



OPEN Identification of UBE2N as a biomarker of Alzheimer's disease by combining WGCNA with machine learning algorithms

Gangyi Feng¹, Manli Zhong¹, Hudie Huang², Pu Zhao¹, Xiaoyu Zhang³, Tao Wang¹, Huiling Gao¹✉ & He Xu²✉

Alzheimer's disease (AD) is the most common cause of dementia, emphasizing the critical need for the development of biomarkers that facilitate accurate and objective assessment of disease progression for early detection and intervention to delay its onset. In our study, three AD datasets from the Gene Expression Omnibus (GEO) database were integrated for differential expression analysis, followed by a weighted gene co-expression network analysis (WGCNA), and potential AD biomarkers were screened. Our study identified UBE2N as a promising biomarker for AD. Functional enrichment analysis revealed that UBE2N is associated with synaptic vesicle cycling and T cell/B cell receptor signaling pathways. Notably, UBE2N expression levels were found to be significantly reduced in the cortex and hippocampus of the Tau^{P301S} mice. Furthermore, analysis of single-cell data from AD patients demonstrated the association of UBE2N and T cell function. These findings underscore the potential of UBE2N as a valuable biomarker for AD, offering important insights for diagnosis and targeted therapeutic strategies.

Alzheimer's Disease (AD) is the most common cause of dementia, accounting for 60–70% of cases¹. The gold standard for AD diagnosis is neuropathologic examination, which identifies hallmark features such as extracellular depositions of senile plaques generated by amyloid β (A β) and neurofibrillary tangles formed by hyperphosphorylated tau in different brain regions². Numerous AD risk factors have been identified, including age, genetic risk variants, stress, immune system dysfunction, and infectious diseases³. Based on the current knowledge of AD pathogenesis, both academic and pharmaceutical industries are actively engaged research and development to achieve a breakthrough in effective AD treatments⁴. A β has been investigated as the primary therapeutic target for many years. The US Food and Drug Administration has approved anti-amyloid antibodies such as aducanumab and lecanemab for AD treatment^{5,6}. Additionally, there is a growing focus on therapies targeting the tau protein⁷. AD biomarkers are confirmatory in clinical decision-making, particularly important for advancing disease-modifying therapies targeting A β and tau proteins. These biomarkers are also essential for improving both AD diagnosis and prognosis⁸.

Recent studies have shown that innate immune genes and cells directly or indirectly affect AD. Even without crossing the blood–brain barrier, T cells regulate brain homeostasis through a cascade of immune signals and secretory molecules. A recent study showed that CD8⁺ T cells are abnormally expanded in the brains of patients with mild cognitive impairment (MCI) and AD, indicating that CD8⁺ T cells may affect neurodegeneration and cognitive impairment in AD⁹. CD4⁺ T cells infiltrate the brain to promote A β clearance and neuronal repair¹⁰. Removal of B cells can significantly reduce A β and reverse memory deficits in a 3xTg AD mouse model¹¹. Blocking the transforming growth factor in the peripheral macrophage β (TGF- β)-mediated signaling pathway can reduce A β levels in the brain of the Tg2576 mouse model, which may be a potential treatment for AD¹². However, the roles of innate and adaptive immune cells have not been fully clarified, and there is an urgent need to identify new immune-related biomarkers to explain the neuroinflammation and pathogenesis of AD further.

The analysis and prediction of biomarkers for AD can significantly increase our understanding of its pathology and improve the development of new drug targets, clinical trials, and overall diagnosis¹³. Weighted gene co-expression network analysis (WGCNA) has been used to screen for potential biomarkers for various

¹Key Laboratory of Bioresource Research and Development of Liaoning Province, College of Life and Health Sciences, Northeastern University, Shenyang, China. ²Department of Anatomy, Histology and Embryology, School of Medicine, Shenzhen University, Shenzhen, China. ³Division of Biotechnology, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China. ✉email: gaohuiling@mail.neu.edu.cn; olivaxu@szu.edu.cn

diseases^{14,15}, as it analyzes the correlation between genes rather than being limited to a single gene¹⁶. Machine learning, a branch of computer science and statistics, is instrumental in the detection, diagnosis, and treatment of diseases¹⁷. The combination of WGCNA with machine learning considerably improves the accuracy of identifying potential disease biomarkers¹⁸. Using this combined approach, small nuclear ribonucleoprotein polypeptide G (SNRPG) has been identified as a critical gene in AD and metabolic syndrome¹⁹. Additionally, guidance receptor Plexin-B1 (PLXNB1), GRAM domain containing 3 (GRAMD3), and gap junction protein alpha (GJA) have been identified in different cortex and cerebellum regions associated with the Braak NFT stage in AD²⁰. Similar genomic transcription patterns in different cortex regions, according to the Braak 0–VI phases, may participate in the pathological progression of AD through the oxidation pathway²¹. Therefore, specific biomarkers have been screened for different research purposes using a combination of these two methodologies.

