Gene expression in brains of Alzheimer's and wild type mice

Team 6

Jonathan Joseph, Emily Kim, Annie Michel, Milan Solanki, Alex Spallone

Abstract

Alzheimer's disease is a neurodegenerative disease that results in amyloid plaques, neurofibrillary tangles, and damage to neurons in the brain (National Institute on Aging, 2023). Studies have revealed that Alzheimer's disease greatly affects specific regions of the brain, such as the hippocampus and cortex (Wenk, 2006), in contrast to other neurodegenerative diseases such as Parksinson's disease which severely affects the substantia nigra (Elbaz et al., 2016). However, more research is needed concerning the expression differences between regions of the brain in relation to Alzheimer's Disease (Lancour et al., 2020). We explore the gene expression differences between regions of the brain, specifically the hippocampus and cortex, in Alzheimer's Disease and wild-type mice. Here we show that gene signatures highly correlate with specific regions of the brain. We identified 214 genes that are differentially expressed between cortex and hippocampus samples. Unsupervised analysis techniques, specifically K-means, hierarchical clustering, and PAM, revealed that the samples separate most clearly among regions of the brain. Supervised analysis techniques, such as support vector machine, logistic regression, K-nearest neighbor, and naive bayes, revealed near perfect classification for brain regions using a varying number of genes from 10 to 10k. ClusterProfiler enrichment analysis found that many biological processes were enriched, often related to the regulation of neural components like neurotransmitters or neuron death. Our results demonstrate that gene expression is highly correlated with brain region, specifically the cortex and hippocampus, and provide insight into differentially expressed genes, as well as enriched biological processes. These results could be used for further research into expression differences between regions of the brain for Alzheimer's Disease and wild-type mice. For example, the expression differences between Alzeihmer's Disease and wild-type mice could be tested and compared to our results. This could provide insight into which processes are unique to specific regions of the brain for mice with and without Alzheimer's Disease. When considered in conjunction with samples taken at different time periods, this could provide further insight into the unique developmental processes that occur within regions of the brain for mice with and without Alzheimer's Disease.

Introduction

With the data, researchers analyzed 483 mice samples from either the hippocampus or the cortex focusing on thelgene expression between these two brain regions in the absence of the TREM2 gene. So our goal was to analyze the differential gene expression, which is essential to understand the molecular mechanisms underlying Alzheimer's disease. In terms of the approach, it followed a three step process: data exploration, clustering by brain section, and then predictive modeling (classifying expression testing data as one of two brain regions). During the data exploration phase, the distribution of the gene data was analyzed and differential gene expression was conducted. The next phase which was clusterin by brain section, which entailed an unsupervised analysis of the gene data, using methods such as K-means clustering, hierarchical clustering, and PAM clustering. These techniques help to observe unique patterns

or characteristics within the expression of both regions respectively. The final phase involved predictive modeling, which used a supervised analysis of the gene data, and methods such as logistic regression and Naive Bayes algorithm were utilized. The goal was to classify expression testing data as one of two brain regions: either the hippocampus or the brain cortex.

Methods

To explore gene expression between the hippocampuses and cortexes of Alzheimer's Disease and wild-type mice, we began with differential expression. Density and PCA plots were generated to view the distribution of the data, which consisted of 41,249 genes and 483 samples. A volcano plot showing $-\log_{10}P$ vs. $\log_2$ fold change was generated and genes passing certain thresholds were highlighted ($\log_2$ fold change > 1, $-\log_{10}P > 1$, and $p < 0.05$). Differential gene expression was done with the R package DESeq2 (Love et al., 2014), with expression data as input. Data was filtered to exclude genes with counts less than 500.

