|  |
| --- |
| **Analysis of Alzheimer’s Disease Gene Expression for Diagnostic Applications**  Zeal Jinwala1, Kristen Norray 1, Henry Hollis 1  1 School of Biomedical Engineering, Drexel University, USA  Course: BMES 543  Instructor: Ahmet Sacan  Date: 2020-06-07  Original Paper(s):  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6060554/>  <https://pubmed.ncbi.nlm.nih.gov/23622250/>  Datasets:  <https://ftp.ncbi.nlm.nih.gov/geo/series/GSE63nnn/GSE63063/matrix/>  <https://ftp.ncbi.nlm.nih.gov/geo/series/GSE44nnn/GSE44772/matrix/> |

[[1]](#footnote-2)\*abstract

This study performs differential expression analysis and classification based on gene expression between patients clinically diagnosed with Alzheimer’s Disease and controls. Gene expression data was microarray-based from the brain and the blood. We find the sets of differentially expressed genes meeting a significance (p<0.05) and fold change (absolute log fold > 0.32) requirement are disjoint between the datasets from blood and brain. Transcripts from the brain are enriched for inflammation, cellular responses to metal ions, and apoptosis. Transcripts from the blood are enriched for translational processes, rRNA processing, and mitochondrial energetics. The pooled differentially expressed genes from all datasets are enriched for pathways involved in neurodegeneration, including Alzheimer’s disease. For classification, we trained a support vector machine classifier to predict disease diagnosis based on gene expression (control vs mild cognitive impairment and control vs Alzheimer’s disease). After feature selection for each task, we were able to achieve an 83.5% Alzheimer’s vs control accuracy and an 80.2% mild cognitive impairment vs control accuracy.

# introduction

Alzheimer’s disease (AD) is a disorder that causes brain cells to deteriorate, leading to a decline in thought processing and day to day independence. There are currently 50 million AD patients worldwide and that number is projected to double every five years until 152 million are affected by 2050. With a growing population of AD patients there has been more emphasis placed on the study of AD prevention and treatment costing the US $1 trillion annually (Breijyeh et al.). Although great strides have been made, clinicians still disagree about what defines AD due to the different biomarkers expressed in AD patients. This provides motivation to understand the omic definition of AD, allowing for the diagnosis of AD from gene expression.

To send messages across the brain, neurons use electrical and chemical signaling through molecules called neurotransmitters. The function of these neurotransmitters is impaired in AD patients causing a loss of connectivity leading to the degeneration of neurons. Amyloid plaque and tau neurofibrillary tangles (NFTs) contribute to synaptic loss in AD patients (**Figure 1**). The presence of amyloid plaques is suspected to be an indicator of AD due to its role in neurotoxicity. The extracellular deposits of amyloid beta (Aβ) protein cause damage to the axons and eventually neuronal death (Breijyeh et al.). NFTs are another characteristic seen in AD patients because they are resistant to proteolysis and cause cell death.

This study aims to aid help aid in defining the omic definition of AD and find differentially expressed genes between the control and AD subjects. The functional annotations of these differentially expressed genes provide biological meaning to the differences in disease states. Furthermore, there is enormous utility in predicting AD disease state from gene expression, doubly so given the lack of biomarkers for AD. The results of this study may help progress medical practice by allowing for AD patients to seek early preventative treatment. The prediction of health risks based on gene expression may provide patients with a more specific treatment plan (Sood et al., 2015). This may also allow researchers more time to discover a treatment for a late onset disease such as AD.

Diagram

Description automatically generated

**Figure 1. Neuron differences in AD patient.** Comparing a healthy brain and brain of an AD patient. a) Healthy compartments of the brain and a diagram of a healthy neuron. b) NFTs and amyloid plaques are shown within the neuron which causing widespread degeneration and shrinkage (Breijyeh et al.).

# Related Work

Previous studies have analyzed these datasets in a manner similar to our methodology. Lunnon et al performed pathway analysis on a subset of the blood data we consider (Lunnon et al., 2012). The same group also performed a classification task on a subset of the blood data (Lunnon et al., 2013). Their goal in that study was to predict the trajectory of mild cognitive impairment (MCI) patients from gene expression derived from blood (blood is easily obtainable tissue, compared to brain). They trained a random forest model on AD and control patients, suppling the classifier with gene expression data and then expression data combined with brain imaging data.

# Dataset

All datasets used in this work are microarray data collected form H. sapiens. Two of the datasets (GSE6306 & GSE63061) are from human blood with a varying severity of AD. While they are from the same experiment, they are split into separate GEO series because they use two different platforms (Illumina HumanHT-12 V3.0/V4.0) (Sood et al., 2015). The other dataset we used is a series of 690 human brain tissue samples divided into three subseries based on the area of brain sampled (Cerebellum, visual cortex and prefrontal cortex). This series has annotated disease states, but they are either control (CTL) or AD (no MCI subjects) (Zhang et al., 2013).