In this study, we aimed to discover new biomarkers for AD by combining WGCNA with machine learning algorithms. Initially, we identified 109 key differentially expressed genes (DEGs) using the GEO database and subsequently applied WGCNA. We then employed a protein–protein interaction (PPI) network and a joint machine learning of the GSE109887 dataset to identify and verify 13 potential biomarkers related to AD. By intersecting immune-related genes selected via the InnateDB database with these 13 potential AD-related genes, we identified UBE2N as an immune-related biomarker. Transcription factor prediction and Gene set enrichment analysis (GSEA) were used to explore the biological processes and pathways involving UBE2N in the context of AD. MFUZZ cluster analysis revealed that UBE2N is involved in T cell and B cell functions, as well as synaptic vesicle cycle signaling pathways. Single-cell data analysis of cerebrospinal fluid (CSF) and peripheral blood of AD patients demonstrated a close association between UBE2N and T cell function, with UBE2N levels being downregulated in CD4⁺ T cells from patients with mild MCI and AD. Additionally, we observed decreased UBE2N expression levels in Tau^{P301S} transgenic mice. All these results suggest that UBE2N is a novel biomarker for AD.

Results

Screening of DEGs in the brain of patients with AD

To screen for DEGs in the brains of patients with AD, we first removed the batch effect of genes between the AD and control groups and crossed the Principal Component Analysis (PCA) dataset that showed separation from each other, laying the foundation for subsequent analysis (Fig. 1A,B). The results included 124 samples from healthy individuals and 157 samples from patients with AD. By setting the screening criteria of $|\text{LogFC}| \geq 0.5$, $p < 0.05$, 86 upregulated and 153 downregulated genes were identified. The expression of DEGs is shown in heat and volcano maps (Fig. 1C,D). GO analysis revealed that DEGs were mainly enriched in vesicle-mediated synaptic transport, the synaptic vesicle cycle, and other related pathways (Fig. S1A). KEGG enrichment analysis showed that DEGs were mainly enriched in Alzheimer's disease, cAMP, and MAPK signaling pathways (Fig. S1B).

Screening characteristic genes of AD by construction of WGCNA

A total of 2,626 genes with expression variance in the top 25% were included in the WGCNA. Analysis of soft threshold selection showed that the average connectivity was high, and the scale-free network distribution reached its optimal level when $\beta = 7$ ($R^2 = 0.85$) (Fig. 2A). Subsequently, we obtained nine independent modules (with the lowest number of genes in the module set to 30 and genes not included in the module shown in gray) (Fig. 2B) by setting the clustering height to 0.25, merging highly correlated modules, and confirming the independence of each module (Fig. 2C). Next, the correlation between the module genes and AD was analyzed. The results showed that the highest correlation module was the turquoise module, which was positively correlated with the control group ($r = 0.43$, $p = 2e-14$) and negatively correlated with the AD group ($r = -0.43$, $p = 2e-14$) (Fig. 2D). The genes inside the turquoise module were screened according to the standard in which $GS \geq 0.35$ and $ME \geq 0.8$ were set as values closer to r , and 195 genes were obtained (Table S1). In total, 109 characteristic genes related to AD were identified by overlapping 195 genes with the DEGs (Fig. 2E). GO analysis showed that the characteristic genes were mainly enriched in exocytosis, glomerular development, and synaptic vesicle cycle (Fig. S2A). KEGG analysis showed that the characteristic genes were mainly enriched in GABAergic synapses, synaptic vesicle cycles, and cAMP signaling pathways (Fig. S2B).

