Alzheimer's at the Cellular Level: Integrative Mapping from GWAS SNPs to Gene Expression

Author : Simranjit Kaur Kang
INFO-B 692 - Bioinformatics Project
Instructor: Sarath Janga
Luddy School of Informatics, Computing, and Engineering, Indiana University

ABSTRACT

This study investigates the genetic groundworks of Alzheimer's disease (AD) by integrating Genome-Wide Association Studies (GWAS) with single-cell RNA sequencing (scRNA-seq). We mapped genetic variants to specific brain cell types, revealing how these variants contribute to AD at a cellular level. Significant expression changes in astrocytes and microglia were linked with key genes such as APOE and SORL1, illuminating their roles in AD pathology. The application of LD pruning refined our analysis, ensuring the genetic associations identified were robust and independent. Through a novel analytical pipeline, we identified potential therapeutic targets, contributing to the personalized medicine approach in AD treatment. This research underlines the importance of cell-type-specific genetic expressions in understanding AD and sets the stage for future work to expand this analysis further, with plans to incorporate more comprehensive omics data to enrich our understanding of AD pathophysiology.

ABBREVIATIONS

AD: Alzheimer's Disease
GWAS: Genome-Wide Association Studies
scRNA-seq: Single-cell RNA Sequencing
APOE: Apolipoprotein E
SORL1: Sortilin-Related Receptor 1
OPC: Oligodendrocyte Progenitor Cells
LD: Linkage Disequilibrium
EUR: European Ancestry
DEGs: Differentially Expressed Genes
APP: Amyloid Precursor Protein

INTRODUCTION

Alzheimer's disease (AD), a neurodegenerative disorder marked by amyloid-β plaque deposition and tau protein aggregation, leads to synaptic degradation and progressive cognitive decline. Genetic variants significantly influence AD's onset and progression, highlighting the critical role of heredity in the disease (Reiman E.M., 2020)[1]. Andrews et al. (2023) elucidate the complex genetic architecture of AD, revealing new directions for research and a deeper understanding of its genetic underpinnings [2].

Despite these advances, the specific cellular consequences of genetic variations identified by GWAS remain to be fully understood. With most loci identified in European ancestry due to larger sample sizes, there is a compelling need to extend these findings across diverse populations [2]. Single-cell RNA sequencing (scRNA-seq) has revolutionized our ability to explore cell heterogeneity, capturing individual cell expression profiles and revealing cellular-specific responses to disease (Wang S, 2023)[3]. The pioneering work of Mathys et al. (2019) has been instrumental in characterizing the transcriptional landscape of AD across various brain cells, offering fresh perspectives on the disease's cellular diversity [4].

Our study extends this exploration, integrating GWAS with scRNA-seq data to further dissect AD's cellular foundations. We have pinpointed critical roles for cells such as microglia, known for their inflammatory response, and oligodendrocyte progenitor cells (OPCs), key in myelination and neuron support. This comprehensive analysis sheds light on gene expression alterations within these cells and provides novel insights into their roles in the progression of AD.

