Raehash Shah and Max Shushkovsky
Graduate Project Final Report
Genomics and Epigenetics of the Brain

Analysis of Blood bulk RNA-seq and scRNA-seq in Alzheimer's Disease

Abstract

Alzheimer's Disease (AD) is a progressive disorder that destroys memory and other mental functions like thinking and the ability to perform daily tasks. Because of the neurodegenerative progression of the disease, it is important to find early diagnostic markers of the disease. RNA-sequencing has emerged as a powerful tool to explore the transcriptomic landscape of diseases, which makes it possible to see how gene expression patterns impact disease progression. In this project, RNA-seq data from peripheral blood samples was used to identify potential biomarkers for AD by analyzing differential gene expression. From those results, GSEA and KEGG pathway enrichment analysis revealed upregulated and downregulated pathways associated with ribosomal transcription and other common neurodegenerative diseases like Parkinson's disease and Amyotrophic lateral sclerosis. Top differentially expressed genes showed high relevance to AD symptoms and cognitive decline. Furthermore, LASSO and elastic net logistic regression predictive models on scRNAseq showed a range from poor (0.5) to high (0.938) AUROC scores, which hints at the potential ability to use bulk RNAseq data as a predictor for scRNAseq samples in AD patients. These results demonstrate that RNAseq data can become a pathway from transcriptomic data to true clinical purposes. New medicines can be developed to target the significant genes expressed from RNAseq data to combat disorders such as AD.

Introduction

Alzheimer's Disease (AD) and other forms of dementia are among the most prevalent neurodegenerative disorders, affecting millions of individuals worldwide. AD alone accounts for 60-80% of dementia cases. The most common symptoms of this disease are presence of amyloid-beta plaques, tau neurofibrillary tangles, and progressive cognitive and memory decline (Alzheimer's Association, 2023). Since it is such a damaging disease, the underlying molecular mechanisms of AD and other dementia disorders need to be studied further to develop effective diagnostic and therapeutic strategies (Querfurth & LaFerla, 2010). Identifying these biomarkers that can reliably detect early stages of these diseases is critical for improving patient outcomes and tailoring treatment strategies.

In recent years, transcriptomics has emerged as a promising approach for understanding the complex biological processes associated with neurodegeneration (Sierksma et al., 2020). RNA sequencing (RNA-seq) technology has become a common practice in the study of gene expression to analyze the entire transcriptome (Stark et al., 2019). This approach provides information about both the coding and non-coding RNAs, which allow researchers to understand disease pathogenesis. Furthermore, RNA-seq enables the identification of long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) that may act as potential biomarkers due to their regulatory roles in gene expression. Integrating RNA-seq with advanced bioinformatics tools, such as differential expression analysis, gene set enrichment analysis (GSEA) and pathway enrichment analysis, allows for the identification of key genomic and molecular pathways associated with disease phenotypes (Zhang & Horvath, 2005). More recent technological advancements in the field have led to the innovation of single cell RNA-seq (scRNA-seq) which is a methodology that can get gene expression at the individual cell level. With a completely different protocol, this cell specific granularity gives more power to analyzing individual cells but may not be as useful for studying overall trends. Therefore, looking at both bulk RNAseq and scRNAseq data may give a complete picture of what is seen in the sample.

In the context of neurodegenerative diseases like Alzheimer's Disease (AD), both bulk RNA-seq and scRNA-seq have been applied to identify differentially expressed genes (DEGs) and molecular signatures that are correlated with disease onset and progression (Zhu et al., 2022). By examining the RNA profiles of brain tissue samples, cerebrospinal fluid, or peripheral blood, researchers were able to detect changes in gene expression linked to neuroinflammatory pathways, synaptic function, and amyloid-beta metabolism, which are key features of AD (Sierksma et al., 2020). Additionally, machine learning models, like logistic regression, can be applied to these large-scale datasets to prioritize candidate biomarkers with high diagnostic and prognostic value. This multi-dimensional approach helps pinpoint specific biomarkers that not only serve as early diagnostic indicators but also offer therapeutic targets, thus supporting the development of precision medicine strategies for AD and other dementia-related disorders (Zhu et al., 2022).