Screening for potential biomarkers of AD by machine learning algorithms

Protein–protein interaction (PPI) networks were constructed based on the 109 potential targets. A total of 48 gene nodes in 80 lines were identified (Fig. 3A). First, 19 genes predicting the incidence of AD were obtained from 48 genes analyzed using the least absolute shrinkage and selection operator regression model (Fig. 3B). Then, 48 genes were analyzed using support vector machine recursive feature elimination, of which, 36 showed a high accuracy rate (0.787) and a low error rate (0.213) (Fig. 3C). The RF results showed that 33 AD-related genes were selected from 48 (Fig. 3D,E). Three machine learning algorithms obtained thirteen overlapping genes (Fig. 3F). ATP6V1E1, CCKBR, DYNC1I1, NRN1, SV2B, SYT1, TUBB2A, and UBE2N were expressed at low levels in the brains of patients with AD, whereas INPPL1, ITPKB, ITSN1, RAPGEF3, and TBL1X were highly expressed (Fig. S3). The validation set exhibited similar results (Fig. S4).

Identifying UBE2N as a biomarker for AD

The diagnostic value of the 13 biomarkers analyzed by ROC curves was greater than 0.7 for all AUC (Fig. 4C). A total of 1,696 immune genes were retrieved from the InnateDB database and overlapped with 13 AD biomarkers, identifying UBE2N as an overlapping gene (Fig. 4B). Biomarker correlation analysis revealed that UBE2N expression positively correlated with seven genes (TUBB2A, SV2B, NRN1, CCKBR, DYNC1I1, ATP6V1E1, and SYT1) and negatively correlated with five genes (INPPL1, ITSN1, ITPKB, RAPGEF3, and TBL1X) (Fig. 4A).