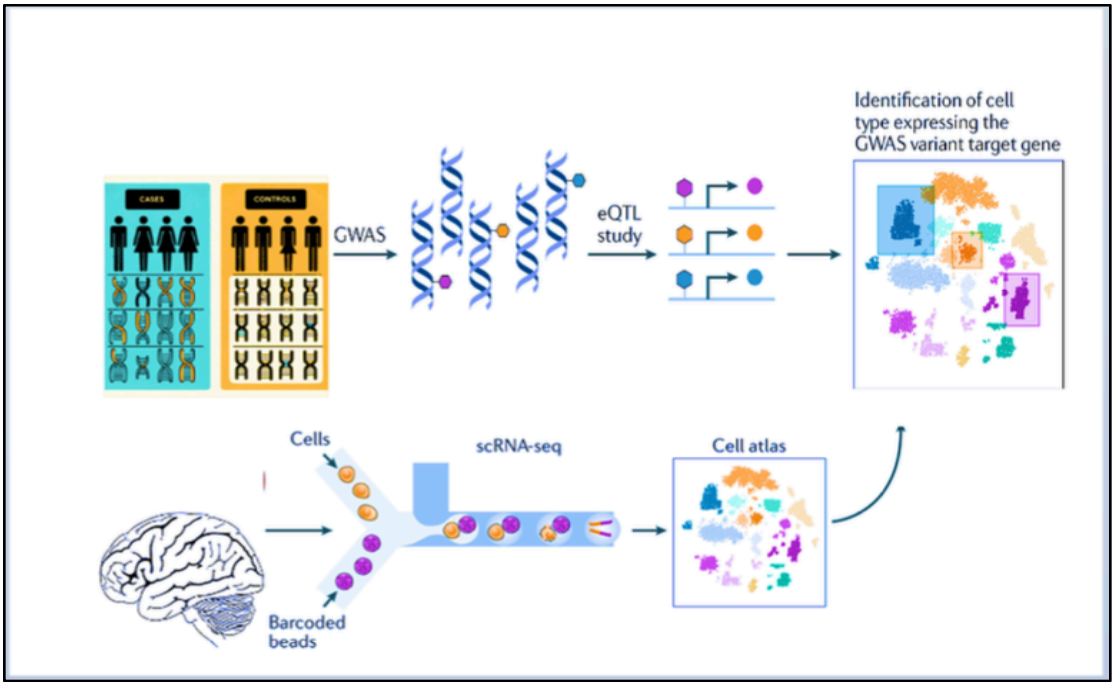


Figure1: This figure illustrates the integrative approach of combining Genome-Wide Association Studies (GWAS) with single-cell RNA sequencing (scRNA-seq) to identify specific cell types that express genes linked to disease variants, contributing to the understanding of complex genetic traits in diseases.

The fusion of GWAS and scRNA-seq data positions us to refine the cellular targets implicated in AD, opening avenues for precision therapeutics. This interaction emphasizes the complexity of glial-neuronal interactions and the pivotal role of cell-specific responses in AD's pathology. Our research represents a significant step towards bridging the gap between genetic susceptibility and cellular dysfunction in AD, potentially transforming the landscape of AD therapeutics.

METHODOLOGY

This section outlines the methodological framework adopted to investigate the genetic foundations of Alzheimer's Disease (AD) through the integration of Genome-Wide Association Studies (GWAS) and Single-cell RNA Sequencing (scRNA-seq) data. We detail the processes of SNP identification, cellular mapping, and the integration of genomic data with cellular function. Our approach is rooted in a precise application of bioinformatics tools for data analysis, stringent statistical validations, and the innovative use of linkage disequilibrium pruning to refine genetic associations. The objective is to present a reproducible and transparent pathway to our findings, ensuring scientific rigor and contributing to the reproducibility of our research.

Datasets

The study utilizes a GWAS Catalog Dataset (GCST007320), comprising data on 24,087 Alzheimer's disease cases and 47,793 individuals with a family history, contrasted with 383,378 controls[6]. This data, derived from genotyping arrays and exome sequencing, offers a comprehensive genetic landscape of Alzheimer's disease. Additionally, the Single-Cell Transcriptomics Dataset (GSE188545) provides an in-depth exploration of cellular responses to Alzheimer's, with 64,845 single-nucleus transcriptomes from the human brain's middle temporal gyrus across 12 subjects. This dataset, accessible on GEO, includes detailed gene expression profiling by high-throughput sequencing, illuminating sex-specific and cell-type-specific gene expression regulation in Alzheimer's disease[7].

Formatting GWAS dataset

In our search of a refined genetic understanding of Alzheimer's disease, we precisely formatted GWAS datasets utilizing the EPIC package in R, aligning with the diverse formats across different associations. This standardization process included a comprehensive suite of 10 essential attributes, from chromosome numbers to minor allele frequencies. Quality control measures were stringently applied, ensuring that our data conformed with the NCBI human genome Build 37 and the 1000 Genomes Project's reference panel, among others, to mitigate any potential discrepancies[8]. Furthermore, to synchronize our GWAS data with the hg38 build used in our single-cell analyses, we employed the liftover package[9], a step critical for maintaining consistency across our genomic datasets. This rigorous approach not only ensures the robustness of our genetic associations but also enhances the interpretability and replicability of our findings.

Mapping SNPs to Genes in Alzheimer's Disease Research

In the quest to unravel the genetic fabric of Alzheimer's Disease (AD), our study has progressed from the raw GWAS data towards a refined gene-centric view. This transformation was achieved by mapping SNPs to genes, a critical step that bridges the gap between wide-scale genetic associations and the cell-specific gene expressions observed in single-cell transcriptomic data. Employing the MAGMA.Celltyping package[10][11], we meticulously matched each SNP to corresponding genes based on the "GRCh37" human genome build. This methodical mapping lays the groundwork for subsequent analyses, including cell-type-specific enrichment tests and the identification of novel therapeutic targets.