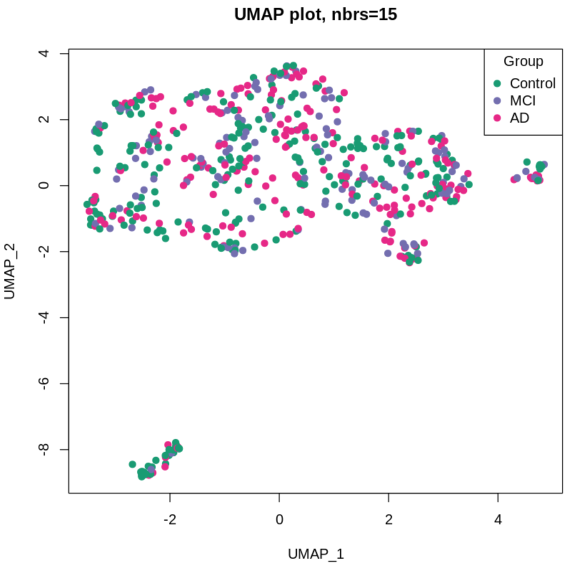


Figure 1: Uniform Manifold Approximation and Projection (UMAP) of the different samples. Each of the samples from GSE140829 (Control = 249 samples, MCI = 134 samples, AD = 204 samples), were plotted in two dimensions to visualize any differences between the normalized gene expression data of each sample in the dataset. The hyperparameters of the UMAP used were a random state and a default number of neighbors equal to 15. The use of UMAP to display the points was to allow for visualization of the data without assuming any linearity in the data to be present. However, an important caveat of a UMAP is that the distances in this space don't truly reflect the true distance in the samples while still preserving the structure of the data.

In this project, peripheral blood bulk RNA expression data in dementia disorder patients from Gene Expression Omnibus (GSE140829) was analyzed for a total of 587 samples. As seen in Figure 1, the patients were labelled as Alzheimer's Disease (AD) patients, Mild Cognitive Impairment (MCI) patients and control patients. The reason for the inclusion of the MCI patients was to see if similar markers found in MCI patients were also found in AD patients. The raw gene expression was normalized by sample and batch. However, after considering these factors, it is hard to distinguish patients as seen in Figure 1, even between dementia disorder patients and the control patients. This motivates the reason for further gene expression analysis needed.

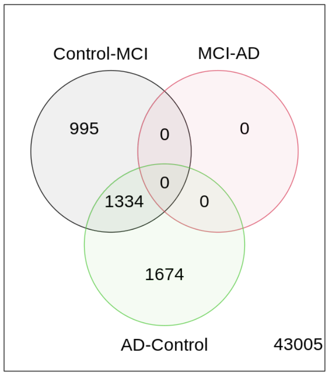
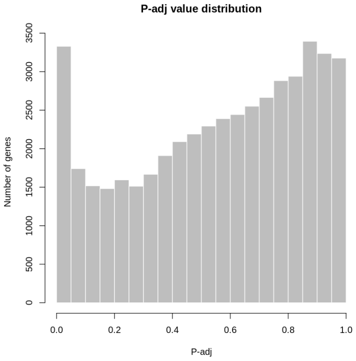
The first stage in the analysis was to perform differential gene expression with the normalized data using the R package "limma" to fit a linear model on the gene expression data and the label of the sample. Using an empirical Bayes method, genes with a log2FoldChange (FC) > 0 and adjusted p value < 0.05, the differentially expressed genes (DEGs) were identified. Using the set of DEGs, gene set enrichment analysis (GSEA) was performed to see if the set of genes that distinguished a dementia disorder patient from a control patient were involved in a larger biological process being activated or suppressed. In addition, the set of DEGs were mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database to identify a more global picture on the processes involved including genomic, chemical and systemic functional information. The goal was to see upregulated pathways involved in neuronal degeneration in AD and MCI patients compared to control patients.

In addition to being informative about the mechanism of action in dementia disorder patients, the DEGs can also be used as a predictive tool to see if a patient is at risk of developing AD. Therefore, a least absolute shrinkage and selection operator (LASSO) regression and an elastic net logistic regression model were fit to the data and analyzed to see how well it could accurately label a patient to have AD based on their gene expression. Lasso regression is a statistical tool that uses L1 regularization to select variables important to prediction, while elastic net regression uses both L1 and L2 regularization. The regularization terms control which DEGs are selected and how important they are for prediction. For lasso regression, a 5-fold cross validation was performed while selecting features based on maximizing Akaike Information Criteria (AIC) score. For elastic net regression, a 5-fold cross validation was performed with a grid search on the best regularization strength value and L1 and L2 ratio. These models were trained on a random 80% of the samples from GSE140829 and tested on the remaining 20%. As well, to test the robustness of the model, scRNAseq data (n = 8 samples) from blood samples (Lefterov Koldamova, 2024) was used as the test dataset. The hope is that the model developed is both an accurate model and a robust model in predicting AD from blood samples.

Results

A. Differential Gene Expression

DEGs were found in GSE140829 between AD and Control patients and MCI and Control patients using a linear regression model from ‘limma’ (diagnostic plots of the linear regression model can be found in the Appendix Figure 5). Using an empirical Bayes method and a log2FC > 0 and p-value < 0.05 (distribution of p-values seen in Figure 2b), 3008 DEGs were identified between AD and Control patients while 2329 DEGs were identified between MCI and Control patients. The number of DEGs found in common and unique in both analyses were plotted as a Venn diagram in Figure 2a. The 1334 common DEGs may have some relationship in a dementia disorder that should be explored further but wasn’t in this analysis. Each of these DEGs for both analyses were plotted as a volcano plot in Figure 2c and 2d. An interesting artifact of the volcano plot pattern is that they seem to follow a similar shape in upregulation and downregulation suggesting that perhaps the DEGS identified may have a similar differential expression pattern in dementia disorder patients compared to control patients.

a.  Control-MCI MCI-AD 995 0 0 0 1334 0 1674 AD-Control 43005	b. 
c.	d.