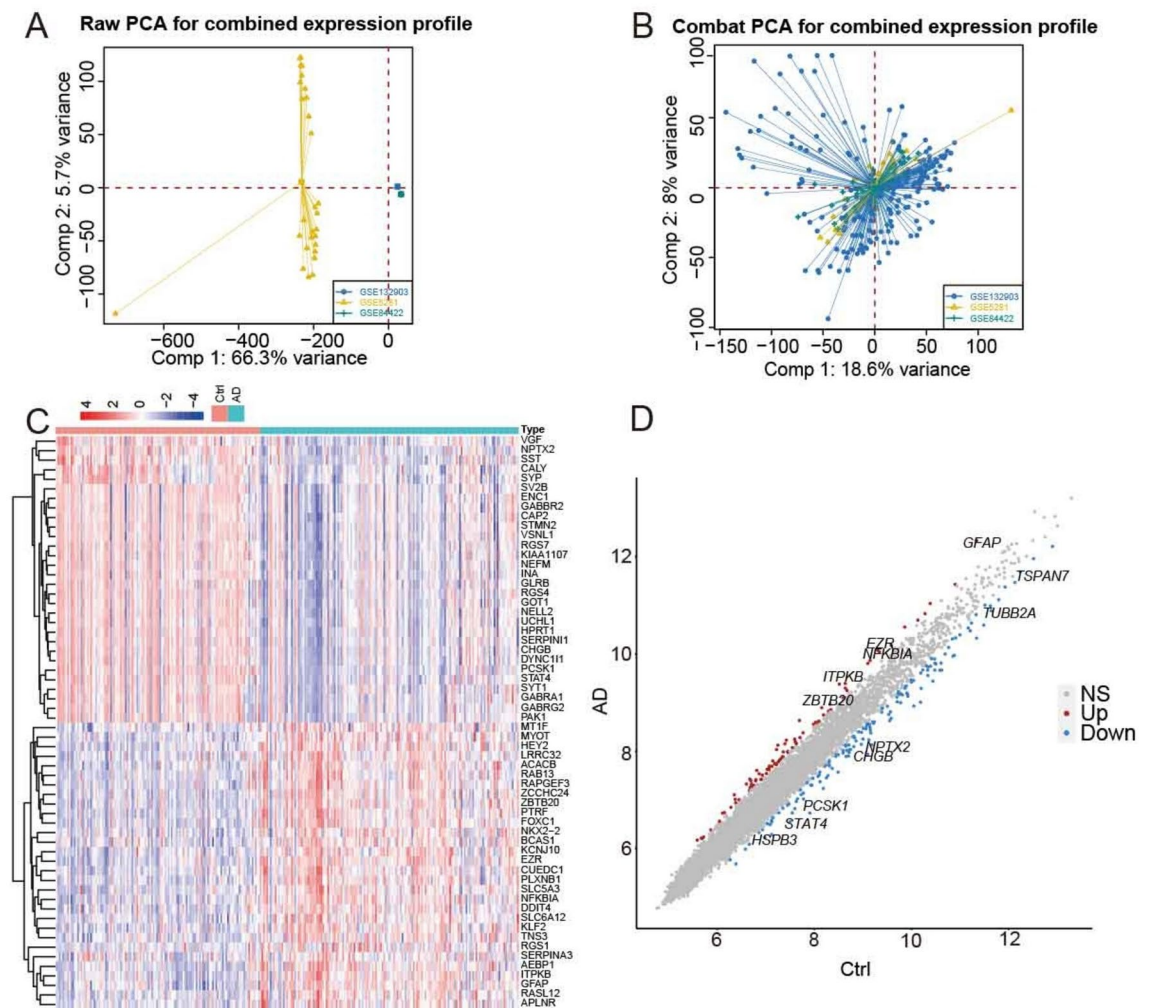


Fig. 1. Data preprocessing for DEG. (A) Raw PCA showing the analysis of the GSE13903, GSE5281, and GSE844223 datasets. (B) Combat PCA shows the three datasets after removal from the batch. (C) Heat map showing the top 30 upregulated genes and the top 30 downregulated genes after logFC sequencing. (D) The volcano map shows DEGs with $|\log_{2}FC| > 0.5$, and the significant DEGs are marked. PCA, principal component analysis; DEGs, differential genes.

The UBE2N column line graph was modeled by analyzing the calibration curves, and the differences between the normal and predicted values were small, indicating that the model was accurate (Fig. 4D,E). In the decision curve analysis (DCA), the model curve was above the grey line, implying that patients could benefit from the model within the threshold (Fig. 4F). The clinical impact curve also demonstrated a better overall net benefit in the threshold range (Fig. 4G). This suggests that the UBE2N columnar line graph model constructed in this study can be used to assess AD prognosis.

Analysis of potential regulatory mechanisms of UBE2N

Based on the pooled median expression values, data from patients with AD were divided into two groups: high- and low UBE2N expression. KEGG analysis was performed on DEGs. The results showed that UBE2N might participate in the activation of five pathways, including nicotine addiction, calcium reabsorption regulated by endocrine and other factors, synaptic vesicle circulation, oxidative phosphorylation, alanine, aspartate, and glutamate metabolism. In addition, UBE2N may be involved in the inhibition of five signaling pathways, including the interaction of the viral protein with cytokines and cytokine receptors, graft rejection, and malaria and *Staphylococcus aureus* infections (Fig. 5A). GSEA analysis showed that the remaining six genes that were positively correlated with UBE2N were mainly enriched in the synaptic vesicle circulation pathway (Fig. S5). These results indicate that UBE2N may be related to vesicular function. In addition, the potential regulatory network analysis of UBE2N showed that hsa-miR-128-3p, hsa-miR-149-5p, hsa-miR-221-3p, hsa-miR-222-3p, hsa-miR-5010-5p, hsa-miR-522-3p, and hsa-miR-96-5p regulated UBE2N expression and the transcription factor ATF1 was also involved in the transcription regulation process of UBE2N (Fig. 5B).