Single-Cell Quality Control and Cell Annotation

In advancing Alzheimer's Disease (AD) research, this study implemented rigorous single-cell quality control (QC) and cell annotation protocols. Initial QC involved the exclusion of cells with high mitochondrial gene expression, indicative of compromised membranes, or those with extreme total RNA counts, potentially signifying doublets[12]. Post-filtering, the DoubletFinder[13] package was deployed to further purify the dataset from artificial doublet signatures. Subsequent integration and normalization of the curated single-cell transcriptomes were performed, ensuring the removal of batch effects. Cell annotation was executed using the SingleR method, referencing a consortium of comprehensive brain cell-type atlases[14], thus providing a robust classification of cellular identities. This meticulous approach ensured the downstream analyses were conducted on a refined, high-quality dataset, establishing a reliable foundation for identifying cell-type-specific disease associations and potential therapeutic interventions.

Differential Expression Analysis Using FindMarkers

Our study has advanced the comprehension of gene expression disparities in Alzheimer’s Disease by utilizing the FindMarkers function from the Seurat package for precise identification of genes that vary

between AD-affected and healthy cells. Integrating this with a composite file of conserved marker genes, received for our research, allows us to enhance the genetic mapping accuracy from GWAS data, thereby sharpening the molecular insights into Alzheimer's Disease and contributing to the discovery of genetic factors pivotal to its pathophysiology.

Integration of GWAS and scRNA-seq Data for Alzheimer's Disease

In this phase of our study, we have meticulously integrated the conserved marker genes from single-cell RNA sequencing with GWAS-derived gene data. This was achieved by utilizing the robust MAGMA software, which allowed for the annotation of SNPs to corresponding genes. Our comprehensive method involved cross-referencing genes from the GWAS dataset with the single-cell data, ensuring that genetic variations linked to Alzheimer's Disease could be pinpointed within specific cellular contexts. The integration of these datasets is pivotal for identifying genetic markers that are not just statistically significant, but also biologically relevant to Alzheimer's pathology. The genes.out file from MAGMA provided a detailed list of genes along with their chromosomal positions and statistics, which we matched with gene names obtained from NCBI to ensure accuracy and reliability in our findings. This integrative approach is set to propel our understanding of the disease forward, paving the way for potential therapeutic targets.

Genetic Analysis and Marker Integration for Alzheimer's Research

In our comprehensive analysis, significant differentially expressed genes were identified using stringent criteria. The threshold for adjusted p-values was set at 0.05, and for greater precision, the p-value threshold was adjusted to 0.1 in further analyses. Gene identifiers (GENE IDs) and their corresponding rsIDs were meticulously extracted from a specifically mentioned MAGMA output file (genes.annot). The process culminated in creating distinct files for each cell type, detailing gene names, GENE IDs, and associated rsIDs, derived from the combination of GWAS and scRNA-seq data appropriate to Alzheimer's disease research. This facilitated an enriched understanding of genetic variations within cellular environments, directly associated with the condition.

Linkage Disequilibrium Pruning in GWAS Analysis

Our methodology extends to refining the genetic data through linkage disequilibrium (LD) pruning, utilizing the LDlinkR package[15]. This process selects representative SNPs from correlated clusters, thereby reducing redundancy and improving the clarity of genetic association signals. By setting population-specific parameters and thresholds for correlation and minor allele frequency, we ensure that only the most informative variants are retained for downstream analysis. This step is crucial for distinguishing genuine associations from those confounded by genetic linkage, streamlining the identification of AD-related genetic markers.

Our methodology provides a robust framework for understanding Alzheimer's Disease through an integrated analysis of GWAS and scRNA-seq data, utilizing cutting-edge bioinformatics tools and statistical rigor to identify and validate genetic markers associated with the disease within cellular contexts.

RESULTS

In this section, we share our key findings from exploring Alzheimer's Disease through genetics and cell data. Our results shed light on how genes may influence the disease and suggest new paths for treatments. We'll look at what our data tells us about the disease's puzzle, aiming to make it clear and easy to understand.

GWAS SNP to Gene Conversion Identifies Key Alzheimer's Disease Genes

Our analysis, which utilized MAGMA to convert GWAS SNP data into gene associations, yielded a comprehensive list of 18,199 genes. Of these, 1,658 demonstrated significant associations with Alzheimer's Disease (AD) at a p-value threshold of less than 0.05. This extensive gene collection provides a detailed genetic landscape to enhance our understanding of AD's complexity. Highlighting the relevance of our findings, Figure 2 showcases a heatmap of the top 15 genes that exhibit the highest significance in relation to AD. This visualization not only details each gene's statistical significance, reflected in the -log10(p-values), but also the count of associated SNPs, thereby emphasizing genes such as APOE and BIN1, which have established links to the disease.