Clustering by brain section was performed in two ways: unsupervised and supervised. Unsupervised clustering was done with K-means clustering, hierarchical clustering, and PAM clustering, using the 5000 most variable genes and expression data as input. All unsupervised clustering was done with base R and R package cluster (Maechler et al., 2022). Elbow plots and alluvial diagrams were generated using the R packages factoextra (Kassambara and Mundt, 2020) and ggalluvial (Brunson, 2020). Elbow plots showed total within sum square vs. number of clusters (k). All unsupervised clustering algorithms required k-selection. Alluvial diagrams showed frequency of brain section vs. clustering setups with variable numbers of genes (10 genes, 100 genes, 1000 genes, 10000 genes). Supervised clustering was done with support vector machines, logistic regression, K-nearest neighbors, and Naïve Bayes, using the 5000 most variable genes and expression data as input. All supervised clustering was done with R packages tidyverse (Wickham et al., 2019), mlr3verse (Lang Schratz, 2023), and caret (Kuhn, 2008). Gene signatures were not collected because none of the supervised clustering algorithms perform automatic feature selection. Changes in the number of genes used in analysis (10 genes, 100 genes, 1000 genes, 10000 genes) and their effects were observed for clustering with both supervised and unsupervised learning. However, only unsupervised results were analyzed for statistical significance with the chi-squared test of independence. P-values and adjusted P-values with Bonferroni correction were observed. Heatmaps of gene expression were generated for both unsupervised and supervised clustering using the R package ComplexHeatmap (Gu, 2016) and DESeq2 results. For unsupervised clustering, annotations were used to compare unsupervised clustering results with actual brain part splits. For supervised clustering, gene expression across all genes included in gene signatures, which consisted of all genes in the DESeq2 results, were compared.

Gene set enrichment analyses were done using R packages topGO (Alexa and Rahnenfuhrer, 2021), clusterProfiler (Yu et al., 2012; Wu et al., 2021), gProfiler2 (Kolberg et al., 2020), and GenomicSuperSignature (Oh et al., 2022). Enriched Gene Ontology (GO) processes were extracted using a list of differentially expressed genes in DESeq2 results. P-values were adjusted using the Benjamini-Hochberg correction with a cutoff of 0.05. All methods and statistical analyses were performed in R version 4.3.1 (R Core Team, 2021). The dataset and all code are available at https://github.com/ehk-kim/CGS4144.

Results

From the differential expression analysis, we can see that the samples split into two groups, according to the PCA plot: cortex and hippocampus (Fig. 1). While the points are scattered across PC2, they remain within their PC1 groups. According to DESeq2 results, 214 genes were differentially expressed between the hippocampus and cortex ($\log_2$ fold change > 1, $p < 0.05$). These genes are represented in red in Figure 2.

Figure 1: PCA plot.

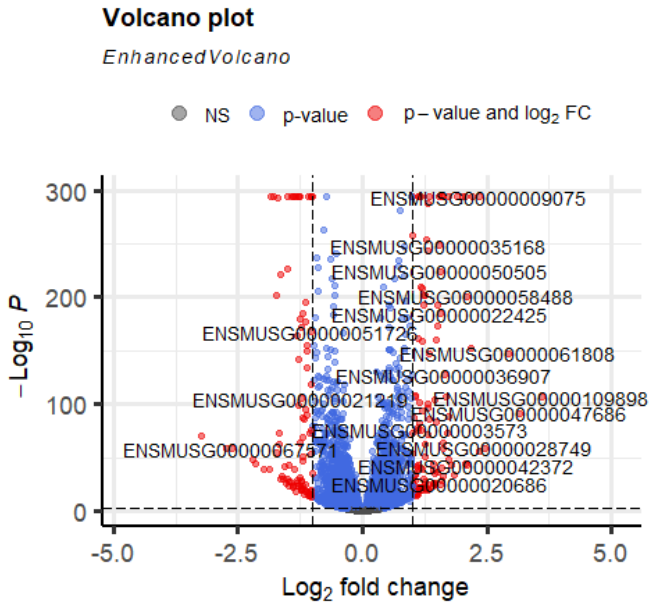

Figure 2: Volcano plot

Clustering results confirmed PCA plot results and showed that genes clustered by brain part. The elbow plots showed that the optimal number of clusters was two, in accordance with the two brain parts: hippocampus and cortex (Figure 3).