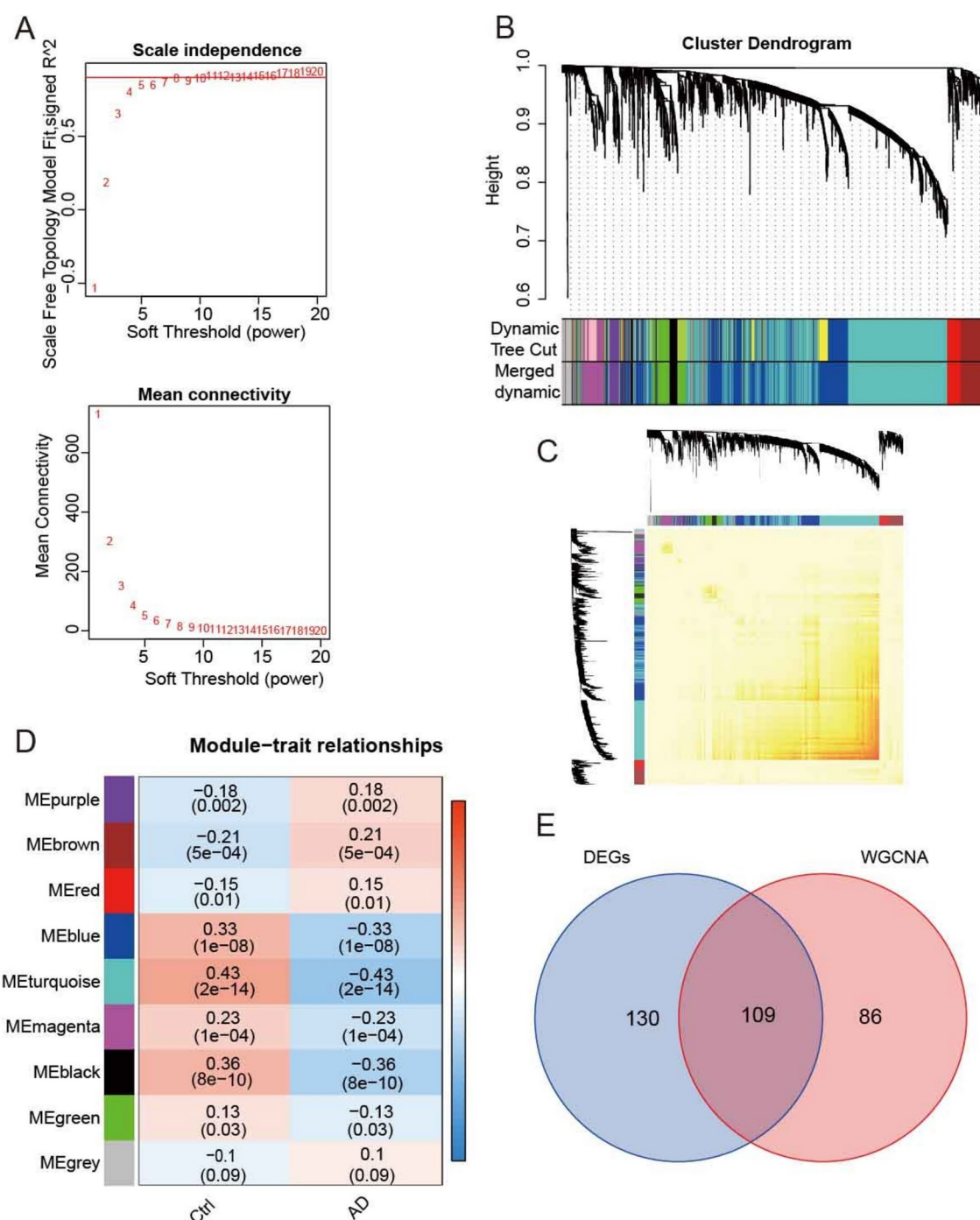


Fig. 2. Construction of the WGCNA co-expression network. **(A)** Soft threshold ($\beta = 7$) and scale-free topological fit index (R^2). **(B)** Original and combined modules using a clustering tree. The cluster dendrogram was the result before module cutting, while the Merged Dynamic was the result after module cutting. **(C)** Cluster tree plot of the module feature genes. **(D)** Heat map of module-trait correlations. The values inside the brackets represent the p-value of significance, whereas the values outside the brackets represent the values of the correlation between the Control and AD groups. **(E)** Venn diagram of key module genes vs. DEGs. The overlapping regions represented the same 109 genes in DEGs and WGCNA. AD, Alzheimer's Disease; DEGs, differential genes; WGCNA: Weighted gene co-expression network analysis.

Cluster analysis of MFUZZ expression patterns

Fifty different clustering results were obtained, based on the expression patterns of UBE2N. The obtained clustering results for ssGSEA scoring were analyzed with the control and AD groups for correlation analysis, and cluster 27 showed the closest module related to UBE2N (Fig. 6A-C). Functional enrichment analysis showed that the cluster 27 module genes were mainly enriched in the T cell receptor, B cell receptor, and synaptic

