Analysis of single-cell RNA-seq data of two APOE genotypes in Alzheimers Disease

Zofia Kochaska

June 26, 2022

Abstract

Alzheimers Disease (AD) is the most common cause of dementia and is classified as a progressive neurodegenerative condition. It touches mostly people older than 65 years and is known to appear more commonly amongst women. One of the alleles of the APOE gene, coding apolipoprotein E has been dubbed the most common genetic risk factor for developing the disease. There are 3 versions of this gene in humans and having one of them can increase the probability of developing AD even 15 times. While the link between the genotype and the disease is known, the mechanism behind the increased risk of the disease still needs to be better studied. Here, I am trying to recreate the research made by Belonwu et. al.[1] on the single-nucleus RNA sequencing dataset created and shared by Grubman et. al.[2] The goal of those analyses is to identify differentially expressed genes (DEGs) among brain cell types between control and test groups and to check if they are connected with APOE genotypes. This might hopefully shed some light on the connection between having a certain genotype of a gene and being diagnosed with this incurable disease.

Introduction

Alzheimers Disease (AD) is the most common type of cognitive disorder.[3] Its defined as a progressive neurodegenerative condition, which means that symptoms appear gradually over a long period. This fact makes it hard to recognize early indicators, such as misplacing of items, as signs of imminent disease.[4] Some trends lead us to believe that AD cases will triple in frequency by 2050[5] which makes it even more important to study its underlying processes and risk factors. More than 90% of patients are older than 65 and as such this age is considered to be a boundary between early and late onset[6]of the disease. Not only is the probability of an appearance of AD higher in the later stages of life, but it is also more probable to appear in females,[7] as more than 2/3 of diagnosed cases of AD happen to be found in that sex. Another demographic risk factor is low education level[8] which is controversial because it has a completely different effect on the pace of progression of the disease depending on its stage.

There is no cure for AD,[9] but there are few known treatment options to help with its symptoms. One of the hardships while looking for therapy for the underlying pathology is the early emergence of related pathologic changes while the symptoms are still not visible. When it is possible to diagnose the disease, the alterations are so progressed that as of right now it is impossible to reverse them. This leaves those suffering from the disease with no hope of recovery and only an option to help with the symptoms. It is crucial to better understand the processes behind the disease and to try to characterize the genes which increase the probability of developing the disease. A better understanding of those factors can hopefully allow us to recognize the disease earlier before

it causes irreparable changes in the brain and make it possible for us to introduce some treatment options to help slow or even stop its progress.

Alzheimers disease is said to be a cause behind 60-80% of dementia cases,[10] where dementia is a general term for symptoms like memory loss and thinking difficulties. AD is linked with the formation of amyloid-β (Aβ) plaques and their deposition in parts of a brain.[11] They are the main characteristic of Alzheimers and as such are often used to diagnose the disease. In the brain of the person suffering from AD, it is possible to notice abnormal aggregations of this protein which form plaques that collect between neurons and disrupt their normal function. Amyloid-β is a result of the breakage of the larger, amyloid precursor protein.

One of the known genetic factors which increases the risk of the late onset of AD is a variant of the apolipoprotein E (APOE) gene[12]placed on chromosome 19. This gene is involved in the production of the protein of the same name, which combines with lipids and forms lipoproteins. They are responsible for metabolism and transporting cholesterol and other lipids in the bloodstream and help to maintain their correct levels. There are 3 known alleles of the APOE gene, which have been linked to the risk of developing AD. All of those forms differ from each other based on the substations of the amino acid residues numbered 112 and 158. The most common allele $\varepsilon 3$ is found in more than half of the population and is said to have no bearing on the probability of developing the disease. The rarest allele of APOE is $\varepsilon 2$, which is found in 8.4% of the general Caucasian population and 3.9% of those suffering from the disease.[13] Researchers believe that the presence of this form of APOE provides some protection against the disease. Most of the AD cases, when people have this allele occur in the later stages of life. The most dangerous form-$\varepsilon 4$ is the largest known genetic factor for developing AD. Observed frequencies of those alleles vary in different populations and even after accounting for those differences, it is impossible to always link the presence of one form of APOE with the higher risk of AD. One example of this phenomenon is the Nigerian population, where the AD is rare and the frequency of $\varepsilon 4$ is among the highest observed.[14] Thanks to those types of observations it is safe to say that APOE4 is not a determinant of disease, but simply increases its risk. It is believed that APOE4 homozygotes have a 12-15 times higher probability of developing the disease than APOE3 homozygotes.[15]

In this project, I decided to analyze samples of single-nucleus RNA-seq (snRNA-seq) applied to the entorhinal cortex of 12 human individuals with varying genotypes of APOE. I reduced the dataset to consist only of the samples having 3/3 (APOE $\varepsilon 3$ homozygotes) or 3/4 (heterozygotes $\varepsilon 3/\varepsilon 4$) APOE genotype. Im trying to identify differentially expressed genes (DEGs) among different cell types and APOE genotypes and compare my results with those obtained by Belonwu et. al.[1] who conducted a similar analysis on the partially overlapping dataset. The goal is to look for differences in the expression of genes across different cell types based on APOE genetic variants and verify if the results obtained by authors conducting similar research are reproducible.