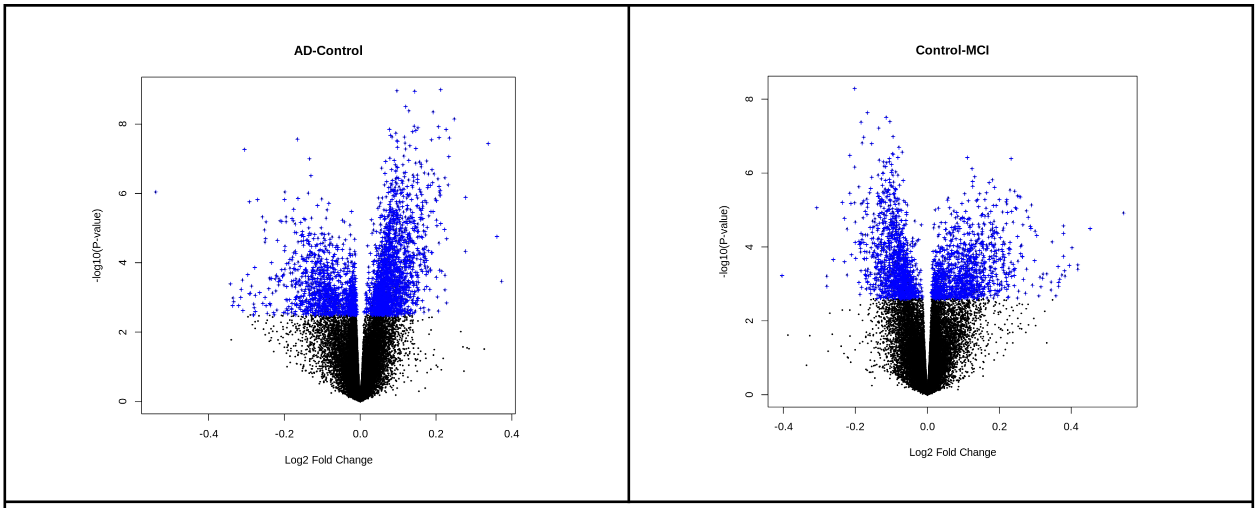


Figure 2: Differentially Expressed Genes (DEGs) identified after performing Differential Expression Analysis. The number of DEGs found in each analysis between (AD and Control) & (MCI and Control) were compared. In Figure 2a, the number of DEGs that are unique and overlapping between both models can be seen. In Figure 2b, the p-values of all the genes can be seen showing a good proportion of them being statistically significant. These statistically significant genes are plotted in the volcano plot for each model in 2c and 2d. An important note is that there is a slight distinction in both experiments. AD-Control means that a positive log2FoldChange means a decrease in expression in AD patients. However, Control-MCI means that a positive log2FoldChange means an increase in expression in MCI patients.

B. GSEA and KEGG Analysis

Using the set of DEGs identified in both experiments, GSEA and KEGG analysis was performed with a background organism of ‘org.Hs.eg.db’ for GSEA and ‘has’ for KEGG. For GSEA, in addition to the DEGs and their corresponding log2FC and p-value, all ontology was visualized with a GS Size (or number of genes included in a specific gene set for a process) between 3 and 800 and a p-value cutoff of 0.05. For KEGG, in addition to the same DEGs as input and the same GS Size and p-value hyperparameter as GSEA, the number of permutations was set to 10000 to ensure the best result is returned from the analysis. The activation and suppression of processes can be seen in Figure 3 and Figure 4 for AD and MCI patients respectively. Further network analysis was performed on these processes and pathways as seen in the Appendix Figure 6, however were not significant to the discussion of the results.