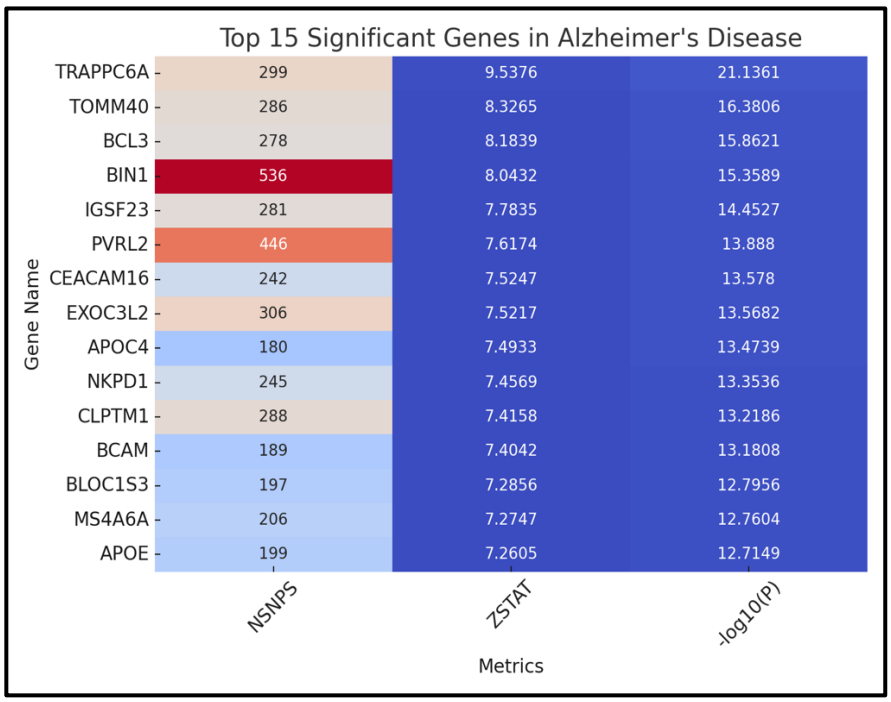


Figure 2: Heatmap of Top 15 Significant Genes in Alzheimer's Disease. The heatmap visualizes genes ranked by their p-value significance and SNP count, emphasizing genes with established and potential roles in AD pathology.

Cellular Insights from Single-Cell RNA Sequencing in Alzheimer's Disease

In-depth single-cell RNA sequencing (scRNA-seq) analysis provided a granular view of the cellular landscape associated with Alzheimer's Disease (AD). Our study processed and integrated scRNA-seq data to discern the transcriptional states of individual cells within AD and healthy control brain tissues. The analysis initially encompassed 40,346 cells, each characterized by a diverse array of 33,538 features. Rigorous quality control measures, including doublet detection and removal, were applied, resulting in a curated dataset of 27,782 cells and 27,393 features. This refinement was crucial to ensure data integrity, enabling the precise characterization of cell types and states relevant to AD pathology. As depicted in Figure 3, the UMAP plot delineates the cellular distribution and highlights the heterogeneity between AD and control samples. Notable is the altered representation of microglia and astrocytes, which are intimately involved in neuroinflammatory responses, potentially implicating their modified roles in AD compared to controls. Such detailed cellular insights are pivotal for unraveling the complexities of AD at the single-cell level and for identifying novel cellular targets for therapeutic intervention.

