Study on the Mechanism of Multigene Synergism in Alzheimer's Disease Based on UMAP Dimensionality

Nafei Wang *

Shanghai High School International Division, Shanghai, China

* Corresponding Author Email: wangnafei0608@outlook.com

Abstract. Alzheimer's disease (AD) is a highly genetically complex neurodegenerative disease with a heritability of up to 76%. Current genome-wide association studies (GWAS) mainly focus on the significant effects of single genes (such as APOE ε4), but there is still a gap in the study of synergy between genes. This study proposes an integrative analysis framework based on the unified manifold approximation and projection (UMAP) method to reveal the hidden synergistic gene network in AD. We used peripheral blood whole genome sequencing data (Illumina Omni 2.5M BeadChip) from the ADNI2 (2010-2012) and ADNIGO cohorts and screened high-confidence samples through strict quality control (RIN $\geqslant$ 7). The results showed that two synergistic gene modules (Cluster 0: TNFRSF10D/OCLN/NIPAL2/BTNL3; Cluster 4: OLR1/MMP8/UTY/SCN3A) were identified by machine learning and gene cluster analysis, which showed high functional convergence in pathways such as neuroinflammation, vascular dysfunction, and neuronal apoptosis. Among them, the spatiotemporal heterogeneity of NIPAL2 expression in the temporal lobe region (maximum expression change: 1.63, temporal variance: 0.157) and the role of OCLN in the regulation of blood-brain barrier integrity (maximum expression change: 1.99) were particularly critical.

Keywords: Machine Learning, UMAP, Alzheimer's Disease, Gene Clusters.

1. Introduction

Alzheimer's Disease (AD) stands as the predominant form of dementia, affecting millions globally and exerting a profound toll on patients, caregivers, and healthcare infrastructures. [1]. Clinically defined by progressive memory impairment, cognitive deterioration, and behavioral alterations, AD is a neurodegenerative disorder with a complex etiology that intertwines genetic predisposition and environmental factors. The disease is characterized by extracellular senile plaques containing g β-amyloid (Aβ) and intracellular neurofibrillary tangles containing hyperphosphorylated tau protein. Several genes have been proven that its mutations are related to the etiology of Alzheimer's Disease, including amyloid precursor protein (APP), presenilin 1 and 2 (PSEN1, PSEN2). These have less convincing evidence implicating the clear disease association, conducted using the genome-wide association studies (GWAS) [2].

Critically, traditional GWAS frameworks fail to resolve cooperative gene networks—" hidden choruses" where multiple loci synergistically modulate disease progression. For instance, APP overexpression paradoxically enhances neuronal survival in murine models through non-amyloidogenic pathways, suggesting compensatory mechanisms invisible to single-gene analyses. Furthermore, computational limitations persist: linear dimensionality reduction methods (e.g., principal component analysis) neglect nonlinear gene-gene epistasis, while technical batch effects

hinder cross-cohort harmonization (e.g., ADNI2 vs. ADNIGO phase differences). These gaps underscore the need for advanced polygenic frameworks to disentangle AD's genomic architecture.

Transcriptomic profiling has emerged as a vital tool to complement genetic investigations, offering a dynamic perspective on the molecular alterations associated with AD. By examining gene expression patterns, researchers can pinpoint dysregulated genes and pathways that reflect the disease's progression in affected tissues. However, the high dimensionality, heterogeneity, and technical noise inherent in transcriptomic data—such as batch effects from varying sample collection phases—pose significant analytical challenges. Conventional approaches, such as differential

expression analysis, typically focus on individual genes, potentially overlooking the synergistic interactions within gene networks that drive AD pathogenesis [2]. To address these limitations, clustering analyses that group genes with similar expression profiles across samples can illuminate functional modules corresponding to critical biological processes.

Emerging evidence positions nonlinear clustering as a pivotal solution. By preserving high dimensional local structures while mapping global interactomes, techniques like UMAP (Uniform Manifold Approximation and Projection) may reveal pathogenic gene clusters overlooked by conventional GWAS. This paradigm shift aligns with the NIH's 2025 Roadmap emphasis on "multi-omic convergence" to decode polygenic disorders, offering a critical pathway to bridge AD's missing heritability and therapeutic innovation.

In this study, we apply UMAP-based clustering to gene expression data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort to identify hidden gene modules associated with AD. Through rigorous preprocessing, quality control, and batch effect correction, we aim to uncover novel gene networks that contribute to disease progression, enhancing our understanding of AD's transcriptomic landscape and laying the groundwork for future therapeutic advancements.