Materials

All of the samples used in this study were acquired from a repository made by the researchers (Grubman et. al.[2]) who sequenced the samples taken from the entorhinal cortex (`http://adsn.ddnetbio.com/`). This site stores data which was already pre-processed while the raw snRNA-seq dataset is publicly available at the Gene Expression Omnibus with the accession number GSE138852. The filtered count matrix provided by the authors consists of 10,850 genes and

13,214 cells sampled from the entorhinal cortex. Originally there were 16 samples where 8 came from Alzheimers disease patients and the rest were used as a control group, as they had no history of AD or cognitive impairment, but 4 of those samples have distinctly different distributions of cell types and as such were excluded from the analysis. This preliminary reduction left 12 samples (Table 1) to be analyzed, which came from the age-matched individuals (mean age 77.6, range 67.391 years). Even after the scaling down of the dataset (exclusion of 2 samples from test and control groups), the categories were sex-matched, where two out of 6 samples in each group were collected from females. The samples were processed in 6 10x libraries where each one of them contained 2 individuals who were matched by the diagnosis and gender. Information about their APOE genotypes was available for all of them and allowed to subtract three more samples from the analysis based on the lack of other samples sharing the same APOE genotype, due to the relative rarity of the alleles.

ID	Batch	Gender	Age	APOE	Amyloid pathology
AD1	AD1_AD2	Male	91	3/4	Numerous diffuse and neuritic Aβ plaque
AD2	AD1_AD2	Male	83,8	3/4	Numerous diffuse and neuritic Aβ plaque
AD3	AD3_AD4	Female	67,8	4/4	Numerous diffuse and neuritic Aβ plaque
AD4	AD3_AD4	Female	83	3/4	Numerous diffuse and neuritic Aβ plaque
AD5	AD5_AD6	Male	73	4/4	Numerous diffuse and neuritic Aβ plaque
AD6	AD5_AD6	Male	74,6	3/4	Numerous diffuse and neuritic Aβ plaque
Ct1	Ct1_Ct2	Female	67,3	3/3	Occasional diffuse plaque in cortex
Ct2	Ct1_Ct2	Female	82,7	3/3	None
Ct3	Ct3_Ct4	Male	72,6	3/3	None
Ct4	Ct3_Ct4	Male	75,6	3/4	None
Ct5	Ct5_Ct6	Male	77,5	3/3	None
Ct6	Ct5_Ct6	Male	82,7	2/4	None

Table 1: Table showing original 12 samples and information about them.

Only APOE genotypes 3/3 and 3/4 were present in control and test groups and as such during the downstream analysis only they were considered. This downsizing of a dataset left 9 samples to analyze out of which 4 shared 3/4 genotype and the rest had 3/3 APOE genotype (Table 2). For all of those samples original researchers, who sequenced the data, provided information about their amyloid pathology. They assigned samples to 3 categories which characterized the number and type of Aβ plaques[16] found in the brain of the deceased during autopsy. They differentiated between two types of Aβ plaques- diffuse (DPs) and neuritic (NPs) where the first ones lack dystrophic neurites, while the second group is more centered and has them. All of the samples in the control group had no signs of any of those changes. Based on those descriptions Belowu et. al.[1] added Consortium to Establish a Registry for Alzheimers Disease (CERAD) scores[17] to the dataset. This number describes the risk of AD based on Aβ plaques where a score higher than 3 means no AD, and a score equal to 1 means definite AD.

ID	Batch	Gender	Age	APOE
AD1	AD1_AD2	Male	91	3/4
AD2	AD1_AD2	Male	83,8	3/4
AD4	AD3_AD4	Female	83	3/4
AD6	AD5_AD6	Male	74,6	3/4
Ct1	Ct1_Ct2	Female	67,3	3/3
Ct2	Ct1_Ct2	Female	82,7	3/3
Ct3	Ct3_Ct4	Male	72,6	3/3
Ct4	Ct3_Ct4	Male	75,6	3/4
Ct5	Ct5_Ct6	Male	77,5	3/3

Table 2: Table showing samples chosen for downstream analysis.