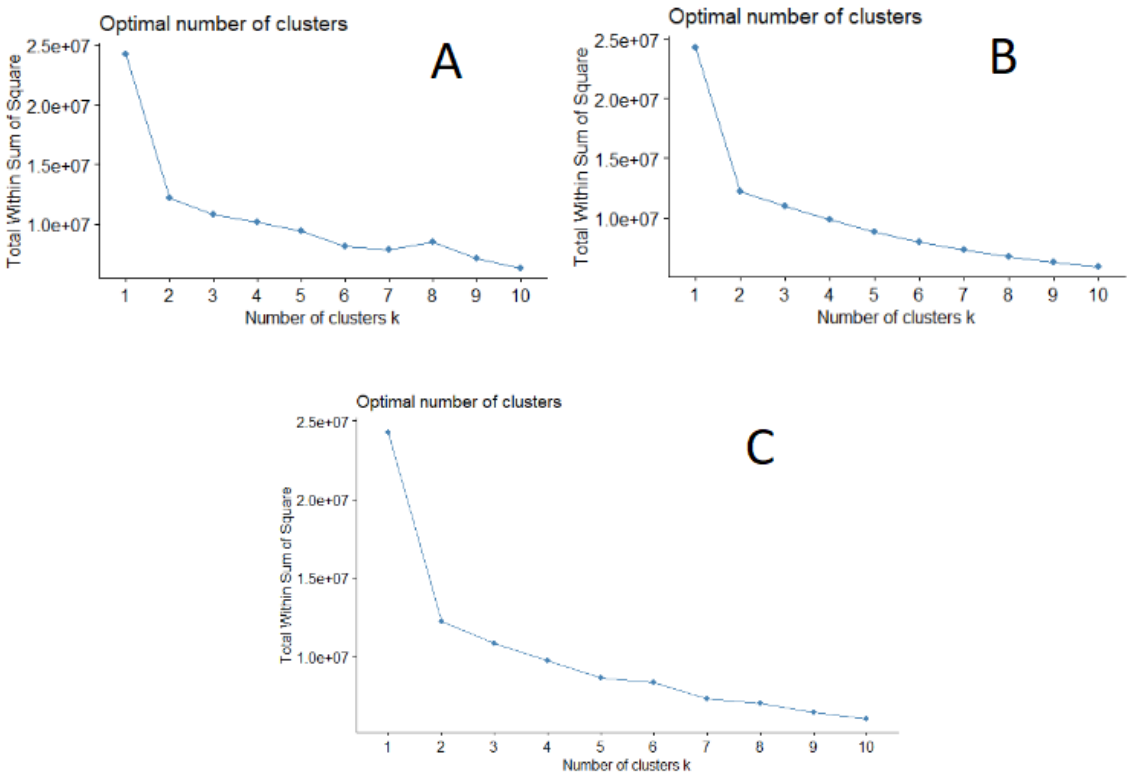

Figure 3: Elbow plots. Figure 3A is from K-means clustering, figure 3B is from hierarchical clustering, and figure 3C is from PAM clustering.

Alluvial diagram results showed cluster memberships for each of the different clustering setups: 10 genes used, 100 genes used, 1000 genes used, and 10000 genes used for analysis. While hierarchical and PAM clustering results showed relatively consistent clustering memberships throughout all clustering setups, K-means results constantly switched cluster membership between cluster 1 and cluster 2 (Figure 4). This was unexpected, as samples should cluster into one consistent group. However, despite the switch in membership, the samples consistently clustered into their respective brain parts.

