe 3. Ftl1 4. Cst3 5. Ctsb	g12	1. Apoe 2. Ftl1 3. Tmsb4x 4. C1qc 5. Itm2b
g1	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa	g7	1. Malat1 2. Cst3 3. Hexb 4. C1qa 5. Tmsb4x	g13	1. Cst3 2. Tmsb4x 3. Malat1 4. Apoe 5. Ctsd
g2	1. Apoe 2. Ftl1 3. Tmsb4x 4. Cst3 5. Itm2b	g8	1. Apoe 2. Ftl1 3. Spp1 4. Ctsb 5. Tmsb4x	g14	1. Malat1 2. Apoe 3. Cd74 4. Cst3 5. Tmsb4x
g3	1. Malat1 2. Cst3 3. Hexb 4. C1qa 5. Tmsb4x	g9	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. Ctsd	g15	1. Apoe 2. Ftl1 3. Tmsb4x 4. Cst3 5. Ctsb
g4	1. Tmsb4x 2. Apoe 3. Ftl1 4. Lgmn 5. Cst3	g10	1. Ftl1 2. Tmsb4x 3. Apoe 4. Ctsb 5. Lgmn		
g5	1. Tmsb4x 2. Ftl1 3. Apoe 4. Ctsb 5. C1qb	g11	1. Cst3 2. Malat1 3. Hexb 4. C1qa 5. Tmsb4x		

Hierarchical Clustering (GLP-1RA Data)
See GLP_Hierarchical_Gene_Clusters.csv for more detailed expression numerics

Cluster	Top 5 Genes Expressed	Cluster	Top 5 Genes Expressed	Cluster	Top 5 Genes Expressed
g0	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa	g6	1. Cst3 2. Malat1 3. Tmsb4x 4. m-Atp6 5. C1qa	g12	1. Apoe 2. Malat1 3. Cd74 4. Tmsb4x 5. Fth1
g1	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa	g7	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. Hexb	g13	1. Cst3 2. Tmsb4x 3. mt-Atp6 4. C1qb 5. C1qa
g2	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. C1qb	g8	1. Cst3 2. Tmsb4x 3. Malat1 4. C1qa 5. C1qb	g14	1. Tmsb4x 2. Cd74 3. Malat1 4. Actb 5. Cst3
g3	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. C1qb	g9	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. C1qb	g15	1. Malat1 2. mt-Atp6 3. mt-Co2 4. mt-Co1 5. Tmsb4x
g4	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. mt-Atp6	g10	1. Cst3 2. Malat1 3. Tmsb4x 4. mt-Atp6 5. C1qa	g16	1. Malat1 2. Tmsb4x 3. mt-Atp6 4. Bsg 5. mt-Co2
g5	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa	g11	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa		

Hierarchical Clustering with Pruning (GLP-1RA Data)
See GLP_Hierarchical_Pruning_Gene_Clusters.csv for more detailed expression numerics

Cluster	Top 5 Genes Expressed	Cluster	Top 5 Genes Expressed
g1	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa	g6	1. Cst3 2. Tmsb4x 3. mt-Atp6 4. C1qb 5. C1qa
g2	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. C1qb	g7	1. Tmsb4x 2. Cd74 3. Malat1 4. Actb 5. Cst3
g3	1. Cst3 2. Malat1 3. Tmsb4x 4. mt-Atp6 5. C1qa	g8	1. Malat1 2. Tmsb4x 3. mt-Atp6 4. Bsg 5. mt-Co2
g4	6. Cst3 7. Malat1 8. Tmsb4x 9. Hexb 10. C1qa	g9	1. Malat1 2. mt-Atp6 3. mt-Co2 4. mt-Co1 5. Tmsb4x
g5	1. Apoe 2. Malat1 3. Cd74 4. Tmsb4x 5. Fth1	g10	6. Cst3 7. Malat1 8. Tmsb4x 9. mt-Atp6 10. C1qa

GMM Clustering (Hammond Data)
See Hammond_GMM_Gene_Clusters.csv for more detailed expression numerics

Clust.	Top 5 Genes Expressed	Uncertainty	Clust.	Top 5 Genes Expressed	Uncertainty
g1	1. Apoe 2. Ftl1 3. Spp1 4. Tmsb4x 5. Ctsb	325.24	g7	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa	1390.39
g2	1. Malat1 2. Cst3 3. Hexb 4. Tmsb4x 5. Ctsd	195.94	g8	1. Apoe 2. Malat1 3. Tmsb4x 4. Cst3 5. Ftl1	37.01
g3	1. Malat1 2. Cst3 3. Hexb 4. Tmsb4x 5. C1qa	686.39	g9	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa	1434.30
g4	1. Cst3 2. Malat1 3. Hexb 4. Tmsb4x 5. C1qa	13.10			
g5	1. Tmsb4x 2. Ftl1 3. Apoe 4. Ctsb 5. Lgmn	208.36			
g6	1. Apoe 2. Tmsb4x 3. Ftl1 4. Cst3 5. C1qc	505.40			