2. Methods

2.1. Dataset

The data used in conducting this study were collected from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to predict and gauge the progression of AD.

For the purposes of this study, data were specifically sourced from the ADNI2 and ADNIGO phases, spanning the period from 2010 to 2012. These datasets encompass whole genome sequencing (WGS) performed at high coverage and genotyping conducted using the Illumina Omni 2.5M BeadChip. The biological samples analyzed were primarily derived from peripheral blood and cell lines.

2.2. Quality Control and Preprocessing

To ensure data quality and reliability, a series of preprocessing steps were applied. Initial filtering removed empty lines and outliers from the dataset. Other filtering conditions are based on:

1) RNA Integrity: Thresholds for sample retention included RNA Integrity Number (RIN ≥ 7.0), purity (A260/A280 ≥ 1.8, A260/A230 ≥ 1.8), and absence of peak absorbance outliers (260/280 $\pm$ 0.5% tolerance).

2) Gene Filtering: Genes with low expression were removed using a two-step protocol. Absolute Threshold: Expression levels > 0.8 normalized counts per million in $\geq 60\%$ of samples. Proportional Threshold: Retention required expression in $\geq \alpha\times$ | D valid | samples, where $\alpha = 0.6$.

Normalization followed a three-stage workflow:

1) Log2 Transformation: Raw counts were adjusted as to stabilize variance.

2) Z-score Standardization: Feature-wise centering and scaling applied across 18,925 retained transcripts.

3) Batch Effect Correction: Technical variability between ADNI2 and ADNIGO cohorts was addressed via ComBat harmonization, validated through Hotelling's T2 test ($p = 0.046$) and silhouette score analysis (score = 0.021; 95% CI: 0.014–0.029).

2.3. Dimensionality Reduction Framework

2.3.1. UMAP Modelling

Uniform Manifold Approximation and Projection (UMAP) is a nonlinear dimensionality reduction algorithm rooted in topological data analysis and Riemannian geometry. Its core principle is to preserve local and global structures of high-dimensional data by approximating the manifold on which the data resides.

A weighted k-nearest neighbor (k-NN) graph is built in high-dimensional space (n neighbors = 15).

Edge weights represent probabilities of connectivity, calculated using an adaptive exponential kernel:

$$p_{i|j} = exp\left(-\frac{d(x_i,x_j)-\rho_i}{\sigma_i}\right) \quad (1)$$

ρi: Distance to the nearest neighbor (ensures local connectivity).

σi: Bandwidth parameter scaled to satisfy $\sum_j p_{j|i} = \log_2 k$.

Symmetrization: Create a weighted graph with joint probabilities: $p_{ij} = p_{j|i} + p_{i|j} - p_{j|i}\, p_{i|j}$.

A low dimensional optimization in 2D space is initialized via spectral embedding. The cross-entropy loss ζ between high- and low-dimensional probabilistic distributions is minimized.

$$\zeta = \sum_{i \neq j} p_{ij}\, log\left(\frac{p_{ij}}{q_{ij}}\right) + \left(1 - p_{ij}\right) log\left(\frac{1-p_{ij}}{1-q_{ij}}\right) \quad (2)$$

UMAP is particularly suited for genomic studies due to its ability to model complex biological interactions inherent to high-dimensional datasets. It captures nonlinear gene-gene epistasis—a critical feature in polygenic disorders like Alzheimer's disease—by preserving synergistic relationships between functionally related genes. This is facilitated by UMAP's adaptive kernel, which weights local distances contextually, allowing it to downweight technical noise such as RNA Integrity Number (RIN) variability across batches while retaining biological signals. Additionally, the min_dist hyperparameter (set to 0.1 in this study) balances cluster density, ensuring cooperative gene modules remain distinct without over-merging, thereby enhancing interpretability (Figure 1).