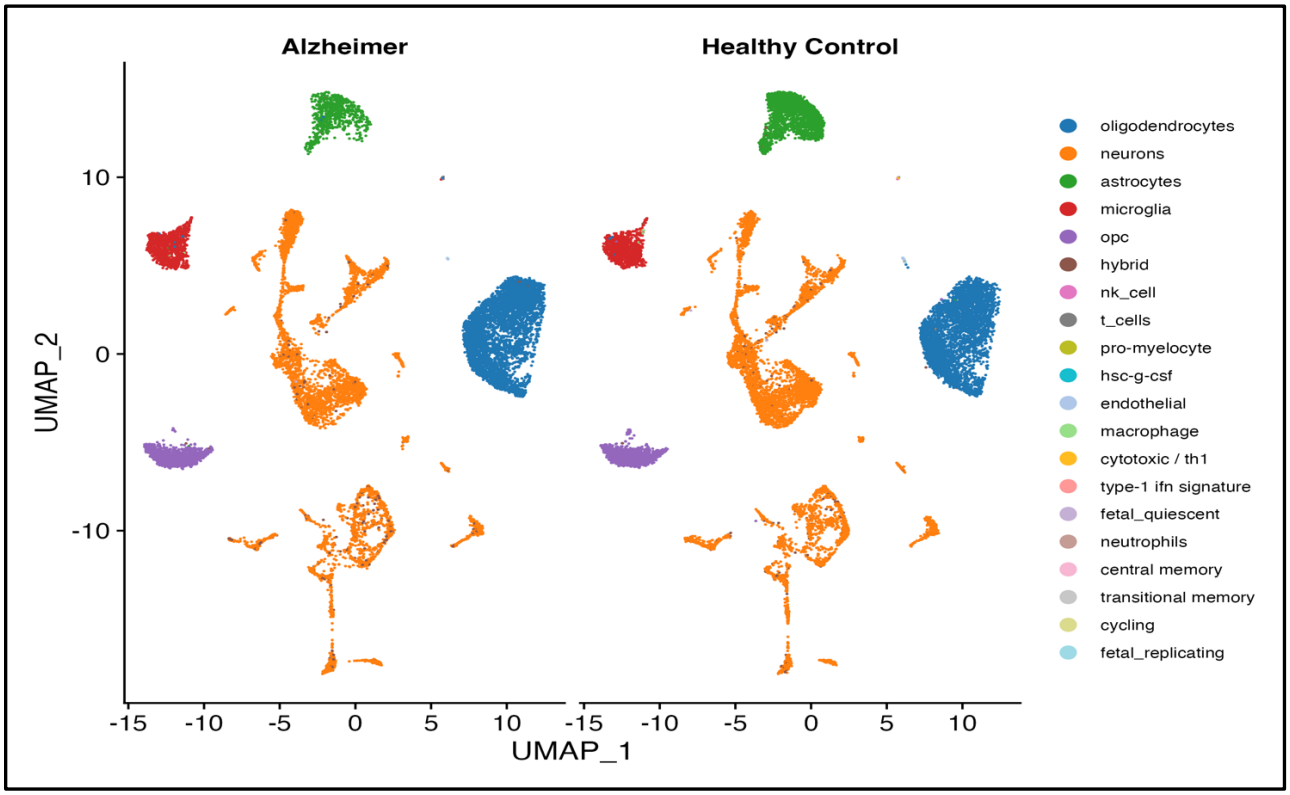


Figure 3: UMAP Clustering of Single-Cell Transcriptomic Data in Alzheimer's Disease Study.

This UMAP plot captures the complex cellular composition of brain tissue from Alzheimer's disease patients compared to healthy individuals. Each point represents a single cell, classified by distinct cell types such as neurons, astrocytes, and microglia, based on their unique gene expression profiles. Notably, the distribution of microglial and astrocytic populations suggests a divergent neuroinflammatory response in Alzheimer's disease, highlighting potential avenues for future therapeutic strategies.

Integration of GWAS Findings with Single-Cell Transcriptomic Data Reveals Cell Type-Specific Gene Expression in Alzheimer's Disease

Our integrated analysis of GWAS data with single-cell RNA sequencing has unearthed cell type-specific gene expression alterations in the Alzheimer's disease (AD) brain. Through meticulous data harmonization, we delineated the gene expression landscape within key neural cell populations implicated in AD pathology. The heatmap in Figure 4 illustrates the log2 fold changes of gene expression across different cell types, including astrocytes, microglia, oligodendrocytes, and OPCs. Notable genes such as APOE and SORL1 exhibit differential expression patterns, suggesting unique roles in cell-specific AD pathophysiological processes. These results bridge the gap between genetic susceptibility and cellular dysfunction, providing a focused lens on the molecular underpinnings of AD.