Methods

All of the following analyses were done using R[18] version 4.0.4, RStudio[19] and Seurat[20] version 4.1.1. While doing them I tried to recreate the work done by Belonwu et. al.[1] as closely as possible while still correcting some parts of the research that were worrisome. Most of the visualizations showed in this work were created using dittoSeq,[21] ggplot2,[22] pheatmap,[23] and UpSetR.[24]

The following steps were done on the pre-filtered file provided by Grubman et. al.[2] in their repository. The expression matrix, which was used, consisted of 10,850 genes and 13,214 cells which were sequenced from 12 samples. From those cells, I kept only those that came from patients with APOE 3/3 or 3/4 genotypes and created a Seurat object based on genes that were found in at least 3 cells, and cells that had at least 200 genes. In their paper, Belonwu et. al. wrote that they had done the same steps, but after analyzing their code I am led to believe that instead of using only cells from patients with genotypes that they were analyzing, they used the cells from every genotype. Later on, they choose only the genes which were differentially expressed in those two groups but in my opinion, this may cause skewing of the results because they normalized the object containing all genotypes, and as such, I decided to follow the outline from the article.

The Seurat object, created this way was later on normalized using SCTransform and integration workflow as the authors of the package claim that it is a more accurate method of normalizing, estimating the variance, and identifying the most variable genes. Those operations help to account for the batch effects introduced by the design of sequencing. Those characteristics are shown in Table 2 and one of them is processing males and females separately and grouping samples based on the diagnosis.

The next step was the assignment of cell types. Matrix downloaded from Grubman et. al.[2] already has information on cell types assigned by them, but I chose to follow the steps outlined by Belonwu et. al.[1] and assign new categories. Using BRETIGEA package,[25] which stores information about markers characterizing specific brain cell types, I calculated the score for each cell type based on markers present in every cell. Only 200 gene markers from BRETIGEA for each of the 6 cell types (astrocytes, neurons, microglia, oligodendrocytes, oligodendrocyte progenitor cells (OPCs), and endothelial cells) were used to assign identities. Later the two cell identifiers with the highest scores were chosen for each cell and based on them identities were assigned. If

the highest and the second highest score absolute values were within 20% of each other cell type was assigned as a hybrid. The next step was to check the separation and homogeneity of created clusters in Uniform Manifold Approximation and Projection (UMAP) plots. As it was done in the article written by Belowu et. al., I excluded endothelial cells from the following analysis based on their small number in comparison to other cell type groups.

The next part of the research focused on looking for differentially expressed genes (DEGs) using Limma package [26] and Voom transformation. In the model used by Belonwu et. al.,[1] only sex and APOE genotypes of the assigned cell types were included, as adding batch as a covariant in the design led to collinearity. Adding age to the used design was not needed because all of the used samples were already age-matched. The next step was to filter out lowly expressed genes in the DEGList object using the filterByExpr function from edgeR package[27] with default parameters. This function disposes of genes that have counts of insufficient value to be retained in statistical analysis. In the next step normalization factors were calculated using the Trimmed Mean of M-values with the singleton pairing (TMMwsp) method. Next, the voom function was applied which estimates the mean-variance relationship of the log counts. After this, the model was fitted with a contrast matrix for every cell-type APOE genotype between the test and control groups, and then, using empirical Bayes, standard errors were moderated. For every cell-type and APOE genotype group, using the prepared model, I looked for DEGs between the control and test groups. Only those genes which had Benjamini-Hochberg (BH) corrected p-value lower than 0.05 and the absolute value of log-fold change higher than 0.25, were saved as DEGs. All of them were then written to two separate files depending on the considered APOE genotype and later on visualized using a variety of plots.

Results

Assignment of cell types using the BRETIGEA package yielded somewhat different results than those obtained by Grubman et. al.[2] and included in the analyzed data. Those differences are mostly results of using different categories, where researchers who originally obtained the data used categories doublet and unId which are not present in my analyses. The number of cells assigned to each cell type is shown in Table 3, while Figure 1 shows differences between percentages of cell types assigned to samples between those provided by Grubman et. al.[2] (Fiqure 1a) and those calculated using the method outlined by Belonwu et. al.[1] (Fiqure 1b). It is noticeable that the results of those two assignments are similar and the main differences are due to the fact that some of the categories are exclusive to one method. In both cases, oligodendrocytes are the most common cell type with astrocytes in second place while there are only a few cells assigned as endothelial.