# methods

An overview of our methods in schematic form is present in Figure 2. Microarray data were downloaded from GEO database with the R package GEOquery (Davis and Meltzer, 2007). Differential expression is performed between the AD and CTL groups using the Limma package (Ritchie et al., 2015). We then restrict each dataset’s probes to the intersection of all probes across all datasets. We then perform Z-normalization and save the normalized data, metadata, and differentially expressed genes to files in a temporary directory specified in MATLAB. This is all handled in GEO\_z\_norm\_de.R.

We then use venn\_diagram\_de\_genes.R to call the vennDiagram R package (Chen and Boutros, 2011). This Venn diagram shows the overlap of the differentially expressed genes from each GEO series (**Figure 3**).

The differentially expressed genes were uploaded to NCBI’s DAVID program (Sherman et al., 2022), which performs an overrepresentation-based approach to determine if gene sets are overrepresented in our lists of differentially expressed genes. After downloading the results from DAVID, we use enrich\_script.R to call the rrvgo R package (Reduce + Visualize GO • rrvgo) which creates tree map plots of the Gene Ontology Biological Processes overrepresented in our differentially expressed genes (**Figure 4, 5**).

We had two classification tasks: predicting control vs AD and control vs MCI. For both tasks we employed support vector machines (SVM). We combined all datasets and withheld 30% of samples for independent validation (for both AD and MCI tasks). The proportion of class labels between the training and validation groups was conserved. The MCI classifier was trained on strictly control or MCI samples and the AD classifier was trained on strictly control or AD samples. We performed sequential feature selection to identify the probes most useful in predicting both disease states. 10-fold cross validation was used to assess performance of both SVMs on their respective training data. Finally, we applied both SVMs to their respective independent validation sets.

Diagram

Description automatically generated

Figure **.** **Workflow Analysis.** This flow chart depicts the order in which the code is organized and what lead up to the results.

# Experiments and REsults

Differentially expressed genes between AD and CTL subjects were compared across datasets (**Figure 3**). GSE63060 and GSE63061 are from the same experiment, in the same tissue (whole blood). This explains the significant overlap in differentially expressed genes. The set of differentially expressed genes from brain tissue (GSE44772) is disjoint, which may be expected by virtue of being collected from a completely different tissue type.

Diagram, venn diagram

Description automatically generated

**Figure 3.** **Differentially Expressed Genes Between Sets**. GSE63060 (green) & GSE63061 (yellow) are two series from the same experiment, using whole blood (Sood et al., 2015). GSE44772 (purple) is from a series of various brain tissue (Zhang et al., 2013).

To further understand the differences between AD and CTL patients in our data, we performed pathway analysis using NCBI’s DAVID (Sherman et al., 2022). We then analyzed the Gene Ontology Biological Processes overrepresented in our differentially expressed genes using the rrvgo R package (Reduce + Visualize GO • rrvgo). The resulting treemap groups the common terms together, where the size of the boxed term is correlated to the abundance of terms in that category (**Figure 4, 5**). When comparing the pathway analysis of the blood (**Figure 4**) vs. the brain (**Figure 5**) the differences between the datasets can be clearly observed: Transcripts from the brain are enriched for inflammation, cellular responses to metal ions, and apoptosis. Transcripts from the blood are enriched for translational processes, rRNA processing, and mitochondrial energetics. We also performed pathway analysis on the pooled differentially expressed genes (union from all datasets) with DAVID (Sherman et al., 2022). This suggested that pathways associated with neurodegenerative diseases such as Parkinson disease, prion diseases, and Alzheimer’s disease were overrepresented (**Figure 6**).

Chart, treemap chart

Description automatically generated

**Figure 4.** **Pathway Analysis of GSE63060-blood.** Terms from the GO biological process library (Sherman et al., 2022) and grouped using the rrvgo R package (Reduce + Visualize GO • rrvgo). This process places common terms together to map out the common terms within the specific pathway.

Chart, treemap chart

Description automatically generated

**Figure 5. Pathway Analysis of GSE4472-brain.** The GO biological process library (Sherman et al., 2022) was used to generate terms related to the specified GSE. The R package rrvgo (Reduce + Visualize GO • rrvgo) was then used to group common terms together to better visualize the most common terms from the data.

Timeline

Description automatically generated

**Figure 6**. **Pathway Analysis of all Differentially Expressed Gene Terms.** Pathways are from the KEGG library (Sherman et al., 2022).

Classification tasks to predict AD, MCI and CTL states based on gene expression were summarized in a confusion matrix. For CTL vs. AD prediction, we obtained a true positive and true negative accuracy of 86.6% and 79.75%, respectively.