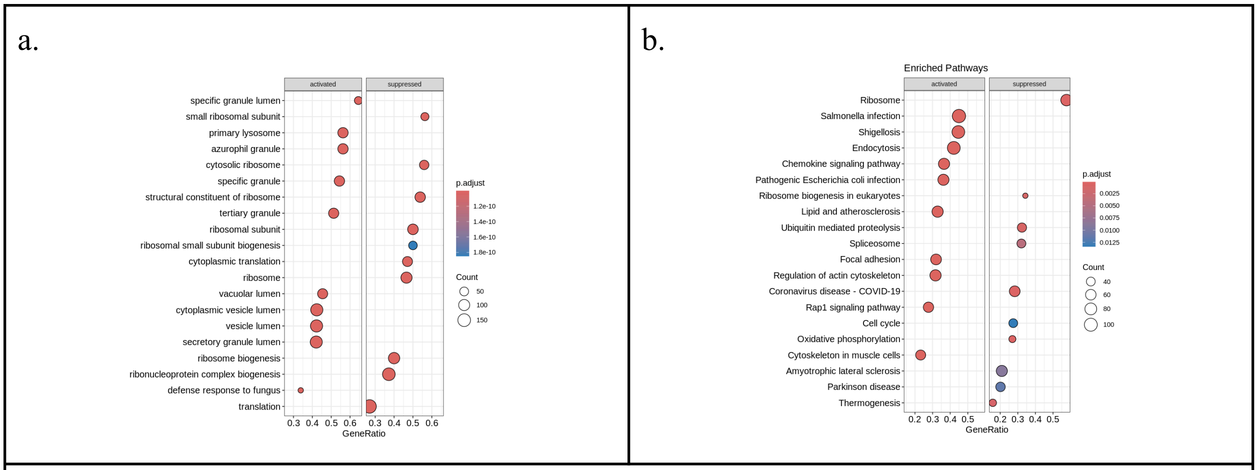


Figure 3: GSEA and KEGG Pathway Analysis for AD. In Figure 3a, the top 10 enriched (activated and suppressed) gene ontologies are visualized. In Figure 3b, the top 10 enriched (activated and suppressed) pathways are visualized. For each process, the associated p-value, number of genes mapped to the process and gene ratio is plotted. Note that based on how the model is defined, activation means downregulated in AD, while suppression means upregulated in AD.

From the GSEA analysis performed comparing AD and control patients, the main observations found were that processes associated with ribosomal transcription and translation in blood samples seem to be upregulated in AD patients. This suggests that there must be an increase in protein production in AD patients which may lead to the symptoms of AD we observe. In addition, there also seems to be a down regulation of lysosomes and cytoplasmic lumen properties suggesting that in AD, a decrease in these cellular components can cause a patient to be more likely to develop AD.

The KEGG analysis showed more intuitive correlation with AD. Parkinson's disease and Amyotrophic lateral sclerosis (ALS) are both other neurodegenerative diseases, whose pathways are upregulated in AD patients. This suggests that similar mechanisms of actions may exist for these neurodegenerative diseases. It was surprising to see COVID-19 as a pathway upregulated in AD patients which may be something that should be explored further to see if there is a true correlation between both pathways. It was hypothesized that the reason infection pathways were downregulated in these patients was that it is unlikely for a patient to get blood drawn during a bacterial infection. However, further exploration into why the Rap1 signaling pathway was downregulated in AD patients may further explain the mechanism of action of AD. Rap1 is a GTPase protein involved in many cellular processes including platelet adhesiveness, and T cell interaction has been linked with cancer biogenesis and metastasis. Inhibition of Rap1 activity has also been shown to impair tumor progression (Looi et al., 2020). Perhaps the opposite may be true in how to treat AD patients.

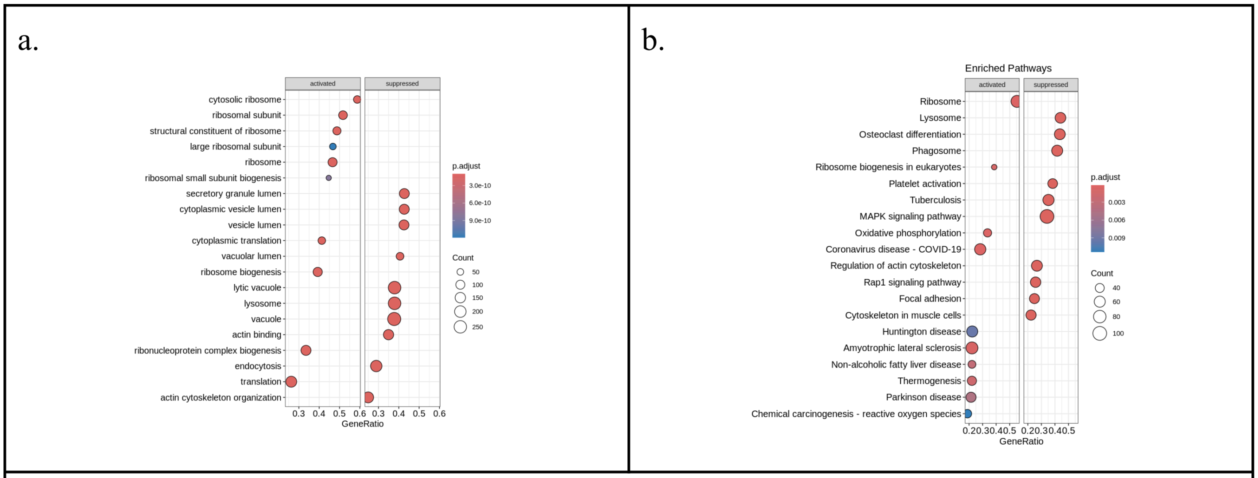


Figure 4: GSEA and KEGG Pathway Analysis for MCI. In Figure 4a, the top 10 enriched (activated and suppressed) gene ontology is visualized. In Figure 4b, the top 10 enriched (activated and suppressed) pathways are visualized. For each process, the associated p-value, number of genes mapped to the process and gene ratio is plotted. Note that based on the model is defined, activation means upregulated in MCI, while suppression means downregulated in MCI.