Astrocytes	Endothelial cells	Hybrid	Microglia	Neurons	Oligodendrocytes	OPCs
2095	60	345	392	688	5519	878

Table 3: Table showing number of cells assigned to each cell type.

Using limma-voom to look for differentially expressed genes allowed to find 985 DEGs for APOE 3/3 and APOE 3/4. Table 4 shows the number of up and down-regulated genes for each cell type and APOE genotype. It is easily noticeable that for APOE 3/3 genotype there are many more

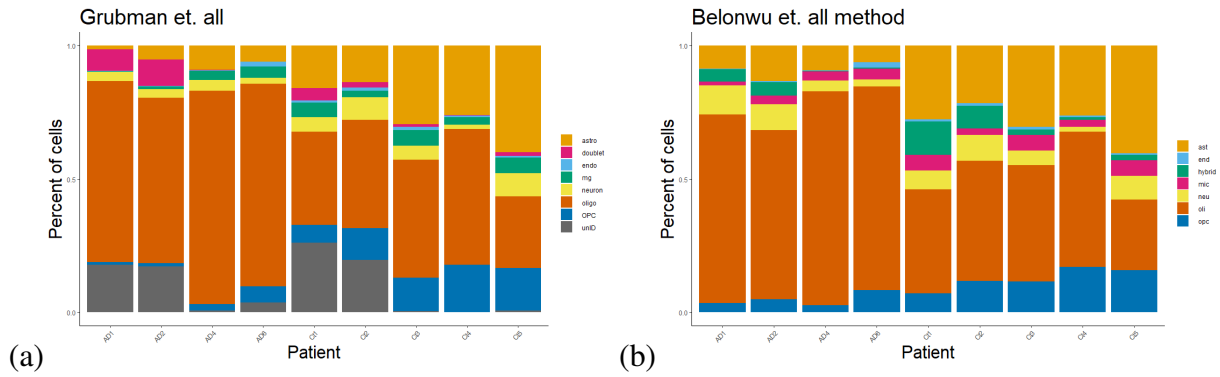

Figure 1: Cell assignments amongst samples.

genes that are downregulated (almost 4.5 times more), while the upregulated are more prevalent (over 3 times more) among samples sharing the APOE 3/4 genotype. The same trends were observed by Belonwu et. al.[1] during their analysis of the samples which came from the entorhinal cortex, but while in my analysis only 175 unique DEGs were found, researchers identified 232 of them.

	Astrocytes		Microglia		Neurons		Oligodendrocytes		OPCs	
APOE	3/3	3/4	3/3	3/4	3/3	3/4	3/3	3/4	3/3	3/4
Down	85	41	43	15	91	33	103	49	48	53
Up	20	123	18	135	10	97	20	112	15	113

Table 4: Table showing number up and down regulated DEGs assigned to each cell type.

Figure 2 shows DEGs specific and shared between cell types, subplot (a) focuses on APOE 3/3 genotype, while (b) depicts genes found in 3/4 genotype. It is clear to see that number of DEGs across cell types in APOE 3/4 is more evenly distributed, while in APOE 3/3 there are many more DEGs in oligodendrocytes and neurons than in OPCs. In both genotypes number of DEGs shared amongst all cell types is higher than in any other grouping of cell types but in APOE 3/4 it is almost 4 times more than the second most popular combination of cell types (all except neurons). In the 3/3 genotype, there are some DEGs unique to every cell type while in APOE 3/4 there is no DEG present in only astrocytes, neurons, or OPCs.

A heatmap of expressions of all genes which are clustered by APOE genotype and cell type is shown in Figure 3. It is easy to notice differences between groups, especially based on APOE genotype, while the boundaries of the cell type categories are a little bit more blurred. This is a natural extension of the trends shown in Table 4, where there is a significant change between the number of up and down-regulated genes between genotypes.

To focus on the genes which showed the highest change across both of the analyzed genotypes, I created a list with the 10 genes which were present in DEGs found for both genotypes and had the highest absolute value of log fold change. Those genes are shown in a heatmap (Figure 4) depicting their expressions, which are grouped by APOE genotype and cell type. Genes *CRYAB*, *HSPA1A* and *LINC00486* are downregulated in APOE 3/3 while the rest of the 10 genes are being upregulated in most of the cell types of 3/3. A completely inverted situation can be seen in APOE