GMM Clustering (GLP-1RA Data)

See GLP_GMM_Gene_Clusters.csv for more detailed expression numerics

Cluster	Top 5 Genes Expressed	Uncertainty
g1	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. C1qb	325.27
g2	1. Malat1 2. Tmsb4x 3. Cst3 4. Cd74 5. mt-Atp6	2.17
g3	1. Cst3 2. Malat1 3. Tmsb4x 4. Hexb 5. C1qa	321.53
g4	1. Cst3 2. Tmsb4x 3. mt-Atp6 4. C1qa 5. Malat1	30.46
g5	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. C1qb	503.67
g6	1. Cst3 2. Malat1 3. Tmsb4x 4. C1qa 5. Hexb	212.57

4. Discussion

Over the course of the project, we have used the Hammond paper data as well as new data from a different set of microglial single-celled RNA sequencing data to confirm the results of the Hammond paper. We filtered both data sets and then used three clustering methods to separate the data into distinct clusters of various microglial states. Finally, we found the highest expressed genes in each cluster. Our analysis showed that the clusters found are different for each approach for both data sets, as expected. However, they are similar enough to where we can confidently conclude that the microglial signatures found in the Hammond paper are valid and that the conclusions drawn in the Hammond paper are sound. We accomplished everything we set out to do except for performing ICA, which was hindered by issues using the *fastICA* package in R, and instead we used PCA for dimensionality reduction and data visualization.

We have successfully filtered and clustered both the raw Hammond paper scRNA sequencing data as well as scRNA sequencing data from a second data set, GLP-1RA data. The data comes from a study looking at scRNA sequencing data from all glial cells in mice at various ages treated with GLP-1 or given placebo. GLP-1R is a GLP-1 agonist drug canonically used to treat obesity. A common brand name is Ozempic®. The drug mimics the effects of Glucagon-like peptide-1, which is a peptide hormone that regulates blood sugar, digestion, and appetite. In this paper, they sought to find the effects of GLP-1R on glial cell expression. For the purposes of our project, we only used the microglial cells found, which were identified computationally and plotted using PCA in the original paper. The paper identified and isolated the microglial cells by identifying cells that expressed canonical microglial genes. These cells were identified as microglial. This method does introduce some uncertainty in this data set and slightly compromises the quality of the data. The Hammond paper, on the other hand, used

biological methods to physically separate out the microglial cells. Then they further filtered the cells computationally leading to higher quality data. While the GLP-1RA data is good quality data from the Broad Institute, it is important to recognize that the quality is not as good as the Hammond data.

We sought to use different clustering methods in both data sets to confirm the clusters found in the Hammond paper. Hammond et al. found 9 distinct clusters with unique expression patterns and signatures. These clusters were highly correlated with age with clusters exhibiting unique expression levels. While slightly unclear based on their descriptions alone, Hammond et al. seemed to use hierarchical clustering based on the methods they reference. Our objective was to use other clustering methods to see if there are other ways to cluster the data that may either validate the clustering used in the Hammond paper or present additionally valuable findings via alternative approaches. We used three approaches to cluster: K-Means, Hierarchical Clustering, and Gaussian Mixture Model (GMM). Our clustering analysis revealed slightly different clusters based on the approach. The K-means method showed 11 distinct clusters (Figure 4a-d). It is important to note that all of our clustering analysis showed two big clusters with multiple subclusters each, while the Hammond paper had one large cluster. This may be due to our using PCA instead of ICA. Additionally, this could be due to the fact that we only did one round of PCA instead of two, which may not have removed all of the contaminants in the data. Upon retrieving the PCA visualizations of the Hammond data and GLP-1RA data, two large clusters were found to be present in the Hammond while 1 was present in the GLP-1RA data. Investigating further led us to believe these clusters were likely formed due to age-related similarities. When plotting the data to color by age (Figures 1f and 1g), it became apparent that the Hammond data was, in fact, heavily clustered by age, as E14 and P4/P5 were found

exclusively in the left-hand cluster. These represent the two youngest cohorts of mice tested, with these samples being pulled from “isolated whole brains from mice at embryonic day 14.5 (E14.5), early postnatal day 4/5 (P4/P5)” (Hammond). All of the other samples were found in the right-hand cluster, encompassing the older samples and the injured samples. Our findings that the clusters correlate highly with age is consistent with the Hammond paper and serves to corroborate their results. The distinctions by age were less clear in the GLP-1RA clustering by age, as we see less clear distinctions by age/cluster between all of the data points. However, it is relevant to note where these cells are located in the PCA as we can see connections to age in later clustering methodologies.