Comparing the GSEA analysis for MCI patients compared to AD patients, there seems to be similar ribosomal transcription and translation processes upregulated in MCI as upregulated in AD. In addition, similar lysosome and cytoplasmic processes are also downregulated in MCI and AD patients. This shows consistency between MCI and AD gene ontology. This was further validated in the KEGG pathway analysis where ALS and Parkinson disease pathways were found upregulated in MCI. It was interesting to see that the Huntington disease (another neurodegenerative disease) pathway was also found to be upregulated in MCI but wasn't in the AD patients. Like AD, the Rap1 pathway is also downregulated in MCI patients. However, there seems to also be more pathways downregulated in MCI patients like MAPK which may indicate a place for more exploratory research as a treatment mechanism.

C. Top 100 DEGs Identified in AD

From the 3008 DEGs identified between AD patients and the Control patients, the top 100 upregulated and downregulated DEGs were identified based on the log2FC value. The values for the top 5 of each set are listed in Table 1 and Table 2.

Table 1: Top 5 Up-regulated DEGs in AD compared to Control patients

Gene Symbol	Log2FC	P-value	Brief Description
HBE1	-0.5399009	1.95E-07	hemoglobin, epsilon 1
HLA-DR86	-0.3402045	0.02747645	major histocompatibility complex, class II, DR beta 6 (pseudogene)

RPS3A	-0.3337411	0.00027931	ribosomal protein S3A
PTPRC	-0.265437	0.00037232	protein tyrosine phosphatase, receptor type, C
RPL17	-0.2820639	0.00061063	ribosomal protein L17

Looking into the top 5 up-regulated DEGs, these genes have lower expression in AD patients versus control patients. HBE1 has been associated with developmental processes and cellular oxygen transport (Martinez-Moro, et al.). This is extremely interesting because of the direct implications with AD, a neurodegenerative disease. Ensuring proper cellular oxygen transport is key to stable function of many cells, like neurons. As a result, it could be lacking in AD patients. HLA-DR86, albeit a pseudogene, could have a link to AD because of immune-related pathway issues. Common issues in the immune system lead to neuronal defects. RPS3A and RPL17 are ribosomal proteins, so they are needed in protein synthesis. Lower expression of ribosomal proteins could mean protein processes in the brain are hampered. Lastly, lower/impaired expression of PTPRC, a regulator of immune cell signaling, could lead to similar effects as HLA-DR86 (Al-Barashdi, et al.). These genes all point to a strong association between lack of expression and AD compared to a higher expression seen in control patients.

Table 2: Top 5 Down-regulated DEGs in AD compared to Control patients

Gene Symbol	Log2FC	P-value	Brief Description
FTLP2	0.37247127	0.00015487	ferritin, light polypeptide pseudogene 2
MMP9	0.36102665	9.58E-05	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)
HIST2H2AA3	0.33704372	1.41E-08	histone cluster 2, H2aa3
RN28S1	0.32712511	0.03072954	RNA, 28S ribosomal 1
DYSF	0.27781606	2.15E-06	dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive)

Looking into the top 5 down-regulated DEGs, these genes have higher expression in AD patients compared to control patients. Elevated expression of the pseudogene FTLP2 hints at overactive iron metabolism, which can lead to oxidative stresses. MMP9 is involved in synaptic plasticity, so increased expression of this gene is interesting (Dziembowska and Wlodarczy). Although initial hypotheses would conclude that more synaptic plasticity is beneficial, perhaps too much expression leads to the inability to retain important information due to a constant rewriting of information. Since HINST2H2AA3 is a histone cluster gene, there may be structural issues/differences in the chromatin between AD and control patients (Kakehashi, et al.). Similar to the implications of some of the up-regulated ribosomal protein DEGs, RN28S1 is a ribosomal

RNA gene; increased protein synthesis could be detrimental for AD patients. DYSF is involved with membrane repairs (Paulke, et al.). This means that neural membrane repair systems could be hampered due to increased expression in AD patients. These genes, when considered in conjunction with each other, highlight additive effects that correlate with the symptoms of AD patients.

D. Risk Prediction with Logistic Regression

Table 3: Accuracy and Area Under Receiving Operating Characteristic (AUROC) Scores displayed from different model arrangements. Below are all of the different models, training data and testing data arrangements and their associated accuracy and AUROC values. The AUROC curves can be found in Appendix Figure 7.

