Understanding Alzheimer's: What Big Data Reveals About APOE and the Brain
Isabelle Donoghue, Francis Alford

1. Introduction

With the popularity of consumer genetic testing services, people are increasingly learning that they have a genetic predisposition to Alzheimer's Disease. For example, 23andMe's Health + Ancestry report for Late-Onset Alzheimer's Disease looks for carriers of the APOE gene, the APOE4 variant in particular, as an indicator of risk (23andMe, n.d.). Receiving this result can feel alarming, but APOE is only one part of the complicated picture that is the mechanisms behind Alzheimer's Disease.

Alzheimer's Disease (AD) is a neurodegenerative disorder that leads to memory loss and cognitive decline affecting over five percent of the world's population aged fifty and above (Gustavsson et al., 2022b). Knowing one's APOE status can inform personal and medical decisions, but it does not guarantee whether AD will develop. In this report, we will explore the role of APOE in Alzheimer's Disease.

To look beyond single-gene risk, our report combines multiple approaches to analysis to get a holistic view of how APOE influences the brain in Alzheimer's. Our hypothesis is that APOE plays an important role in AD on multiple levels of brain function. We review large-scale GWAS analyses to confirm APOE's genetic association, compare gene expression in key brain regions between AD and non-AD brains, map APOE distribution across the healthy brain, and explore structural changes from AD in APOE-heavy brain regions. We also assess whether machine learning methods can use brain volume patterns to distinguish AD brains from non-AD brains.

2. Analysis and Results

2.1 Genome-Wide Association Study (GWAS)

We began by analyzing the genome-wide association study for Alzheimer's Disease (AD) from the FUMA GWAS platform. The dataset contains gene-level p-values and Z-scores for genes associated with AD.

Methods

We first extracted relevant columns from GWAS summary statistics file, which included chromosome number, gene ID, gene symbol, Z-score, and p-value. We then applied a genome wide significance threshold of $p < 10^{-8}$ to identify genes associated with AD. The next step was to highlight the APOE gene and analyze its Z-score and p-value, as well as extract the five most significant genes based on Z-scores.

Key Findings

Over 30 genes met the significance threshold ($p < 10^{-8}$), including APOE, TOMM40, and PVRL2. APOE had one of the highest Z-scores (17.1060) as seen in Table 1, which suggests it has a significant role as a genetic risk factor for AD. Interestingly, we see non-APOE genes that are also highly associated with AD. The list of these significant genes were stored for later analysis, including gene expression profiling.

Chromosome	Gene Code	Z-score	Gene Symbol	P-value
19	ENSG00000130204	21.9210	TOMM40	8.13473-107
19	ENSG00000130202	21.7420	PVRL2	4.1268e-105

19	ENSG00000130208	20.2840	APOC1	8.8846e-92
19	ENSG00000130203	17.1060	APOE	6.6390e-66
19	ENSG00000007047	12.9520	MARK4	1.1511e-38

Table 1: Top 5 Most Significant Genes Associated with AD Ranked by Z-score

2.2 AD Gene Expression in Brain Regions

We continued our analysis by investigating gene expression of significant GWAS genes in specific brain regions implicated in AD, focusing on the entorhinal cortex (EC) which is critical to memory function.

Methods

We began by merging the list of significant genes from the GWAS results with brain tissue gene expression data, using gene symbols as our key. The expression datasets included raw expression values (ExpressionData.txt), sample metadata (SamplesMetaData.txt), and gene probe mappings (ProbeNames.txt). We then performed t-tests comparing expression levels of these genes between patient and control groups across multiple regions, paying special attention to the APOE gene and the EC region.

Key Findings

We found that the difference in APOE expression was statistically significant ($p < 0.05$) in the EC and Middle Temporal Gyrus (MTG) regions (Table 2). These results align with our understanding of AD, since the EC is important for memory, and the MTG is important for facial recognition, functions impaired by AD. Interestingly, the direction of difference was opposite for the two regions. APOE expression was higher in AD than in controls in the EC (Figure 1), but lower in AD than controls in the MTG. Higher APOE expression in the EC could be as a response to declining function from AD, while lower APOE in MTG could be as a result of AD's damage to the region. This pattern shows that APOE's role depends on the context of the disease's progression and how genes react differently to AD. We also see that APOE does not rank particularly high in statistical significance of its difference between AD and control, ranking 6th out of 11 for AD-associated genes in the EC and 7th out of 8 for AD-associated genes in MTG. This complication points to the difficulty of pinpointing causal relationships between a gene and AD, with multiple genes contributing to the picture.