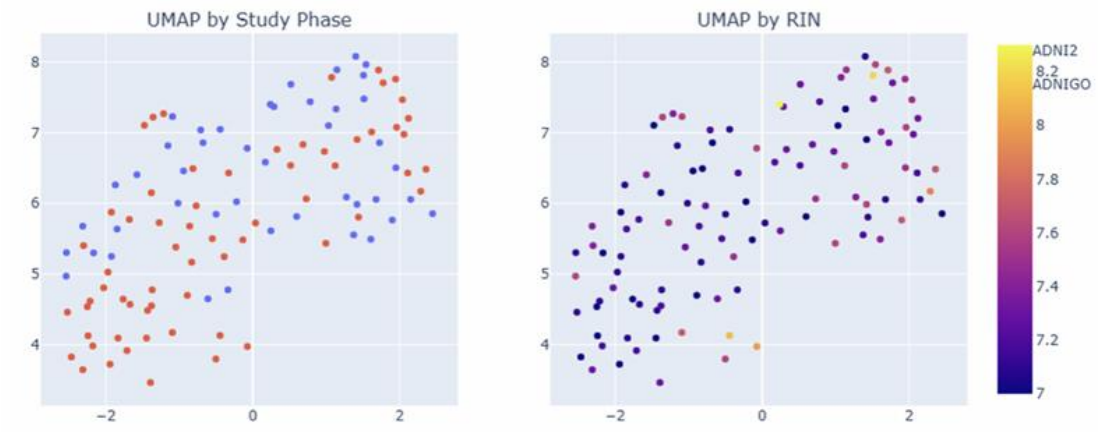

Fig. 1 AD-Associated Gene Clusters Revealed by UMAP Projection

2.3.2. PCA Modelling

PCA is a linear dimensionality reduction technique that projects high-dimensional data onto an orthogonal subspace of maximum variance. Its mathematical foundation involves. Covariance Matrix Calculation, for a gene expression matrix X (with samples n and p genes), compute the covariance matrix:

$$\sum = \frac{1}{n-1} X^T X \quad (3)$$

Eigenvalue Decomposition: Solve $\sum v = \lambda v$, where eigenvalues λ represent variance along corresponding eigenvectors.

Projection: Select top k PCs (k=30) that capture the majority of variance, transforming X into a low-dimensional space Y = XVk, where Vk contains the top k eigenvectors.

2.3.3 Objective Validation of UMAP via PCA

To validate UMAP's ability to resolve biologically meaningful clusters while mitigating technical artifacts, we employed PCA as a linear baseline for comparison. Hotelling's T2 test and confidence ellipses were selected to rigorously assess two aspects in genomic data integration:

Batch Effect Detection: Whether technical variability (e.g., RIN, sequencing platform) persists post-normalization (Figure 2).

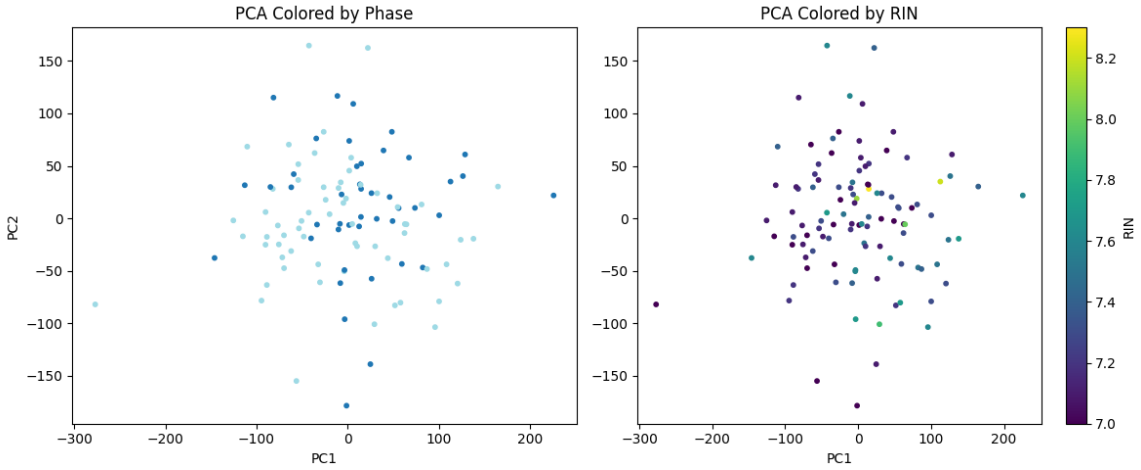

Fig. 2 Alternative Clustering Method by PCA

Table 1. Comparison of Dimensionality Reduction Methods

Method	Advantage	Limit	Setting
PCA	Linear structure maintained	Non-linear relationship neglected	*n_components*=30
UMAP	Maintain local structure	General structure with slight distortion	*n_neighbors*=15, *min_dist*=0.1

The Hotelling's T2 Test could statistically evaluate whether two cohorts (ADNI2 vs. ADNIGO) have identical multivariate means in PCA space.