Training Data	Model	Testing on 20% of Bulk RNAseq Data	Testing on scRNAseq Data
80% of Bulk RNAseq Data	Lasso Regression	AUROC: 0.587	AUROC: 0.938
All Bulk RNAseq Data	Lasso Regression		AUROC: 0.813
80% of Bulk RNAseq Data	ElasticNet Regression	Accuracy: 0.628 AUROC: 0.629	Accuracy: 0.625 AUROC: 0.625
All Bulk RNAseq Data	ElasticNet Regression		Accuracy: 0.5 AUROC: 0.5

Using the top 100 up-regulated DEGs and top 100 down-regulated DEGs, different logistic regressions were created and performed to analyze performance as seen in Table 3 (AUROC curves seen in Appendix Figure 7). In Zhu et al., 2022, the researchers performed a lasso regression to perform gene filtration and create a risk detection model. Therefore, using these 200 DEGs, a lasso regression was trained on 80% of the patients in the bulk RNA-seq data and tested on the remaining 20% of patients. The model achieved a mediocre performance with an Area Under Receiving Operating Characteristic (AUROC) value of 0.587. However, when training an Elastic Net Regression model with the same 80% of patients in the bulk RNA-seq data and testing it on the remaining 20% of patients, the model achieved greater performance with an AUROC score of 0.629 and an accuracy of labelling of 0.628. This suggests that the Elastic Net Regression model performs better in predicting risk for AD patients.

Another question this investigation explored was the robustness of the model it created and if it generalized well to a scRNAseq dataset. To do this, the scRNAseq data was merged across the cells to create a pseudo-bulk RNAseq dataset. Using the lasso regression model, the scRNAseq data achieved a remarkable AUROC value of 0.938. When training the model on the full bulk RNAseq data and testing it on the scRNAseq data the lasso regression model achieved another high AUROC value of 0.813. In comparison, the elastic net regression model trained on 80% of the bulk RNAseq patients only achieved an AUROC value of 0.625 and accuracy of 0.625. Similarly training on the full bulk RNAseq data and testing on the scRNAseq data achieved an AUROC value of 0.5 and accuracy of 0.5. These results suggest that the lasso regression performs better than elastic net regression which is the opposite of what was shown when testing on 20% of the bulk RNA-seq data. This may be since there are only 8 samples in the scRNAseq data. This could mean that the lasso regression model overfit to the data and therefore achieved high performance. It is also possible that the intrinsic difference in nature between bulk RNAseq data and scRNAseq data leads to the model performance difference. A future more robust direction is to explore model performance between bulk RNAseq and scRNAseq data from the same set of patients to see if the difference is truly due to the different data generation process.

Conclusion and Future Work

In this study, DEGs were identified between dementia disorder patients and control patients using bulk RNAseq data from blood samples. These DEGs were used to perform GSEA and KEGG pathway analysis to see if the gene sets identified were associated with known processes and pathways affiliated with AD and MCI. From the 3008 DEGs identified between AD and control patients, the top 100 upregulated DEGs and top 100 downregulated DEGs were explored and used to create logistic regression models. The elastic net regression model had the best performance when tested on the bulk RNAseq data. When tested on the scRNAseq data, the lasso regression model had better performance. However, neither model seems to achieve performance adequate for deployment of prediction for AD. Therefore, further exploration in the DEGs and model development may be necessary to create an optimal model that can be used in practice.

There are multiple exciting avenues to further explore as a result of this investigation. First, there were 1334 common DEGs between AD-control and MCI-patients. This merits a deeper dive into how various expression levels of the genes could cause worsening symptoms and potentially progression from MCI to AD. Next, there could be an analysis of correlation between COVID-19 patients and AD patients, due to it being an upregulated pathway in AD patients. Given how COVID-19 is a relatively quite new disease that took the world by storm, there could be novel insights between the mechanisms of COVID-19 that translate to AD, a disease that has been uncovered for longer. Finally, a different computational approach would be determining model performance from bulk RNAseq and scRNAseq data obtained from the same patients. Furthermore, purely having more patients to obtain scRNAseq from would be extremely beneficial. This would provide a more rigorous and statistical approach in determining if the changes in model performance are due to data generation processes or real biological processes.

In the vein of more data, incorporating temporal data to better understand the progression of AD through changes in gene expression would be exciting to look into. If possible, having more discrete stages of RNA-seq data than just control, MCI, and AD can lead to better understanding the transition of up-regulated and down-regulated genes. Ultimately, the availability of both bulk RNAseq and scRNAseq data mean exciting and statistically rigorous analyses can be conducted to better understand the AD progression pathway.

References

Al Barashdi, Maryam Ahmed et al. "Protein tyrosine phosphatase receptor type C (PTPRC or CD45)." *Journal of clinical pathology* vol. 74,9 (2021): 548-552. doi:10.1136/jclinpath-2020-206927

Alzheimer's Association. (2023). Alzheimer's Disease Facts and Figures.