For control vs MCI prediction, we achieved a true positive and true negative accuracy of 90.78% and 51.7%, respectively **(Figure 7).** We think that because MCI represents an intermediate stage between CTL and AD states, the omic features of MCI are less distinguishable from CTL, hence the prediction of MCI samples has a lower accuracy. The differ

Chart

Description automatically generated

**Figure 7. Classification of Results.** The confusion matrices above show the analysis of AD and MCI data prediction. The y-axis reflects the true class labels, and the x-axis shows the predicted classes.

ences in expression on genes in AD samples are relatively more distinct and hence, a better prediction accuracy is achieved when classifying these two groups.

# DISCUSSION

A study by Lunnon et al looked at the blood gene expression alterations using the same human blood datasets (GSE6306 & GSE63061) we explored in our study (Lunnon et al., 2012). In their comparison of differentially expressed genes between MCI and AD, they observed associations with lower ribosomal function, decreased mitochondrial function and translation, lower secretory and endocytic function, decreased metabolic processes, altered immune function and promotion of transcription in disease condition. Similarly, in our pathway analysis for the same samples (**Figure 4, 5**), we observed that transcripts from the blood samples are enriched for translational processes, rRNA processing, and mitochondrial energetics.

In another study, Lunnon et al developed an AD diagnostic classifier for predicting the trajectory of MCI patients, and to aid in early detection of AD (Lunnon et al., 2013). They present results for classifying AD vs control patients, which is directly comparable to our work. They achieved a performance accuracy of 75% on classifying CTL vs AD subjects from gene expression. Specifically, they achieved a sensitivity of 80.8% and a specificity of 69.2%. Lunnon et al identified 50 important probes (48 genes) after performing feature selection. Our feature selection identified a total of 13 probes (6 for AD, 7 for MCI task). There was only one gene in the intersection of these two sets: RPL36AL, a gene coding for a subunit of the ribosome (RPL36AL - 60S ribosomal protein L36a-like - Homo sapiens (Human) - RPL36AL gene & protein).

One limitation of our MCI classifier is the abovementioned low sensitivity **(Figure 7).** One interesting direction to proceed with this work is introducing the extra MCI metadata that Lunnon et al use in their classifier. Those researchers have access to annotations about the disease state of MCI patients two years after the initial data collection. An optimistic possibility is that some of the patients our algorithm (mis)classified as CTL are in the set of subjects that have no AD pathology after two years. Access to this extra metadata would also be interesting to test our AD classifier on, which would let us further compare our work with Lunnon et al.

Moreover, there are dozens of other statistical tools besides support vector machines that we could apply to this data. Lunnon et al use a random forest classifier for example (Lunnon et al., 2013).

Finally, we could further explore how our classifier, trained on data from the blood and brain, performs on microarray data from other tissues or experiments. We could also apply our classification approach to other types of gene expression data, such as scRNAseq data from the same tissues and identify biomarkers for different cell type associated with AD.

# References

Breijyeh, Z., Karaman, R., Muñoz-Torrero, D., and Dembinski, R. molecules Comprehensive Review on Alzheimer’s Disease: Causes and Treatment. doi: 10.3390/molecules25245789.

Chen, H., and Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. doi: 10.1186/1471-2105-12-35.

Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. 23, 1846–1847. doi: 10.1093/bioinformatics/btm254.

Lunnon, K., Ibrahim, Z., Proitsi, P., Lourdusamy, A., Newhouse, S., Sattlecker, M., et al. (2012). Mitochondrial Dysfunction and Immune Activation are Detectable in Early Alzheimer’s Disease Blood. *Journal of Alzheimer’s Disease* 30, 685–710. doi: 10.3233/JAD-2012-111592.

Lunnon, K., Sattlecker, M., Furney, S. J., Coppola, G., Simmons, A., Proitsi, P., et al. (2013). A Blood Gene Expression Marker of Early Alzheimer’s Disease. *Journal of Alzheimer’s Disease* 33, 737–753. doi: 10.3233/JAD-2012-121363.

Reduce + Visualize GO • rrvgo Available at: https://ssayols.github.io/rrvgo/ [Accessed June 2, 2022].

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43. doi: 10.1093/nar/gkv007.

RPL36AL - 60S ribosomal protein L36a-like - Homo sapiens (Human) - RPL36AL gene & protein Available at: https://www.uniprot.org/uniprot/Q969Q0 [Accessed June 2, 2022].

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., et al. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*. doi: 10.1093/nar/gkac194.

Sood, S., Gallagher, I. J., Lunnon, K., Rullman, E., Keohane, A., Crossland, H., et al. (2015). A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol* 16, 185. doi: 10.1186/s13059-015-0750-x.

Zhang, B., Gaiteri, C., Bodea, L. G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer’s disease. *Cell* 153, 707–720. doi: 10.1016/J.CELL.2013.03.030.

1. [↑](#footnote-ref-2)