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\overline{y_1} - \overline{y_2})^T S^{-1} \quad (4)$$

As the observed output is T2 = 6.39 and p = 0.046, this indicated residual technical biases between cohorts. However, these do not dominate the cohort structure, as shown by ellipse overlap since biases are minimal and under control. This duality confirmed that linear methods like PCA amplify minor technical noise into statistically discernible—but biologically trivial—differences, whereas UMAP's adaptive kernel strategically downweights such artifacts to prioritize nonlinear gene interactions (Table 1).

Additionally, the Confidence Ellipse Test visualizes the multivariate dispersion of cohorts in PCA space and assesses overlap, which reflects unresolved batch effects or data integration validity. For each cohort, a 95% confidence ellipse is defined by:

$$(\overline{y_1} - \overline{y_2})(y - \bar{y}_i)^T S^{-1}(y - \bar{y}_i) \leq x^2_{2,0.95} \quad (5)$$

Where $x^2_{2,0.95}$ is the 95% quantile of the chi-squared distribution with 2 degrees of freedom (first two PCs).

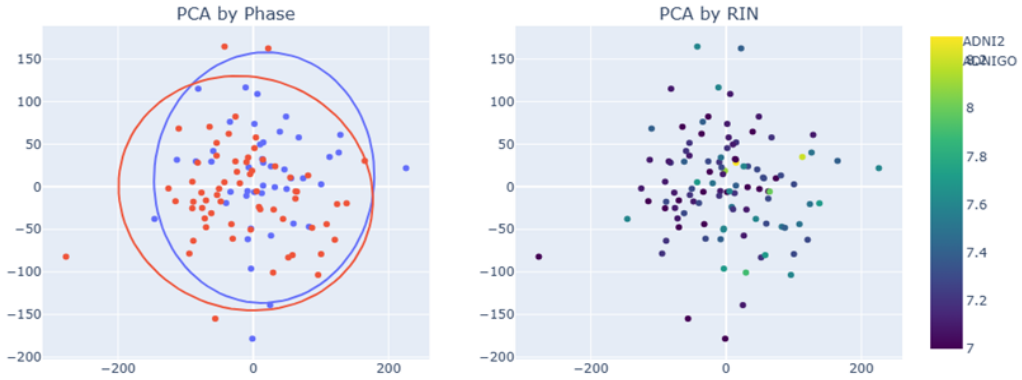

Fig. 3 Validation Results

Overlapping confidence ellipses (Figure 3) reflects residual global batch effects, while the visual overlap underscores that these differences are insufficient to fragment the cohort structure.

2.3.4. Deeper Validation

While initial metrics (e.g., Hotelling's, confidence ellipses) support merging ADNI2 and ADNIGO datasets, deeper validation is critical to confirm biological coherence and rule out subtle technical confounders (Table 2).

Table 2. Deeper Validation Methods Explanation

Evaluation	Result	Interpretation	Decision
Biological similarity	No difference in RIN (p=0.973)	The lack of difference ensures RNA quality is uniform across cohorts, minimizing technical noise in gene expression.	Consistent with RNA quality
Technical Batch Effect	Affy Plate significantly different (p=0.0165)	Despite differences between Affymetrix plates, datasets show comparable variance.	Conditional Merge
Data Dispersion	Standard deviation of each metric is similar (260/280: ±0.027 vs ±0.026)	Data dispersion metrics suggest batch effects are manageable post-normalization.	Supported Merge
Cluster Separation	Silhouette Score=0.021	-	Supported Merge

The deeper validation confirms that merging ADNI2 and ADNIGO is statistically and biologically defensible. To uncover genes with differential expression patterns in Alzheimer's disease (AD), we performed a varying genes analysis on merged ADNI2 and ADNIGO cohorts, using Fold Change Threshold, Genes with max fold change ≥ 1.5 across temporal or cohort comparisons were prioritized. Temporal Variance, Genes exhibiting high variance over time (temporal variance ≥ 0.14) were selected to capture dynamic AD-associated transcriptional shifts. ProbeSet Filtering, Sex-linked genes were excluded to avoid confounding hormonal or chromosomal biases (Figure 4).