Dziembowska, Magdalena, and Jakub Wlodarczyk. "MMP9: a novel function in synaptic plasticity." *The international journal of biochemistry & cell biology*vol. 44,5 (2012): 709-13. doi:10.1016/j.biocel.2012.01.023

Kakehashi, Anna et al. "Cytokeratin 8/18 overexpression and complex formation as an indicator of GST-P positive foci transformation into hepatocellular carcinomas." *Toxicology and applied pharmacology* vol. 238,1 (2009): 71-9. doi:10.1016/j.taap.2009.04.018

Looi, C., Hii, L., Ngai, S. C., Leong, C., & Mai, C. (2020). The Role of Ras-Assocated Protein 1 (Rap1) in Cancer: Bad Actor or Good Player? *Biomedicines,* 8(9) 334.

Martínez-Moro, Álvaro et al. "RNA-sequencing reveals genes linked with oocyte developmental potential in bovine cumulus cells." *Molecular reproduction and development* vol. 89,9 (2022): 399-412. doi:10.1002/mrd.23631

Paulke, Nora Josefine et al. "Dysferlin Enables Tubular Membrane Proliferation in Cardiac Hypertrophy." *Circulation research* vol. 135,5 (2024): 554-574. doi:10.1161/CIRCRESAHA.124.324588

Querfurth, H. W., & LaFerla, F. M. (2010). Alzheimer's disease. *The New England Journal of Medicine*, 362(4), 329-344.

Sierksma, A., Escott-Price, V., & De Strooper, B. (2020). Translating genetic risk of Alzheimer's disease into mechanistic insight and drug targets. *Science*, 370(6512), 61-66.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11), 631-656.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Zhu, M., Hou T., Jia L., Tan Q., Qiu C., Du Y., (2022) Development and validation of a 13-gene signature associated with immune function for the detection of Alzheimer's disease. *Alzheimer's Disease Neuroimaging Initiative*, 125. 62-73.

Appendix

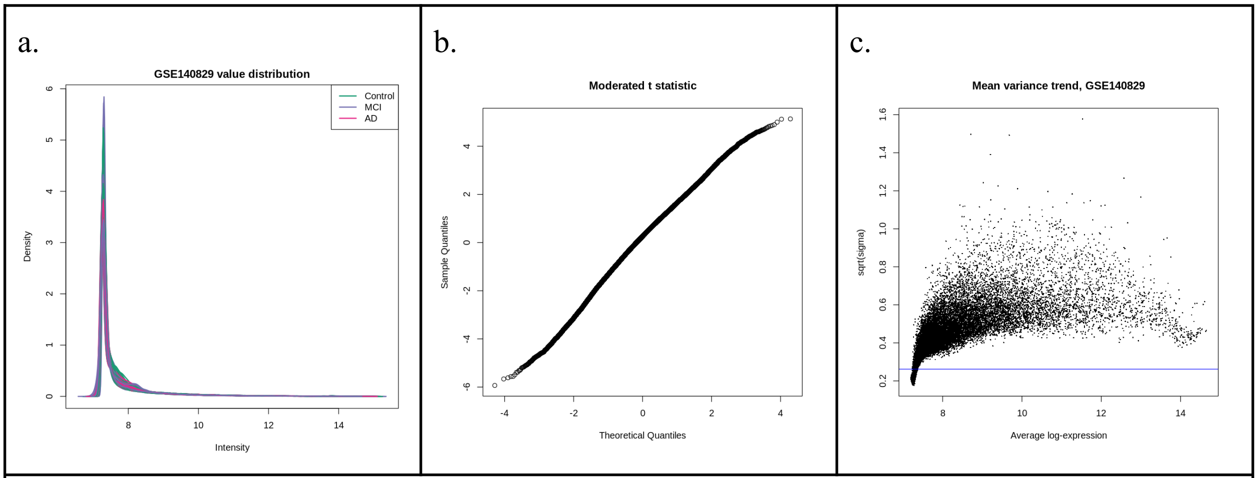


Figure 5: Diagnostic Plots of the Linear Model in performing Differential Expression Analysis. In Figure 5a, the distribution of the normalized gene expression is visualized. It is clear there is some separation in intensity after normalization especially for smaller values which the linear model of ‘limma’ captures. In Figure 5b, the moderated t statistic or Q-Q Plot for the linear model is displayed. Based on the plot the residuals of the linear model seem to fall on a straight line meaning they are normal. In Figure 5c, the mean-variance trend or the scale-location plot shows how the residuals are spread along the ranges of predictors. From the plot since the points don’t seem to be equally spread, it suggests that the assumption of homoscedasticity doesn’t hold, which perhaps is something to consider as a future direction.