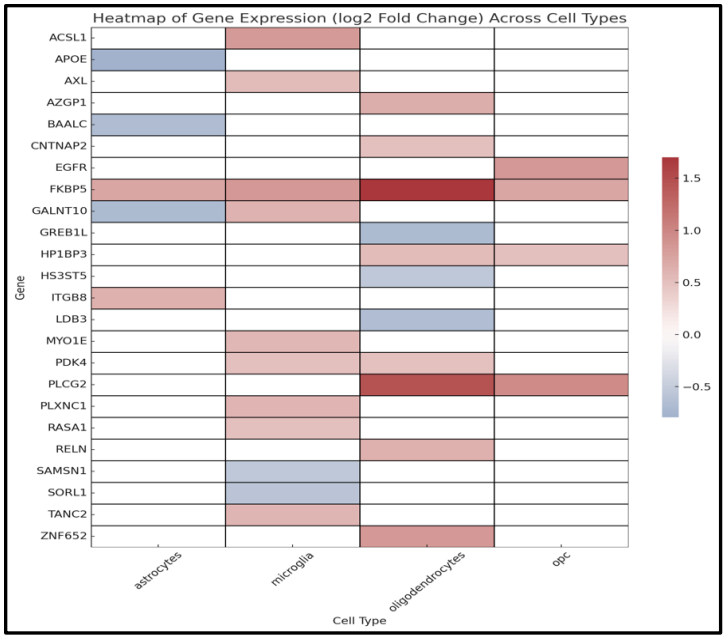


Figure 4: Heatmap of Cell Type-Specific Gene Expression Changes in Alzheimer's Disease. The heatmap displays log2 fold changes of genes across various cell types, correlating with their GWAS significance. The color scale reflects the degree of upregulation (red) and downregulation (blue) of gene expression. Genes like APOE show significant expression changes in astrocytes, while FKBP5 and PLCG2 stand out in microglia and oligodendrocytes, underscoring their potential involvement in AD pathology. This visualization highlights the intricate relationship between genotype and phenotype within the cellular milieu of AD, providing insights into the disease's complex biology.

Impact of LD Pruning on rsID Counts by Cell Type in Alzheimer's Disease

Cell Type	Gene Name	rsID Count Before Pruning	rsID Count After Pruning
Astrocytes	APOE	199	25
Astrocytes	BAALC	659	46
Astrocytes	FKBP5	661	35
Astrocytes	GALNT10	1237	65
Astrocytes	ITGB8	521	37
Microglia	ACSL1	603	43
Microglia	AXL	312	33
Microglia	FKBP5	661	35
Microglia	GALNT10	1237	65
Microglia	MYO1E	1151	82
Microglia	PDK4	278	19
Microglia	PLXNC1	843	65
Microglia	RASA1	482	34
Microglia	SAMSN1	636	54
Microglia	SORL1	571	52
Microglia	TANC2	1051	42

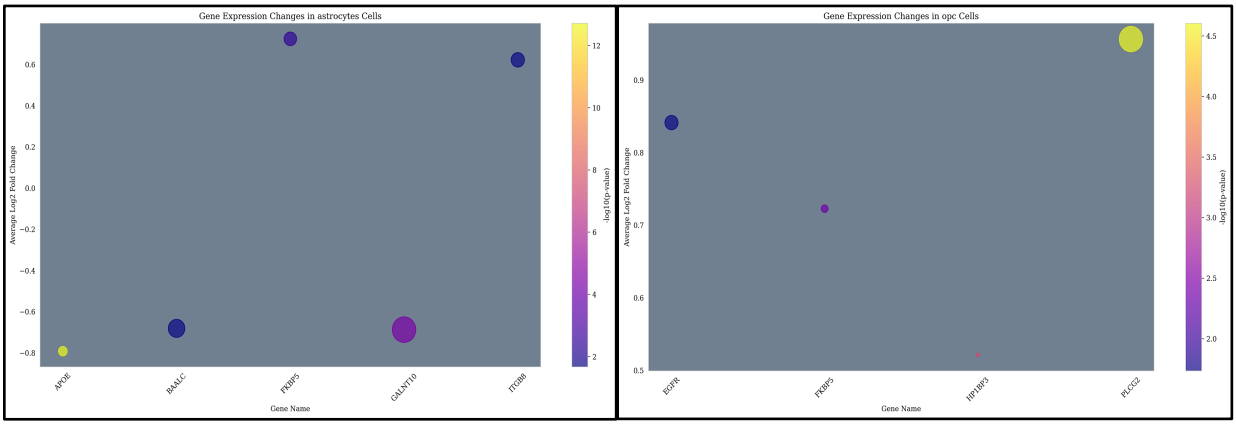
Cell Type	Gene Name	rsID Count Before Pruning	rsID Count After Pruning
Oligodendrocytes	AZGP1	150	21
Oligodendrocytes	CNTNAP2	11341	6
Oligodendrocytes	FKBP5	661	35
Oligodendrocytes	GREB1L	433	30
Oligodendrocytes	HP1BP3	234	19
Oligodendrocytes	HS3ST5	893	57
Oligodendrocytes	LDB3	596	35
Oligodendrocytes	PDK4	278	19
Oligodendrocytes	PLCG2	1745	117
Oligodendrocytes	RELN	2648	119
Oligodendrocytes	ZNF652	421	21
OPC	EGFR	941	66
OPC	FKBP5	661	35
OPC	HP1BP3	234	19
OPC	PLCG2	1745	117

Table 1 It provides a clear demonstration of the impact of linkage disequilibrium (LD) pruning on our dataset. For each cell type, we list the genes of interest alongside the rsID counts before and after pruning. Notably, the rsID count for the APOE gene in astrocytes dropped significantly from 199 to 25 post-pruning, underscoring the gene's potential importance in Alzheimer's disease pathology. Similar reductions across other key genes such as FKBP5 and GALNT10 across various cell types also reflect the targeted nature of our analysis post-pruning, ensuring the genetic variants we consider are likely to be the most relevant to the disease.