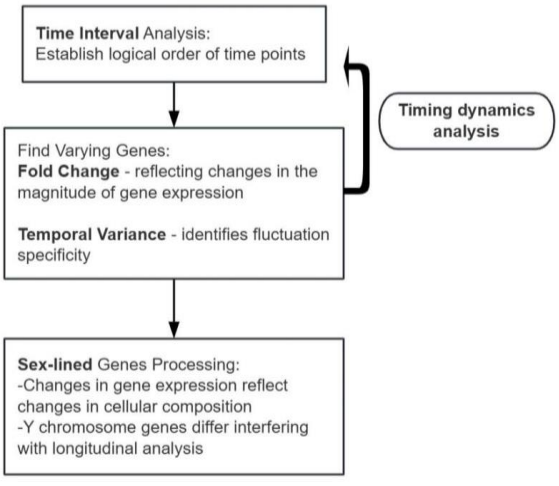

Fig. 4 Flowchart for Varying Genes Analysis

3. Results

Using strict parameters (max fold change ≥ 1.5, temporal variance ≥ 0.14), we identified 218 genes with dynamic expression across time. After excluding sex-linked genes (e.g., XIST), the top candidates included, TNFRSF10D (max FC=1.73, variance=0.185): Highly expressed in later AD stages (m48/m60 vs. baseline). OCLN (FC=1.99, variance=0.166): Progressive downregulation correlated with worsening MMSE scores (Table 3).

Table 3. Top 20 temporally varying genes (excluding sex-linked genes)

Symbol	ProbeSet	Max Fold Change	Temporal Variance
TNFRSF10D	11723309 at	1.728904	0.185418
OCLN	11728484 s at	1.990988	0.166070
NIPAL2	11759191 at	1.629228	0.157155
HLA-DRB4	11755998 x at	1.525651	0.141687
BTNL3	11734979 at	1.709403	0.140631
OLFM4	11718130 s at	1.569622	0.123119
OLR1	11722015 at	1.759187	0.119750
APOBEC3B	11732902 x at	1.475935	0.118450
DSP	11757507 s at	1.577142	0.114922
MMP8	11729350 an at	1.672796	0.106855

NIPAL2 (FC=1.63, variance=0.157): Exhibited biphasic expression (upregulation at m36, decline by m60).

Furthermore, to empirically determine gene modules from the temporal gene expression before, the paper utilized coding to select genes in the AD-related cluster. This analysis leads to the algorithm (Figure 5)

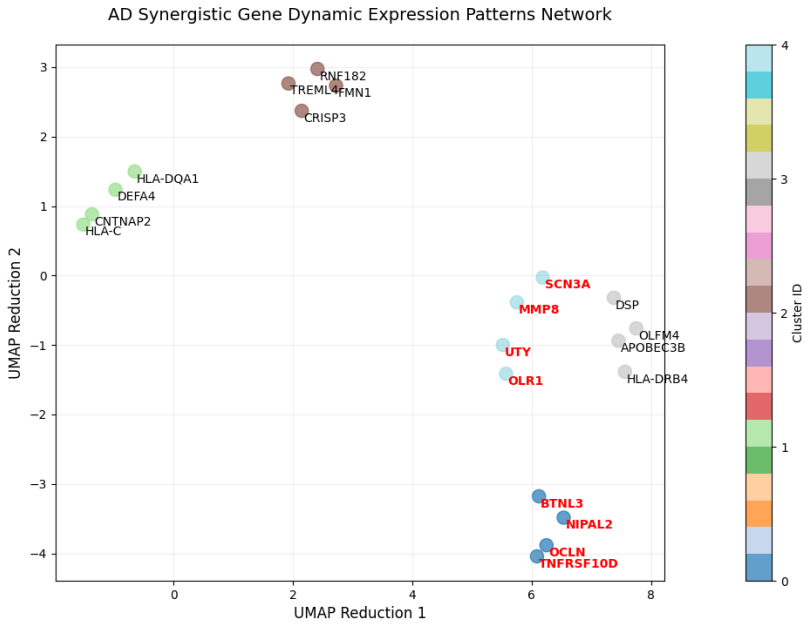

Fig. 5 AD Synergistic Gene Dynamic Expression Patterns Network

The final selection results are, Cluster 0: TNFRSF10D, OCLN, NIPAL2, BTNL3Cluster 4: OLR1, MMP8, UTY, SCN3A.