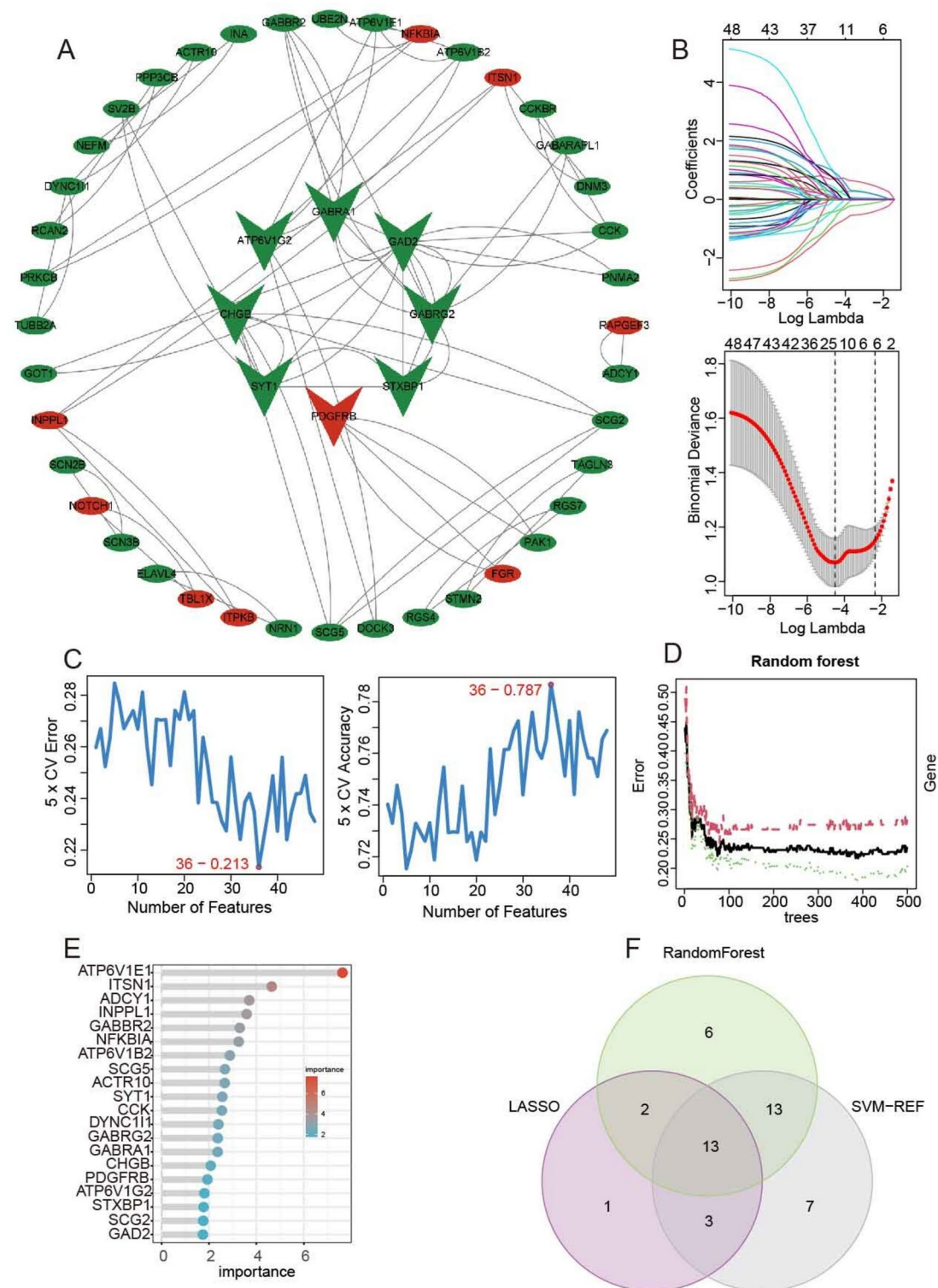


Fig. 3. Screening for potential biomarkers of AD using a machine learning algorithm. (A) PPI network diagram; red and green represent upregulated and downregulated genes, respectively. (B) The minimum absolute contraction and selection operator models (LASSO) were used to select the characteristic genes. (C) The SVM-RFE algorithm selected biomarker feature genes. The red circle at the lowest point in the left image indicates a minimum error rate of 0.213 for 36 genes, whereas the red circle at the highest point in the right image represents a maximum accuracy rate of 0.787 for 36 genes. (D) The random forest tree algorithm was used to evaluate characteristic genes. (E) The top 20 most important genes were identified using the random forest tree algorithm. (F) Venn diagram of the three algorithms used to screen genes. The overlapping parts of the three circles represent 13 common genes obtained from the three machine learning algorithms. AD, Alzheimer's Disease; SVM-RFE, support vector machine recursive feature elimination.