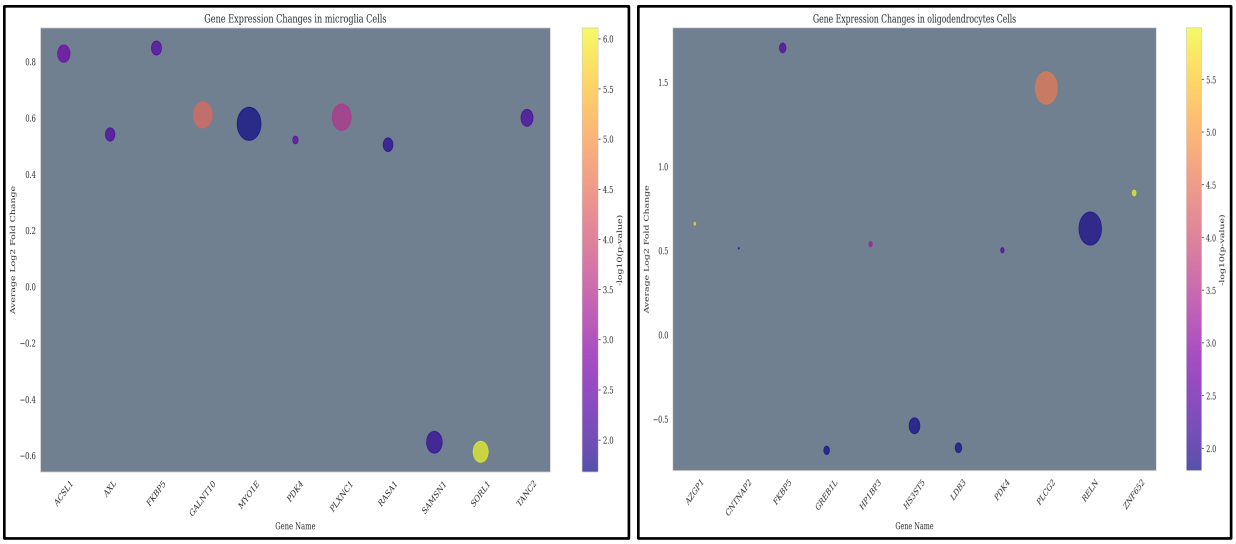
GWAS Signals and Cellular Expression in Alzheimer's Disease

Our analysis presents a consolidated view of Alzheimer's Disease (AD) by integrating GWAS signals with expression data from neural cells. We identify key genes with significant expression alterations and genetic associations, highlighting their potential impact on AD pathology. The following figures illustrate these relationships, offering insights into the genetic and cellular dimensions of AD.

ASTROCYTES (A) **OPC. (B)**



MICROGLIA (C) **OLIGODENDROCYTES. (D)**



Figures 5 (A – D): Combined Analysis of GWAS Significance and Gene Expression in Neural Cell Types

The series of figures from 5 to 9 provide an insightful correlation between the genetic predisposition and the expression changes of genes in different neural cell types, including astrocytes, OPCs, microglia, and oligodendrocytes. The size of the dots in these figures indicates the number of associated SNPs and the significance of GWAS P-values, with larger dots denoting a stronger genetic association with Alzheimer's Disease. Concurrently, the color gradient illustrates the direction and magnitude of log2 fold changes, revealing genes that are upregulated or downregulated.

In astrocytes, genes like APOE show significant expression changes, despite a lower fold change, suggesting a robust genetic association with Alzheimer's Disease.
Within OPC cells, EGFR exhibits a notable expression change, potentially influencing progenitor cell behavior, while PLCG2 in oligodendrocytes stands out for both expression changes and genetic significance, hinting at its impact on myelination.
The microglia cell type reveals genes such as SORL1 and TANC2 with prominent expression changes that may be linked to the immune response and synaptic function in Alzheimer's pathology.
These visualizations underscore the intricate relationship between the genetic landscape and cellular expression profiles in the context of Alzheimer's Disease, providing invaluable insights into the molecular mechanisms underpinning this complex condition.