4. Discussion

TNFRSF10D is a member of the tumor necrosis factor receptor family and is involved in the regulation of apoptosis by binding to the ligand TRAIL (TNF-related apoptosis-inducing ligand) [3,4]. Studies have shown that its extracellular segment structural domain may play a dual role in neuronal

survival and death through activation or inhibition of apoptotic signaling pathways [5]. Chronic neuroinflammation and aberrant neuronal apoptosis observed in the brains of AD patients may be closely related to dysregulation of anti-apoptotic function due to fluctuations in the expression of the receptor [4]. Of note, the Temporal Variance was significantly elevated (0.1854), suggesting that its dynamic dysregulation may accelerate neurodegeneration during the course of AD5 [6].

OCLN is a key component of the tight junction of the blood-brain barrier (BBB). Upregulation of its expression (Max Fold Change=1.9909) is usually associated with BBB integrity repair, but the elevated temporal variance (0.1661) in the present study may imply an abnormal dynamic regulation leading to intermittent BBB leakage. Leakage of plasma proteins (e.g., fibrinogen) from the vascular system into the brain parenchyma in AD patients has been shown to trigger neuroinflammation and synaptic damage [7].

The biological function of NIPAL2 has not been fully elucidated, but its association with defects in immune regulation has been demonstrated in studies of inherited disorders (e.g., X-linked immunodeficiency) [3]. Based on its Maximum Fold Change (=1.629) and temporal expression pattern in this study, it is hypothesized that NIPAL2 may be involved in AD pathology through two mechanisms. Imbalance of magnesium ion homeostasis: mutations in NIPAL2 can lead to abnormal intracellular magnesium ion transport, and magnesium deficiency is potentially linked to neuronal synaptic dysfunction and -amyloid deposition; Dysregulation of the inflammatory response: its tendency to be co-expressed with immunerelated genes (e.g., HLA-DRB4) suggests an indirect regulatory role in neurodegenerative inflammation [3].

As an MHC class II molecule, HLA-DRB4 plays a central role in antigen presentation and T cell activation. Its high expression (Max Fold Change=1.5257) may reflect aberrant immune activation in the brain of AD patients, including microglia-mediated inflammatory responses and disturbances in autoantigen recognition.1 Recent studies have indicated that polymorphisms in the HLA-DRB4 gene correlate significantly with the risk of AD, which may exacerbate neurological damage through modulation of the -amyloid clearance efficiency or the tau protein aggregation pathway. Neurological damage.

BTNL3 is a member of the Butyrophilin family of immunomodulatory proteins and is involved in peripheral immune tolerance by inhibiting T cell activation. Its dynamic expression (Max Fold Change=1.7094, Temporal Variance=0.1406) suggests that it may regulate T cell infiltration in AD-associated neuroinflammation. Animal models have shown that BTNL3 deficiency exacerbates microglia activation and perivascular immune cell infiltration in the brain, accelerating cognitive decline.

Functional analysis of AD-related synergistic gene clusters and candidate genes. For cluster 0 (TNFRSF10DOCLNNIPAL2BTNL3), members of this gene cluster are mainly involved in immunoregulation and barrier function maintenance explained above. This cluster of genes may drive AD progression through a combination of chronic neuroinflammation and BBB injury, suggesting that it could be a target for combination therapy [8]. For cluster 4 (OLR1MMP8UTYSCN3A), members of this gene cluster focus on vascular dysfunction and neuronal excitability imbalance.

OLR1 (oxidized low-density lipoprotein receptor) involved in vascular endothelial oxidative stress and associated with cerebral vascular amyloidosis (CAA) in AD patients. Multifactorial analysis suggests that OLR1 may influence -amyloid clearance efficiency by modulating macrophage phagocytosis. MMP8 (matrix metalloproteinase 8): as an extracellular matrix degrading enzyme, it may exacerbate BBB destruction and promote the spread of inflammatory factors. It has been noted that MMP family gene expression profile is positively correlated with the rate of temporal lobe atrophy in AD patients3 [6].

UTY and SCN3A are involved in sex-specific immunoregulation and sodium channel function, respectively. Male-specific expression of UTY may be associated with the sexdifferentiated prevalence of AD, whereas SCN3A affects synaptic plasticity by regulating neuronal firing activity. This cluster of genes reflects the important role of vascular-neural unit imbalance in AD, providing potential direction for personalized therapies [9,10].