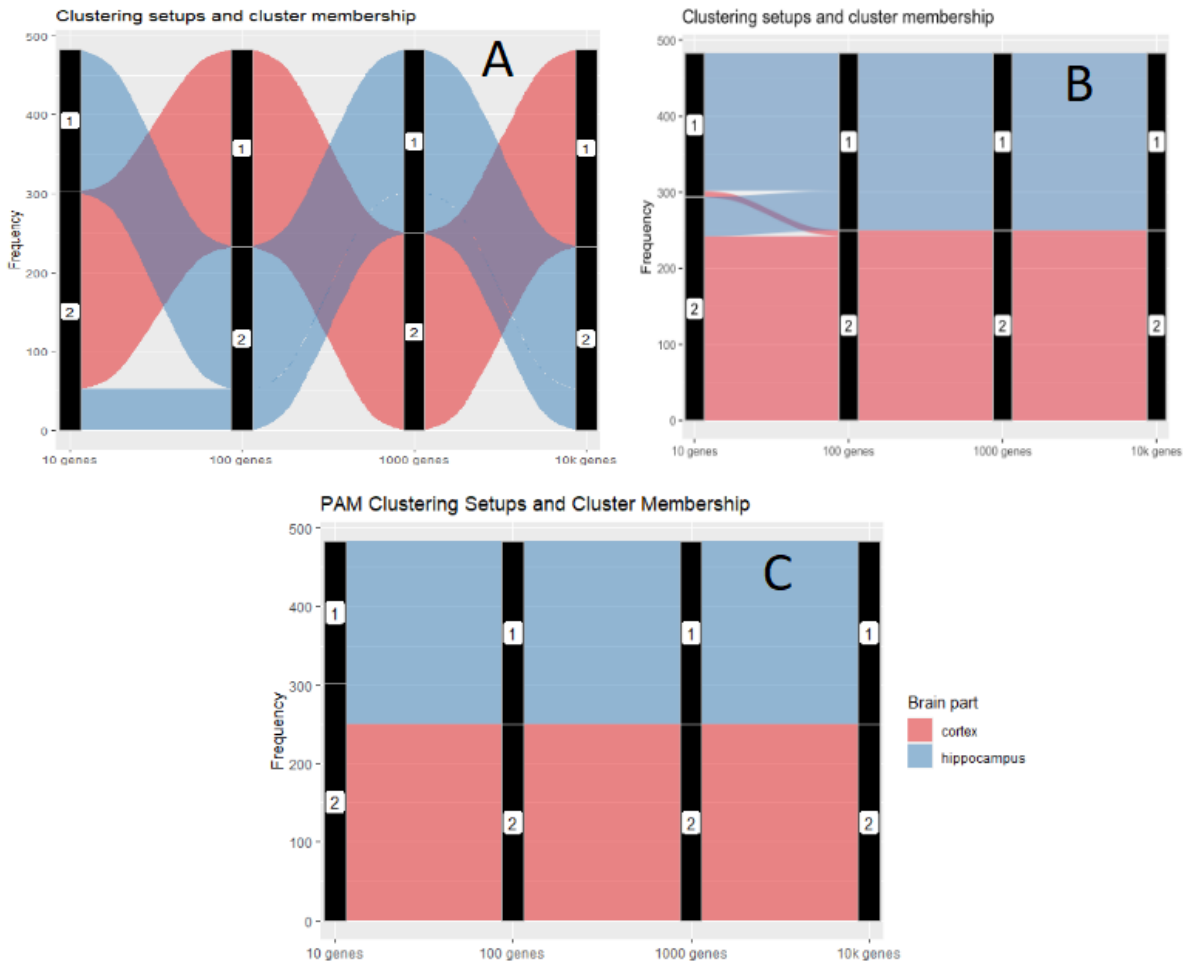

Figure 4: Alluvial diagrams. Figure 4A is from K-means clustering, figure 4B is from hierarchical clustering, and figure 4C is from PAM clustering

Results from the chi-squared test of independence showed extremely strong correlation between cluster membership and brain part. Even with ten genes used for analysis, P-values and adjusted P-values were well below 0.05. With 100, 1000, and 10000 genes used for analysis, P-values and adjusted P-values were the same (Table 1). While it was expected that the P-values and adjusted P-values would be low due to the clustering memberships, a degree of e-106 was unexpected. Because the raw P-values are so low, a Bonferroni correction does not change the P-value.

Number of genes	P-value (K-means)	Adjusted P-value (K-means)	P-value (hierarchical)	Adjusted P-value (hierarchical)	P-value (PAM)	Adjusted P-value (PAM)
10	8.48e-69	8.48e-69	2.29e-62	2.29e-62	8.48e-69	8.48e-69
100	3.52e-106	3.52e-106	3.52e-106	3.52e-106	3.52e-106	3.52e-106

1000	3.52e-106	3.52e-106	3.52e-106	3.52e-106	3.52e-106	3.52e-106
10000	3.52e-106	3.52e-106	3.52e-106	3.52e-106	3.52e-106	3.52e-106

Table 1: Unadjusted and adjusted P-values from chi-squared test of independence of cluster membership for K-means, hierarchical, and PAM clustering.

Supervised clustering results showed near perfect classification (AUC = 1), even when changing the training/testing data split and number of genes used. This indicates that gene expression is strongly correlated with regions of the brain, supporting the unsupervised clustering results.

While some enrichment analyses showed no significant enriched GO terms, clusterProfiler found that many GO terms related to biological processes were enriched compared to cellular component or molecular function GO terms. Of the enriched terms, many were related to regulation of neural components like neurotransmitters or neuron death. A few were related to fear response and cognition.

Overall, our results showed that gene expression is highly correlated with regions of the brain. Both supervised and unsupervised clustering showed that, using the 5000 most variable genes, samples clustered strongly with hippocampus or cortex. Additionally, results from the chi-squared test of independence showed that cluster membership is representative of brain region.