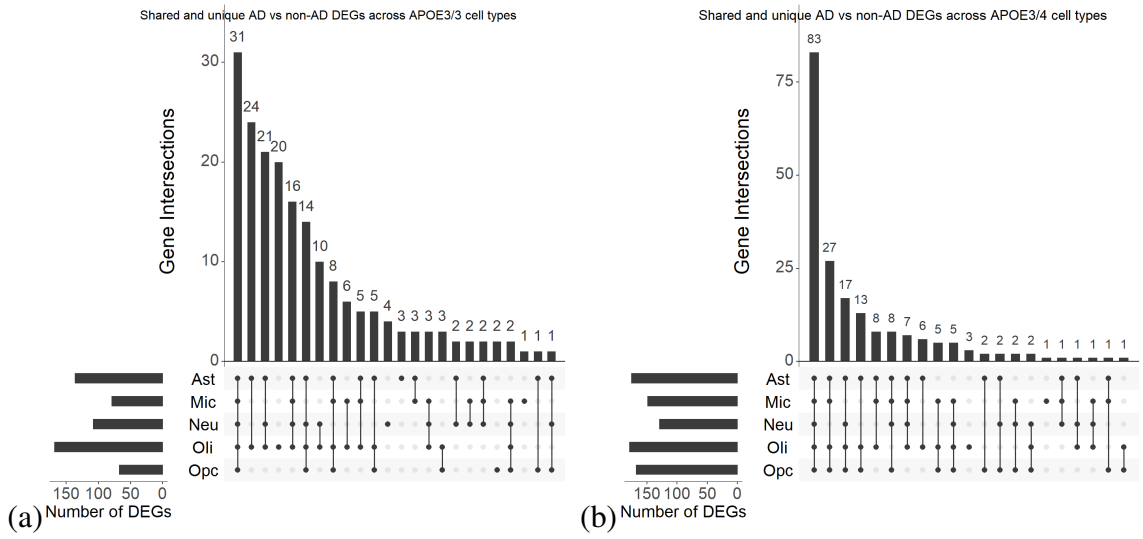

Figure 2: Cell assignments amongst samples.

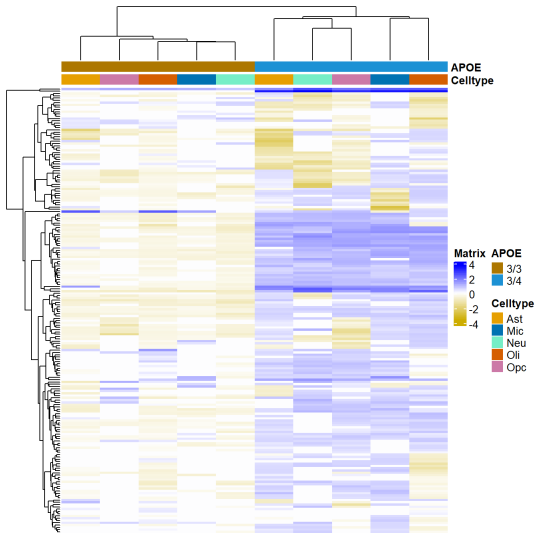

Figure 3: Heatmap of all genes grouped by APOE genotype and cell type.

3/4 genotype. Of course, there are some cell types where the genes are showing a different type of expression than in the rest, but those are the observed trends. In APOE 3/3 most of the genes with the biggest changes in expressions seem to be found in microglia and neurons while in the 3/4 genotype astrocytes seem to have the most expressed genes in comparison to the control group.

For 5 genes, which had the highest absolute value of log fold change, violin plots showing their expressions were created. They are shown in Figure 5 where each subplot adheres to the same color scheme so the legend is shown for only one of them. In the first subplot, which shows *DOCK4* expression across genotypes, the most visible difference can be seen in its expression in microglia, which is heightened in both genotypes in the control group and AD with APOE 3/3. *LINC00486* and *FRMD4A* seem to be evenly distributed among all cell types with not many outliers. In *NEAT1* its expression in astrocytes in AD with APOE 3/3 genotype seems to be centered in the

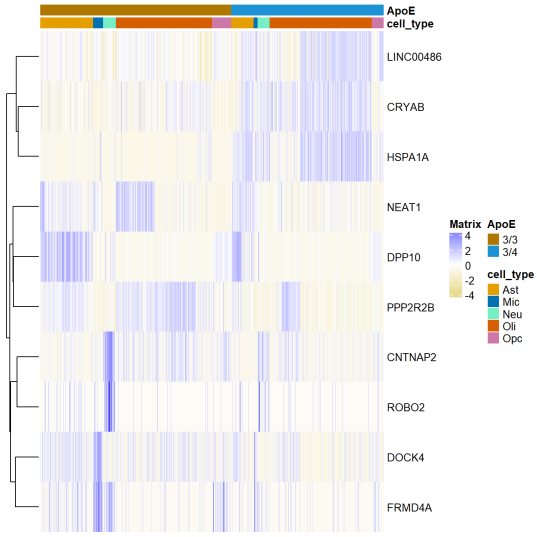

Figure 4: Heatmap of 10 genes with highest absolute value of log fold change grouped by APOE genotype and cell type.