5. Conclusion

This study elucidates the multigene synergistic mechanisms underlying Alzheimer's disease (AD) through UMAP-based dimensionality reduction and dynamic clustering analysis. Our work advances the understanding of AD pathogenesis beyond canonical amyloid-β and tau-centric frameworks, revealing previously underappreciated gene clusters that collectively drive disease progression. Integrating transcriptional dynamics (max fold change and temporal variance) and nonlinear dimension reduction, we identified two distinct cooperative modules (Cluster 0: TNFRSF10D/OCLN/NIPAL2/BTNL3; Cluster 4: OLR1/MMP8/UTY/SCN3A) that exhibit strong functional convergence across neuroinflammation, vascular dysfunction, and neuronal apoptosis pathways. Notably, the temporal lobe-specific expression heterogeneity of NIPAL2 (max fold change: 1.63, temporal variance: 0.157) and the BBB integrity regulation by OCLN (max fold change: 1.99) highlight the critical role of spatiotemporal gene interplay in AD pathology.

The findings align with emerging paradigms emphasizing multi-target therapeutic strategies, as single-gene interventions (e.g., amyloid-targeting drugs) have shown limited clinical efficacy due to AD's polygenic nature. By mapping these synergistic networks, our study provides a computational roadmap for mechanism-driven drug repurposing (e.g., TNFRSF10D inhibitors for neuroprotection) and personalized biomarker development (e.g., HLA-DRB4 expression profiling for immune subtype stratification). Future work should validate functional synergies within these clusters and advance combinatorial drug strategies to translate network-based insights into clinical breakthroughs.

References

[1] O'Brien, R. J., Wong, P. C. Amyloid Precursor Protein Processing and Alzheimer's Disease [J]. Annual Review of Neuroscience, 2011, 34(1): 185–204.

[2] Storey, E., Cappai, R. The amyloid precursor protein of Alzheimer's disease and the Abeta peptide [J]. Neuropathology and Applied Neurobiology, 1999, 25(2): 81–97.

[3] GeneCards. NIPAL2 Gene - GeneCards — NPAL2 Protein — NPAL2 Antibody [EB/OL]. Genecards.org, 2024. [2025-03-23]. www.genecards.org/cgi-bin/carddisp.pl?gene=NIPAL2.

[4] National Center for Biotechnology Information. TNFRSF10D TNF Receptor Superfamily Member 10d [Homo Sapiens (Human)] - Gene NCBI[EB/OL]. Nih.gov, 2025. [2025-03-23]. www.ncbi.nlm.nih.gov/gene/8793.

[5] Cao, Jixin, et al. Integrating Rare Pathogenic Variant Prioritization with Gene-Based Association Analysis to Identify Novel Genes and Relevant Multimodal Traits for Alzheimer's Disease [J]. Alzheimer's Dementia, 2024, 21(2). [2025-03-23].

[6] Huang, Xiaoqing, et al. Predicting Alzheimer's Disease Subtypes and Understanding Their Molecular Characteristics in Living Patients with Transcriptomic Trajectory Profiling[J]. Alzheimer's Dementia, 2025, 21(1). [2025-03-17]. pmc.ncbi.nlm.nih.gov/articles/PMC11772740/abstract1.

[7] Hu, Jichang, et al. Multi-Omics Analysis of Druggable Genes to Facilitate Alzheimer's Disease Therapy: A Multi-Cohort Machine Learning Study[J]. The Journal of Prevention of Alzheimer's Disease, 2025, (Nov.): 100128. pubmed.ncbi.nlm.nih.gov/40074652/.

[8] National Institute on Aging. Beyond Amyloid: Targeting a Broader Range of Alzheimer's Disease Factors [EB/OL]. Nia.nih.gov, 2025-02-26. [2025-03-23]. www.nia.nih.gov/research/blog/2025/02/beyond-amyloid-targeting-broader-range-alzheimers-disease-factors.

[9] Brigham, Mass General. Sixteen New Alzheimer's Disease Susceptibility Genes Discovered [EB/OL]. Medicalxpress.com, 2025-02-25. [2025-03-23]. medicalxpress.com/news/2025-02-sixteenalzheimer-disease-susceptibility-genes.html.

[10] OMIM. Entry - *603614 - Tumor Necrosis Factor Receptor Superfamily, Member 10D; TNFRSF10D[EB/OL]. Omim.org, 2025. [2025-03-23]. www.omim.org/entry/603614.