Conclusion

In conclusion, the hypothesis we sought out to investigate entailed whether gene expression patterns varied in the context of Alzheimer's Disease when factoring in different locations of the brain (in our case the hippocampus and cortex). Our results demonstrate that gene expression is highly correlated with regions of the brain in that both supervised and unsupervised clustering showed that, using the 5000 most variable genes, samples clustered strongly with hippocampus or cortex. We could have improved the analysis by not only fine tuning algorithm parameters of clustering algorithms to optimize performance and give more reliable clustering results, but also, we could have normalized the gene expression data to increase accuracy and comparability of results. This would help mitigate batch effects that transpire due to variations in sample processing. One form of additional analysis we would like to do would entail a more temporal analysis approach where we take samples taken at different time periods which could potentially demonstrate changes in gene expression over the course of Alzheimer's Disease progression. This could allow us to see the disease's response developmentally to time or interventions and potentially give us more clues in regards to gene expression in the hippocampus and cortex. Potentially critical time points of therapeutic interventions could be understood from this research to help patients suffering from the illness.

References

Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.

Alexa, A. and Rahnenfuhrer, J (2021). topGO: Enrichment Analysis for Gene Ontology. R package version 2.46.0.

Brunson JC (2020). ggalluvial: Layered Grammar for Alluvial Plots. Journal of Open Source Software, 5(49), 2017. https://doi.org/10.21105/joss.02017.

Elbaz, A., Carcaillon, L., Kab, S., & Moisan, F. (2016). Epidemiology of Parkinson's disease. Revue neurologique, 172(1), 14-26. https://www.sciencedirect.com/science/article/abs/pii/S0035378715009224.

Gu, Z. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849. https://doi.org/10.1093/bioinformatics/btw313.

Kolberg L., Raudvere U., Kuzmin I. et al (2020). gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset. F1000Research, 9(ELIXIR):709. https://doi.org/10.12688/f1000research.24956.2.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1 - 26. http://dx.doi.org/10.18637/jss.v028.i05.

Lancour, D., Dupuis, J., Mayeux, R. et al (2020). Analysis of brain region-specific co-expression networks reveals clustering of established and novel genes associated with Alzheimer disease. Alz Res Therapy 12, 103. https://doi.org/10.1186/s13195-020-00674-7

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

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K (2022). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.4.

Michel Lang and Patrick Schratz (2023). mlr3verse: Easily Install and Load the 'mlr3' Package Family. R package version 0.2.8. https://CRAN.R-project.org/package=mlr3verse.

National Institute on Aging (2023). Alzheimer’s disease fact sheet. National Institute on Aging. https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet

Oh S, Geistlinger L, Ramos M, Blankenberg D, van den Beek M, Taroni J, Carey V, Waldron L, Davis S (2022). GenomicSuperSignature facilitates interpretation of RNA-seq experiments through robust, efficient comparison to public databases. Nature Communications, 13. https://doi.org/10.1038/s41467-022-31411-3.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wenk G. L. (2006). Neuropathologic changes in Alzheimer's disease: potential targets for treatment. The Journal of clinical psychiatry, 67 Suppl 3, 3–23. https://www.psychiatrist.com/read-pdf/12705/.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu x, Liu S, Bo X, Yu G (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141.

Yu G, Wang L, Han Y, He Q (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5), 284-287. https://doi.org/10.1089/omi.2011.0118.

Bioinformatics Project Teammate Evaluation Rubric

Rating system: 5-strongly agree, 4-agree, 3-neutral, 2-disagree, 1-strongly disagree

Category	Scoring Criteria	Alexandre Spallone	Milan Solanki	Emily Kim	Anne Michel	Jonathan Joseph
Teamwork	Responded quickly to group work	5	5	5	5	5
	Learned background information that enhanced understanding of the project.	5	5	5	5	5
	Framed the critical goals and/or questions being asked in the project.	5	5	5	5	5
Content	Data preparation	5	5	5	5	5
	Did one clustering approach	5	5	5	5	5
	Differential expression analysis	5	5	5	5	5
	Heatmap	5	5	5	5	5
	Ontology enrichment analysis	5	5	5	5	5
Participation	Finished their part of the assignment code	5	5	5	5	5
	Finished their part of the assignment writeup	5	5	5	5	5
Score	**Total Points (of 50)**	**50**	**50**	**50**	**50**	**50**
Comments:	Keep:					
	Improve:					