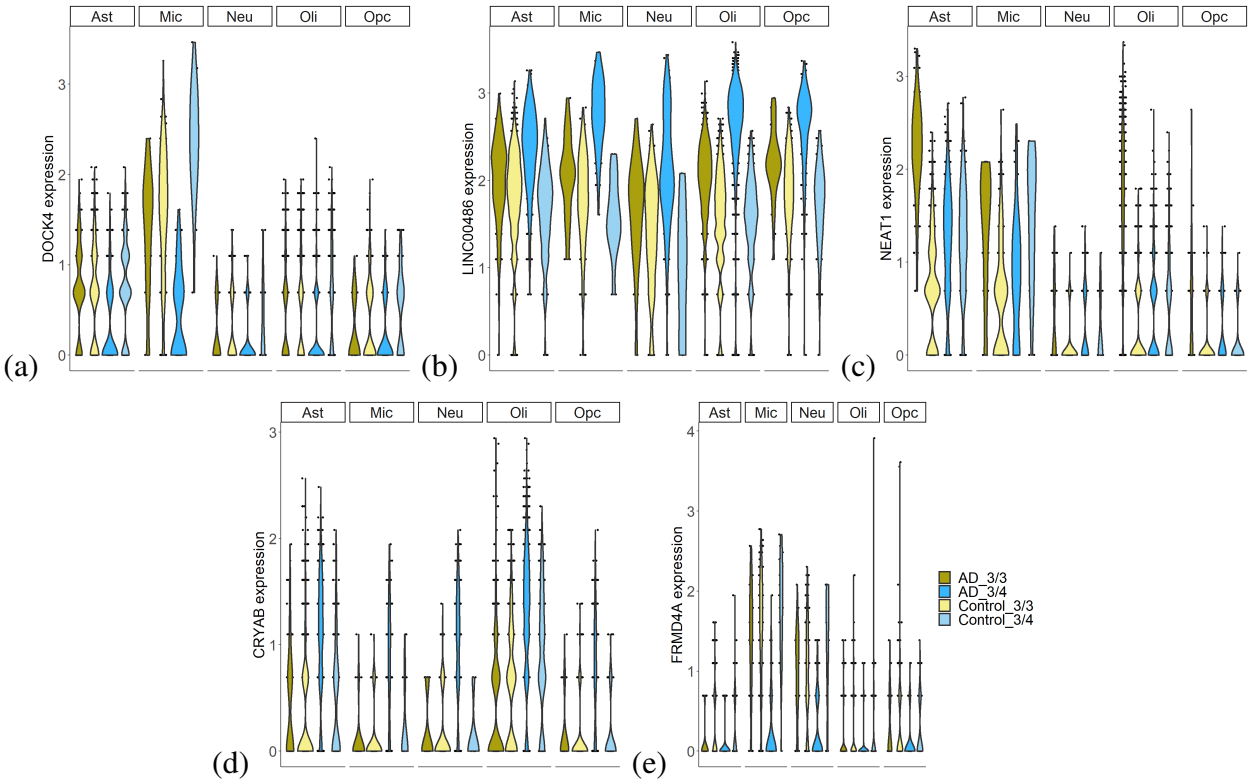

Figure 5: Violin plots for 5 genes with highest absolute value of log fold change.

higher range than in other groups, while *CRYAB* seems to be overly expressed in oligodendrocytes.

Figure 6 shows genes and their log fold changes and adjusted p-values across different cell types. The y-axis of the plots shows values that quantify changes between expressions in AD

patients having APOE 3/4 genotype and control groups. The x-axis shows a comparison between genes in APOE 3/4 and non-AD patients, while the colors define adjusted p-values. The results showed in those plots are similar to those obtained by Belonwu et. al.[1] where the most noticeable difference is in *LINGO1* gene which in my results was placed only in those DEGs found for APOE 3/4 while authors identified it as DEG in both genotypes. What is also noticeable is the placement of genes in each cell type where it is seen that DEGs found in OPCs are more centered which means that they have lower values of log fold change between APOE 3/3 and non-AD as well as APOE3/4 and non-AD.