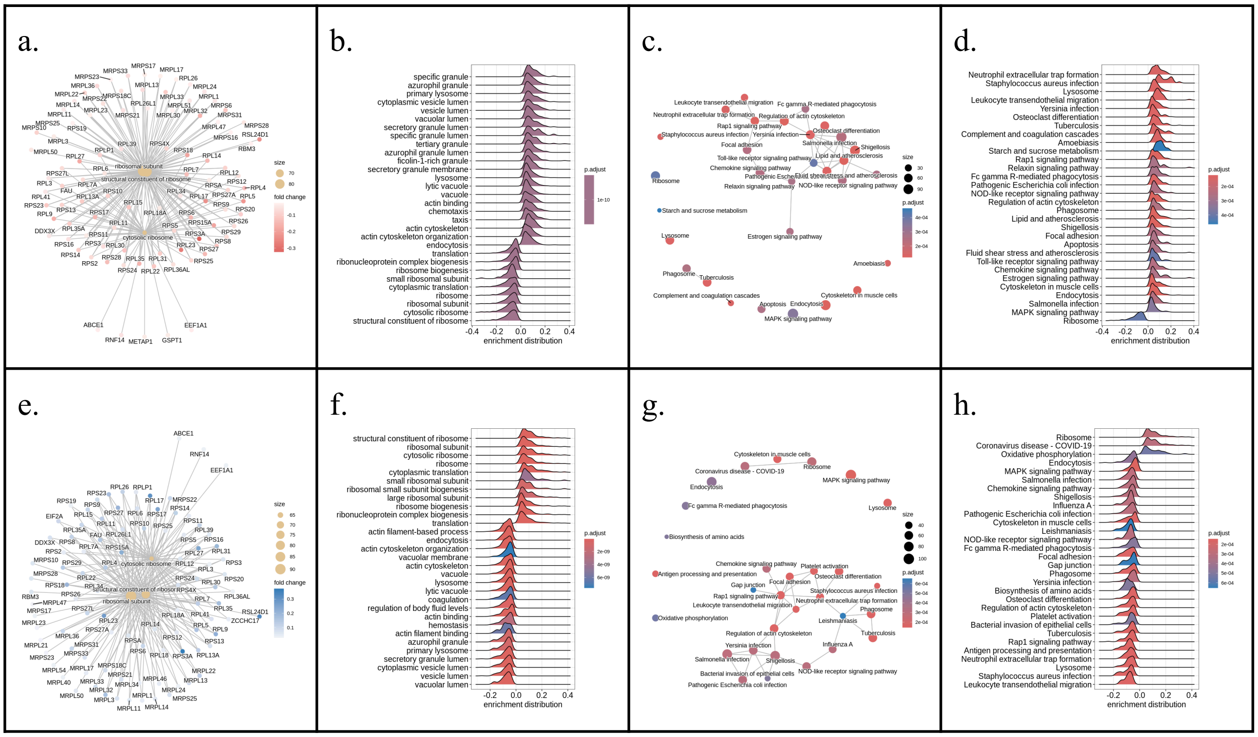


Figure 6: Further exploration of GSEA and KEGG enrichment processes. Figure 6a-d are results from GSEA and KEGG from AD and Control analysis while Figure 6e-h are results from GSEA and KEGG from MCI and Control analysis. After identifying the processes, the genes mapped were paired to connect them together. In Figure 6a and 6e, the GSEA processes are plotted with the associated genes that are mapped to the process. Figure 6b and 6f is a ridge plot showing the level of enrichment distribution for the process. Figure 6c and 6g is an expression map plot of the KEGG pathways connected together and annotated based on enrichment. Figure 6d and 6h is a similar ridge plot for the KEGG pathways in enrichment.

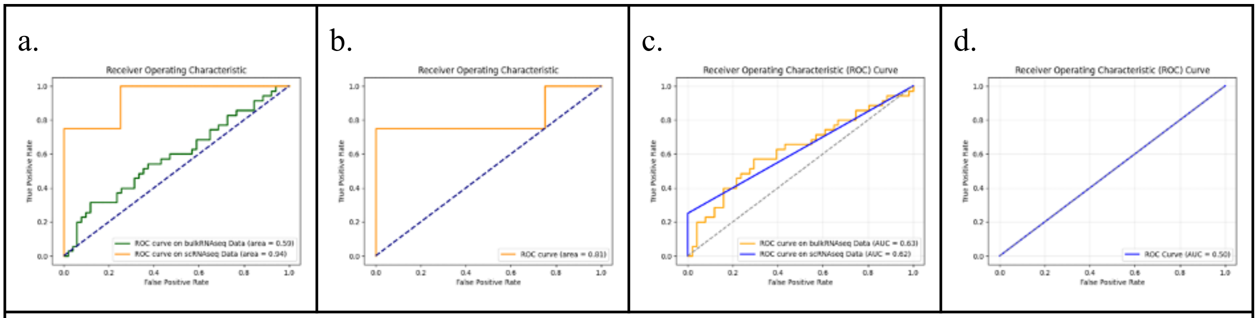


Figure 7: Receiver Operating Characteristic (ROC) Curves for each of the logistic regression models developed. A plot of the true positive rate and false positive rate for each of the models is shown above. The order of the ROC curves are the same as Table 3 where Figure 7a and 7b are performed using a Lasso Regression and Figure 7c and 7d are performed using an Elastic Net Regression.