DISCUSSION

Our study offers a novel insight into the cellular mechanisms underpinning Alzheimer's Disease (AD) through the integration of Genome-Wide Association Studies (GWAS) and single-cell RNA sequencing (scRNA-seq). The identification of key genes like APOE and SORL1 across specific neural cell types underscores the intricate interplay between genetic susceptibility and cellular function in AD pathology. The substantial reduction in rsID count post-LD pruning for these genes confirms their robust association with AD and supports their potential as therapeutic targets.

While our findings are promising, there are limitations to consider. The study's reliance on data from predominantly European ancestry might limit the generalizability of the results across different populations.

Moreover, the cross-sectional nature of this data precludes insights into the temporal progression of gene expression changes in AD.

Future research should aim to include more diverse populations to build a more inclusive genetic profile of AD. Longitudinal studies could provide valuable information on the dynamics of gene expression changes over the course of the disease. Additionally, functional studies are required to confirm the mechanistic roles suggested by our genetic and cellular findings.

By illuminating the cellular context of genetic associations in AD, our study contributes to the precision medicine approach, potentially leading to tailored therapies that target specific cell types or pathways implicated in the disease process.

Supplementary Materials

For access to the complete dataset, analytical methods, and scripts used in this study, please refer to our GitHub repository: https://github.com/SKVirk27/Alzheimer-s-at-the-Cellular-Level-Integrative-Mapping-from-GWAS-SNPs-to-Gene-Expression/blob/main/README.md

References

1. Reiman, E.M., et al. (2020). Exceptionally low likelihood of Alzheimer's dementia in APOE2 homozygotes from a 5,000-person neuropathological study. Nat Commun, 11, 667. [PMC free article] [PubMed]

2. Andrews, S. J., Renton, A. E., Fulton-Howard, B., Podlesny-Drabiniok, A., Marcora, E., & Goate, A. M. (2023). The complex genetic architecture of Alzheimer's disease: novel insights and future directions. EBioMedicine, 90, 104511. https://doi.org/10.1016/j.ebiom.2023.104511

3. Wang S, Sun ST, Zhang XY, Ding HR, Yuan Y, He JJ, Wang MS, Yang B, Li YB. (2023). The Evolution of Single-Cell RNA Sequencing Technology and Application: Progress and Perspectives. Int J Mol Sci, 24(3):2943. https://doi.org/10.3390/ijms24032943

4. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, Martorell AJ, Ransohoff RM, Hafler BP, Bennett DA, Kellis M, Tsai LH. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature, 570(7761):332-337. https://doi.org/10.1038/s41586-019-1195-2

5. Dato S, De Rango F, Crocco P, Pallotti S, Belloy ME, Le Guen Y, Greicius MD, Passarino G, Rose G, Napolioni V. (2023). Sex- and APOE-specific genetic risk factors for late-onset Alzheimer's disease: Evidence from gene-gene interaction of longevity-related loci. Aging Cell, 22(9):e13938. https://doi.org/10.1111/acel.13938

6. Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., ... & Andreassen, O. A. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature Genetics, 51(3), 404-413. https://doi.org/10.1038/s41588-018-0311-9

7. Zhang, L., He, C. H., Coffey, S., Yin, D., Hsu, I. U., Su, C., ... & Strittmatter, S. M. (2023). Single-cell transcriptomic atlas of Alzheimer's disease middle temporal gyrus reveals region, cell type and sex

specificity of gene expression with novel genetic risk for MERTK in female. medRxiv. https://doi.org/10.1101/2023.02.18.23286037

8. Wang R, Lin D, Jiang Y. EPIC: inferring relevant tissues and cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing. PLOS Genetics, 2022.

9. MRC Integrative Epidemiology Unit. (n.d.). IEU GWAS R package. Retrieved from https://mrcieu.github.io/ieugwasr/.

10. Skene, N. G., Bryois, J., Bakken, T. E., et al. (2018). Genetic identification of brain cell types underlying schizophrenia. Nature Genetics, 50(6), 825-833. https://doi.org/10.1038/s41588-018-0129-5

11. de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. PLOS Computational Biology, 11(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

12. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology, 36(5), 411–420. https://doi.org/10.1038/nbt.4096

13. McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Systems, 8(4), 329–337.e4. https://doi.org/10.1016/j.cels.2019.03.003

14. Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., ... & Butte, A. J. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology, 20(2), 163–172. https://doi.org/10.1038/s41590-018-0276-y

15. Myers, T. A., Chanock, S. J., & Machiela, M. J. (2023). LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. LDlinkR Package Documentation. Retrieved May 31, 2023