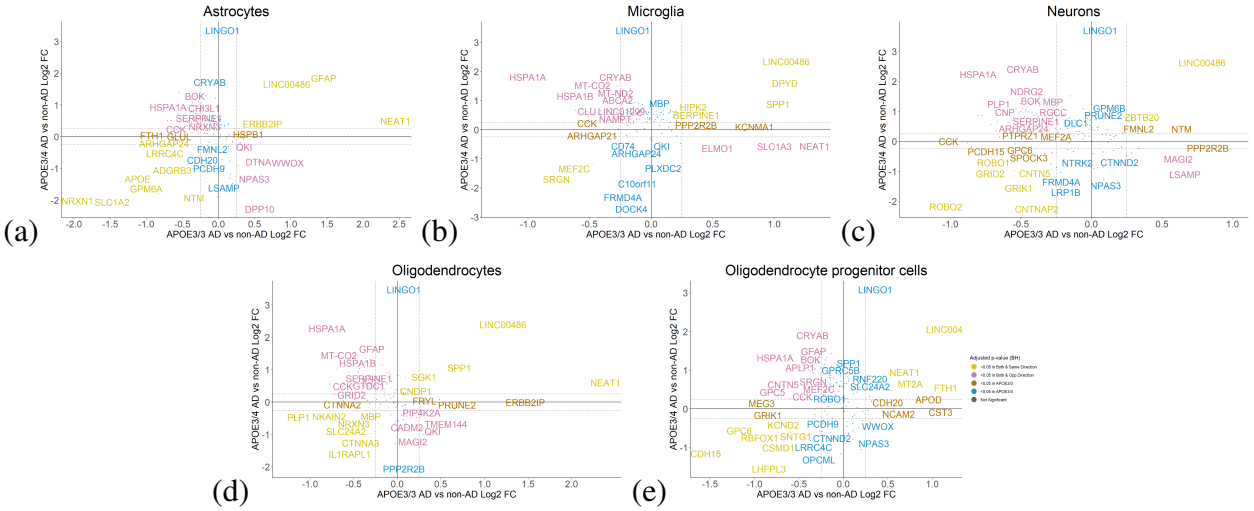

Figure 6: Log fold change and adjusted p-values for genes found in different cell types.

Discussion

Analyses described in this work show DEGs found across different APOE genotypes and cell types. They were done following the workflow designed by Belonwu et. al.[1] and in some cases yielded similar results. Small change- earlier exclusion of the rest of the APOE genotypes, managed to completely change some of the results. For example, it altered the list of genes with the highest absolute log fold change value and changed mutual DEGs found in both of the analyzed genotypes. This shows that attention to detail is crucial in this line of work, where even a small mistake can completely modify obtained results and potentially lead to wrong conclusions.

Most of the DEGs in the APOE 3/4 genotype were found in oligodendrocytes and astrocytes which can point to their involvement in the disease. What is worth noting is the percentage of those cell types in the dataset which is significant and as such can skew the results as the number of analyzed genes for those cells might have been higher. On the other hand, genes which had the highest absolute value of log fold change were mostly prevalent in astrocytes, neurons, and microglia which are amongst the smallest cell types categories.

Results, where the number of DEGs in different APOE genotypes and AD diagnosis were considered are in my opinion not very informative as there were many more samples with APOE 3/3 genotype in the control group than in the test group. Another part of the design that can

potentially change the results is the different number of cells in each cell type category. Those inequalities may skew the results as it is possible that the used samples did not show the trends in the disease. I believe that the size of the used dataset was not ample enough to capture all the changes in gene expressions in the population and may lead to wrong conclusions based on consideration of changes that are occurring only in some of the people suffering from AD. Keeping this in mind, analyses done by Belonwu et. al.[1] are better planned as the dataset used here was only part of the data that they were analyzing. This can hopefully deal with the changes which were only introduced because of the dataset size and make their results more believable.

I cannot in all certainty say if researchers planned to use other genotypes in the analyses and simply forgot to mention this in their article or if the mistake was done when they omitted the exclusion of other samples. I believe that those other APOE genotypes should have been excluded earlier, before the normalization of the Seurat object, and chose to do it in my analyses. Changes, caused by these modifications are in the details of gene expression and not in the observed trends of the results. As such, they may not completely invalidate the knowledge glimpsed from their results, where the general tendencies were considered, but the detailed results about gene expression values are not, in my opinion, accurate.

Data and Code Availability

All of the raw samples are available at Gene Expression Omnibus under the accession number GSE138852. It is also possible to download already pre-processed data from the website provided by the original researchers (Grubman et. al.[2]) who sequenced them, where they also publish results from their analyses (`http://adsn.ddnetbio.com/`). The additional file, which contains the information about each sample APOE genotype, is available on Github at `https://github.com/zofiakk/MoCBS/tree/main/final_project/data` in the supp _data.csv file.