Most notably, in the case of GMM on GLP-1RA data (Figure 3b), Cluster 3 seems to align directly with the young5 age group. Interestingly, the GLP-1RA paper set out to investigate potentially potent anti-aging therapeutic results across diverse brain cell types as a result of seeing improved human condition (cognitive abilities) in Parkinson’s and diabetic patients. If the GLP-1RA paper did uncover anti-aging properties associated with the treatment, we would expect to see more of the GLP-1RA cells exhibiting similar expression patterns to the younger Hammond clusters/cells. However, the youngest Hammond samples were found to uniquely express certain genes as follows, “Uniquely expressed genes found predominantly at the youngest ages (E14.5 and P4/P5) included arginase 1 (Arg1, Cluster 1), ribonucleotide reductase M2 (Rrm2, Cluster 2a), ubiquitin-conjugating enzyme E2C (Ube2c, Cluster 2b), centromere protein A (Cenpa, Cluster 2c), fatty acid binding protein 5, epidermal (Fabp5, Cluster 3), osteopontin (Spp1, Cluster 4), heme oxygenase 1 (Hmox1, Cluster 5), and membrane-spanning 4-domains, subfamily A, member 7 (Ms4a7, Cluster 6)” (Hammond). While we initially hypothesized that these genes would be more prominent among all GLP-1RA data should its

anti-aging properties be realized, upon further investigation it seems as though many of these genes are associated with development, and while it's anticipated they would be prevalent in juvenile cells, it's not necessarily true that anti-aging results in cells re-expressing crucial *developmental* genes, as such cells would have surpassed the stage where the expression of such genes is necessary. While we also initially expected that the mice classified as "young" in the GLP-1RA samples would be most similar in their expression patterns/clustering to the youngest mice sampled by the Hammond paper. However, the "young" mice sampled in the GLP-1RA paper were 2–3 months old while the youngest mice sampled by the Hammond paper were taken during embryonic development day 14.5, and so there is a large gap in sampling ages to be aware of even between mice of the youngest category.

While we saw that in the GMM on GLP-1RA data, Cluster 3 seems to align directly with the young5 age group (comparing Figures 3b to 1g), the old5 and oldEx5 groups don't align exactly to any of the clusters but fall into the same overall large visual cluster. The "hook" part of the classical GLP-1RA visualization, however, and all of the data points to the left of the large visual cluster containing young5, old5, and oldEx5 seems to contain a mix of all the age ranges. We hypothesize that this region potentially is demonstrative of debris associated with cells of all ages.

A gene of note is Apoe, which has been shown to be involved in brain debris clearing. The protein it expresses is involved in making a protein that helps carry cholesterol and other types of fat in the bloodstream, and issues with this protein have been linked to an increased risk of Alzheimer's disease (AD). Our clustering analysis showed that there is no significant difference in Apoe expression levels between old and young mice microglia in both the Hammond data set and the GLP-1RA data, indicating that old age itself likely does not lead to

development of AD. These results could also be explained by the fact that Apoe plays a protective role as well as an injury resolution role, so in young mice, Apoe is highly expressed to protect the cells, while in older mice it is expressed to counteract the harmful effects of disease states. There are, however, significant differences in Apoe expression between the Hammond and GLP-1RA data. The Hammond cells have significantly higher expression levels of Apoe compared to the GLP-1RA treated cells, demonstrating that there could be a correlation between younger cells and higher Apoe expression since the GLP-1RA cells appear younger than they actually are. The Apoe gene remains a gene of interest for AD research and further research must be done to determine its effects on AD prevalence and Apoe expression with age

Another large difference between the Hammond data and GLP-1RA data is the higher prevalence of the gene Cst3 in the GLP-1RA cells than the Hammond cells. Cst3 has been found to play a neuroprotective role in Parkinson's disease. We can therefore attribute this difference in the data to the GLP-1RA treatment itself, rather than differences in the inherent microglial signatures associated with aging. The most common genes in each cluster were Apoe, Malat1, Tmsb4x, Cst3 and Ftl1, which are all common genes expressed characteristically by microglial cells. These cells were highly expressed in both the GLP-1RA data and the Hammond data. This further corroborates the similarities between the clustering.

Hierarchical clustering allowed us to merge clusters based on their most expressed genes and their relative position in the dendrogram. The Hammond dataset did not appear to require any pruning while the GLP-1RA dataset appeared to have redundant top genes. Using these details we were able to merge (from the previous hierarchical GLP-1RA clusters) clusters 0, 11, 5 and 1 into a new cluster and merge clusters 2, 4, 7, 3 and 9 into a new cluster while retaining

the independence of the remaining clusters. These new clusters displayed a higher level of distinction between the highest expressed genes.