Brain Region	Expression T-statistic	P-value for t-test	T-stat Rank (out of)
EC	2.5047	0.0206	6 (11)
MTG	-2.5933	0.0154	7 (8)

Table 2: Brain Regions with Significant APOE Gene Expression

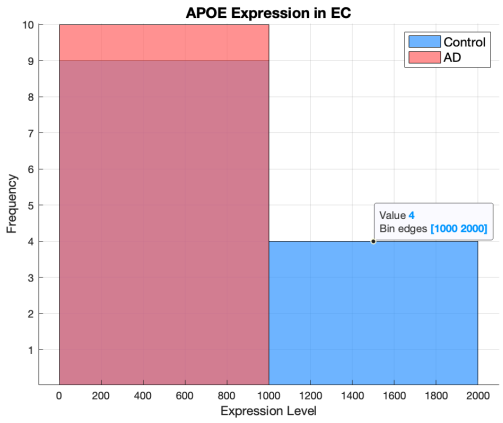

Figure 1: EC Region Expression of APOE Gene in AD vs Control samples.

2.3 Transcriptomic Profile (AHB Expression)

To see where APOE is most active in a healthy brain, we used data from the Allen Human Brain Atlas (citation). This dataset measured the amount of messenger-RNA (mRNA) in six post-mortem brains. Since mRNA is the copy of DNA that the cell uses to make protein, high mRNA levels for APOE mean that its instructions are being used while low mRNA levels mean that APOE is rarely used. This will allow us to see what brain region's cells rely the most on APOE, giving us a view of APOE's role at the microscopic level.

Methods

We focused on the left side of the cortex of the brain (the outer region) and measured average APOE expression in 57 regions of the cortex defined by the DK-114 atlas. We mapped areas of high expression on the brain to visualize APOE expression in the brain, and ranked top areas (Figure 2) (Table 3).

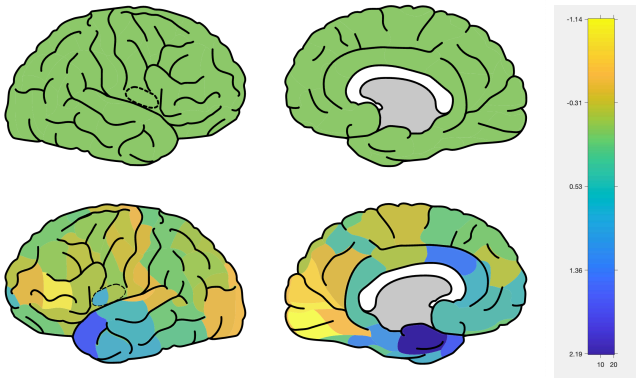

Figure 2: Cortical heatmap showing APOE expression across 57 left hemisphere regions

Rank	Region	Expression Level
1	ctx-lh-entorhinal_1	2.193
2	ctx-lh-temporalpole_1	1.5508
3	ctx-lh-caudalanteriorcingulate_1	1.29
4	ctx-lh-fusiform_2	1.1532

Table 3: Top 5 Cortical Regions with Highest APOE Expression

Key Findings

We see that areas of high APOE expression (dark blue) are concentrated in the temporal lobe, specifically in the entorhinal cortex. These regions are responsible for functions that are most affected by AD, including memory and perception of time. The fact that APOE transcription peaks in these regions suggests that these areas depend on the proper function of APOE. Carriers of APOE variants that affect its functioning may be more vulnerable to the symptoms of AD.

2.4 MRI and Structural Changes

We further examined structural brain changes in AD by analyzing MRI volumetric data and exploring its relationship with APOE gene expression levels.

Methods

We used SURASIS volumetric MRI data, as well as the corresponding clinical and session metadata, to analyze gray matter volume across 68 cortical regions in patients and controls. We then combined transcriptomic data on APOE expression levels with regional MRI measurements. Finally, we performed a linear regression analysis to see whether APOE expression levels are associated with increased atrophy in AD affected regions.

Key Findings

We observed that regions with higher APOE expression levels tend to have greater levels of gray matter loss for patients, as shown in Figure 4. The scatter plot shows a positive correlation between APOE expression and regional atrophy, suggesting that APOE likely plays a role in structural changes in affected regions. This relationship was found to be statistically significant with a p-value of 0.0279 for the slope in our linear regression model. This finding suggests a link between APOE gene expression and AD pathology in the DK68 region, highlighting APOE's potential role in disease progression.