To allow easier recreation of all of the analysis and figures R code written for those tasks is available on Github at `https://github.com/zofiakk/MoCBS/tree/main/final_project`, where an HTML file showing all steps and plots is also placed. In the same repository, there is a renv.lock file which stores all of the information about necessary packages and their versions used to process the data. It is possible to load them using the commands for renv package[28] (renv::restore()). While running this code it might be necessary to increase the memory limit in RStudio, which can be checked and then changed with memory.limit(size=new_value) command.

References

[1] Belonwu, S. A. *et al.* Genotype-Specific Changes Across Cell Types in Two Brain Regions. *Front Aging Neurosci* **14**, 749991 (2022).

[2] Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat Neurosci* **22**, 2087–2097 (2019).

[3] Oboudiyat, C., Glazer, H., Seifan, A., Greer, C. & Isaacson, R. S. Alzheimer's disease. *Semin Neurol* **33**, 313–329 (2013).

[4] Lyketsos, C. G. *et al.* Neuropsychiatric symptoms in Alzheimer's disease. *Alzheimers Dement* **7**, 532–539 (2011).

[5] Hebert, L. E., Weuve, J., Scherr, P. A. & Evans, D. A. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology* **80**, 1778–1783 (2013).

[6] Kumar, A., Sidhu, J., Amandeep Goyal, A. & Tsao, J. W. *Alzheimer Disease* (StatPearls Publishing, 2021).

[7] Ferretti, M. T. *et al.* Sex differences in Alzheimer disease - the gateway to precision medicine. *Nat Rev Neurol* **14**, 457–469 (2018).

[8] Kim, K. W. *et al.* Disease progression modeling of Alzheimer's disease according to education level. *Sci Rep* **10**, 16808 (2020).

[9] Khan, S., Barve, K. H. & Kumar, M. S. Recent Advancements in Pathogenesis, Diagnostics and Treatment of Alzheimer's Disease. *Curr Neuropharmacol* **18**, 1106–1125 (2020).

[10] Scheltens, P. *et al.* Alzheimer's disease. *Lancet* **397**, 1577–1590 (2021).

[11] Zhang, H. & Zheng, Y. [Amyloid Hypothesis in Alzheimer's Disease:Pathogenesis,Prevention,and Management]. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* **41**, 702–708 (2019).

[12] Serrano-Pozo, A., Das, S. & Hyman, B. T. APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches. *Lancet Neurol* **20**, 68–80 (2021).

[13] Farrer, L. A. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* **278**, 1349–1356 (1997).

[14] Sepehrnia, B. *et al.* Genetic studies of human apolipoproteins. X. The effect of the apolipoprotein E polymorphism on quantitative levels of lipoproteins in Nigerian blacks. *Am J Hum Genet* **45**, 586–591 (1989).

[15] Long, J. M. & Holtzman, D. M. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell* **179**, 312–339 (2019).

[16] Malek-Ahmadi, M., Perez, S. E., Chen, K. & Mufson, E. J. Neuritic and Diffuse Plaque Associations with Memory in Non-Cognitively Impaired Elderly. *J Alzheimers Dis* **53**, 1641–1652 (2016).

[17] Fillenbaum, G. G. *et al.* Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimers Dement* **4**, 96–109 (2008).

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2020). URL `https://www.R-project.org/`.

[19] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA (2019). URL `http://www.rstudio.com/`.

[20] Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902 (2019).

[21] Bunis, D. G., Andrews, J., Fragiadakis, G. K., Burt, T. D. & Sirota, M. dittoseq: universal user-friendly single-cell and bulk rna sequencing visualization toolkit. *Bioinformatics* (2020). Btaa1001.

[22] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016). URL `https://ggplot2.tidyverse.org`.

[23] Kolde, R. *pheatmap: Pretty Heatmaps* (2019). URL `https://CRAN.R-project.org/package=pheatmap`. R package version 1.0.12.

[24] Gehlenborg, N. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets* (2019). URL `https://CRAN.R-project.org/package=UpSetR`. R package version 1.4.0.

[25] McKenzie, A., Wang, M. & Zhang, B. *BRETIGEA: Brain Cell Type Specific Gene Expression Analysis* (2021). URL `https://CRAN.R-project.org/package=BRETIGEA`. R package version 1.0.3.

[26] Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).

[27] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

[28] Ushey, K. *renv: Project Environments* (2022). URL `https://CRAN.R-project.org/package=renv`. R package version 0.15.5.