The Gaussian Mixture Model approach was most interesting for its uncertainty metric. We took the cumulative level of uncertainty for each cell in each GMM cluster to understand how well each of the clusters captures the structure of the data. In the Hammond dataset, clusters 4 and 8 have the least levels of uncertainty. Since cluster 8 generally represents the control mice, we can conclude that microglial states are more stable in the absence of injury and disease. Cluster 4 represents the same age as cluster 8 but with treatment which suggests that micro glial cells for mice near this age tend to have a similar expression pattern. In the GLP1-RA dataset, clusters 2 and 4 have the least level of uncertainty. These clusters represent a mix of ages and treatment which might similarly suggest that the microglial states with these expression patterns are stable.

While the methods we used, including but not limited to Gaussian Mixture Model (GMM), Hierarchical, and K-Means clustering, allowed us to cluster the GLP-1RA and Hammond data in multiple different manners and describe gene expression for each found cluster, there are absolutely further avenues to take this project further. Some limitations associated with these methods is that they provide a lot of high-level clustering analysis and information but without looking for specific results like age-related clustering tend to lead away from more specific findings. If someone were to continue the project one lead to explore would be finding the marker genes for each cluster, specific to that cluster only, for all of the clustering methods we used. This would have allowed us to further confirm the Hammond clustering methods. In addition to this route of further confirming the Hammond clustering methods, further applications of this process would undergo the same clustering methods and instead of

using PCA for dimensionality reduction use ICA. While the Hammond paper used ICA, two rounds of it, to visualize their data, and we would have found it valuable to do the same, issues with the *fastICA* R package prevented us from being able to do the same on both the Hammond dataset and the GLP-1RA dataset.

Overall, we found that our clustering of the Hammond data and new data from GLP-1RA data has provided evidence that the Hammond clustering and their results are largely valid. Microglial signatures and states are highly continuous and it is difficult to deduce that these states are the only ones present in microglial cells from this data. However, since our clustering analysis for the Hammond data as well as for a second data set is similar to the clustering done in the Hammond paper, we can confirm that the conclusions drawn from the Hammond paper and from the GLP-1RA paper hold.

5. References

Broad Institute Single Cell Portal. GLP-1R Brain Aging Reversal. 2024, Single Cell Portal, https://singlecell.broadinstitute.org/single_cell/study/SCP1182/glp1ra-brain-aging-reversal#study-download

Hammond, Timothy R et al. "Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes." Immunity vol. 50,1 (2019): 253-271.e6. doi:10.1016/j.immuni.2018.11.00

Jamail, I., Moussa, A. "Current State-of-the-Art of Clustering Methods for Gene Expression Data with RNA-Seq." Applications of Pattern Recognition, IntechOpen, 2020, pp. 1-12, doi:10.5772/intechopen.94069.

Li, Z., Chen, X., Vong, J.S.L. et al. Systemic GLP-1R agonist treatment reverses mouse glial and neurovascular cell transcriptomic aging signatures in a genome-wide manner. Commun. Bio 4, 656 (2021). https://doi.org/10.1038/s42003-021-02208-9

Olah, M., Menon, V., Habib, N. et al. Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. Nat Commun 11, 6129 (2020). https://doi.org/10.1038/s41467-020-19737-2

R Core Team. 2024, R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria.

Saadeh, Heba, Reem Q. Al Fayez, and Basima Elshqeirat. "Application of K-Means Clustering to Identify Similar Gene Expression Patterns during Erythroid Development." International Journal of Machine Learning and Computing, vol. 10, no. 3, May 2020, pp. 327-332, doi:10.18178/ijmlc.2020.10.3.956. <https://www.R-project.org/>.

Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, Goeva A, Nemesh J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll SA. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell. 2018 Aug 9;174(4):1015-1030.e16. doi: 10.1016/j.cell.2018.07.028. PMID: 30096299; PMCID: PMC6447408.

Scrucca L, Fraley C, Murphy TB, Raftery AE (2023). Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman and Hall/CRC. ISBN 978-1032234953, doi:10.1201/9781003277965, https://mclust-org.github.io/book/.

Zou, J., Chen, Z., Wei, X. et al. Cystatin C as a potential therapeutic mediator against Parkinson's disease via VEGF-induced angiogenesis and enhanced neuronal autophagy in neurovascular units. Cell Death Dis 8, e2854 (2017). https://doi.org/10.1038/cddis.2017.240

6. Acknowledgements

We would like to acknowledge the sacrifice of the mice who gave their microglial cells to the Hammond and GLP-1RA papers on which this project was based. We recognize their contribution to our computational and biological understanding and regret that their lives were lost in the process of furthering this important scientific work.