Figure 4: Scatterplot showing significant positive correlation between APOE expression and MRI regional atrophy, suggesting APOE contributes to structural degeneration in AD

2.5 Machine Learning (ML)

For the last part of our analysis, we evaluated whether large-scale changes in the brain could be used to distinguish AD patients from controls. Using gray-matter volume from MRI scans, we trained a ridge-regularized linear model. When applied to a test set the models had never seen before, the ridge model scored an accuracy of 97.50%.

	Predicted positive	Predicted negative
True positive	39	1
True negative	1	39

Table 4: Confusion Matrix for AD Classification by Linear Model with Ridge Regularization

We also trained a neural network in batches. By the final epoch, the model scored a training accuracy of 92.97% and validation accuracy of 86.39%. After adjusting parameters, we found that changing the value of initial learning rate from 0.05 to 0.01 and momentum from 0.95 to 0.99 lead to the best test set accuracy of 84.02%. The decline in scores from training to validation to testing points to possible overfitting, as the neural network struggles to generalize to data it is not familiar with.

	Predicted positive	Predicted negative
True positive	137	9
True negative	18	5

Table 5: Confusion Matrix for AD Classification by Neural Network
These performances, especially the ridge model's, show that gray-matter volume, combined with machine learning, is an effective marker for AD, even with no information on genes like APOE.

3. Discussion

Our hypothesis was that APOE plays an important role in AD on multiple levels of brain function. Our analysis mostly supports this idea, but shows that APOE is only part of a complicated system. At the genetic level, APOE had one of the strongest signals in the GWAS. This result confirmed why consumer genetic tests screen for APOE variants. Further analysis of gene expression showed that APOE expression differed significantly between patients and controls, but the direction of change depended on the brain region. This result showed that APOE's role changes based on brain region and possibly based on AD's progression. At the microscopic level, we saw that APOE transcription is highest in the EC for healthy brains, showing that carriers of APOE variants are more vulnerable to changes in this region's functions. At the structural level, regions with higher APOE expression showed greater gray-matter loss in AD. Using this MRI-based data our machine learning model classified AD vs control brains without genetic data, showing that many other factors also drive the disease. Overall, the analysis confirms our hypothesis that APOE is influential at the genetic, microscopic, and structural level, but its causal connection is not certain because of the multitude of factors at each level.

4. Conclusion

Future research should explore ML models' ability to combine information at different levels to better detect early signs of AD. Our analysis shows that signs of AD in the brain occur at

multiple levels, but our ridge model was accurate when looking at only one level. APOE carriers at risk should remember that APOE is only one factor and not at all a guarantee. Focusing on healthy lifestyle habits can help your brain health. and being a participant in AD studies can help the brain health of the future.

5. Data Management Plan (DMP)

This report follows the FAIR principles to ensure proper data management. Findability: All datasets are publicly available: GWAS data from course material, gene expression data from NCBI GEO, MRI data from OASIS, as well as information from Allen Brain Atlas and HENA datasets. Accessibility: Data and results files are stored in common formats (.txt, .csv, .mat) and are clearly labeled for easy access. Interoperability: Variables are matched across datasets, and data is saved in standard table formats compatible with various tools including MATLAB and Python. Reusability: Code is thoroughly commented and separated to make analysis easy to understand and reuse. All patient data is fully anonymized with no personal identifiers. The project uses simple folder structure, separating raw data and scripts, and version control is used to track all changes and allow for reproducibility.

6. References

23andMe. (n.d.). 23andMe DNA testing kit for health + ancestry. https://www.23andme.com/dna-health-ancestry/

Affymetrix. (2006). Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133_Plus_2] (GPL570). NCBI Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570

Allen Institute for Brain Science. (2006). Allen Brain Atlas: Gene Expression. Retrieved May 22, 2025, from https://portal.brain-map.org/gene-expression

Gustavsson, A., Norton, N., Fast, T., Frölich, L., Georges, J., Holzapfel, D., Kirabali, T., Krolak‐Salmon, P., Rossini, P. M., Ferretti, M. T., Lanman, L., Chadha, A. S., & Van Der Flier, W. M. (2022b). Global estimates on the number of persons across the Alzheimer's disease continuum. Alzheimer's & Dementia, 19(2), 658–670. https://doi.org/10.1002/alz.12694

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS). Washington University in St. Louis. https://sites.wustl.edu/oasisbrains/

Sügis, E., Dauvillier, J., Leontjeva, A., Adler, P., Hindie, V., Moncion, T., & Kuznetsov, D. (2019). HENA, heterogeneous network-based data set for Alzheimer's disease. Scientific Data, 6, 151. https://doi.org/10.1038/s41597-019-0152-0

Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nature Communications, 8, 1826. https://doi.org/10.1038/s41467-017-01261-5