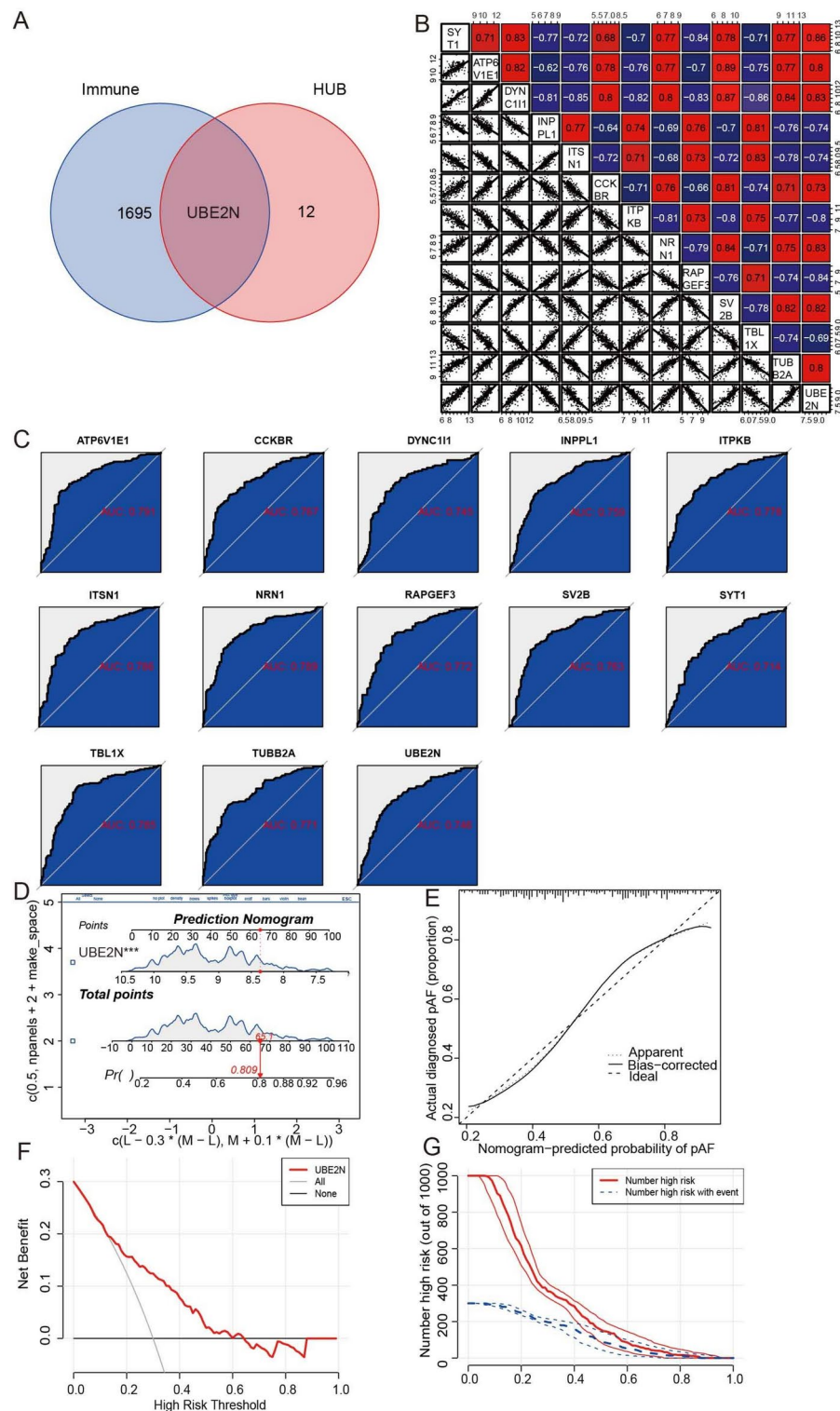


Fig. 4. Evaluation of UBE2N as a biomarker of AD. **(A)** Venn plots of the immune and diagnostic markers. **(B)** Correlation between genes. Red squares show the positive correlation of genes and blue squares represent the negative correlation of genes. **(C)** Training focused on the ROC curves for the diagnostic markers. **(D)** Norman diagrams were used to predict AD incidence. **(E)** The ROC curve evaluates the clinical application value of the Norman diagram model. **(F)** DCA curve evaluates the clinical application value of the Norman diagram model. **(G)** Clinical impact curve: the red curve (number of high-risk individuals) represents the number of individuals classified as positive (high-risk) by the model at each threshold probability; the blue curve (the number of at-risk individuals with results) is the number of true positives at each threshold probability. ROC, Receiver Operating Characteristic; DCA, decision curve analysis.

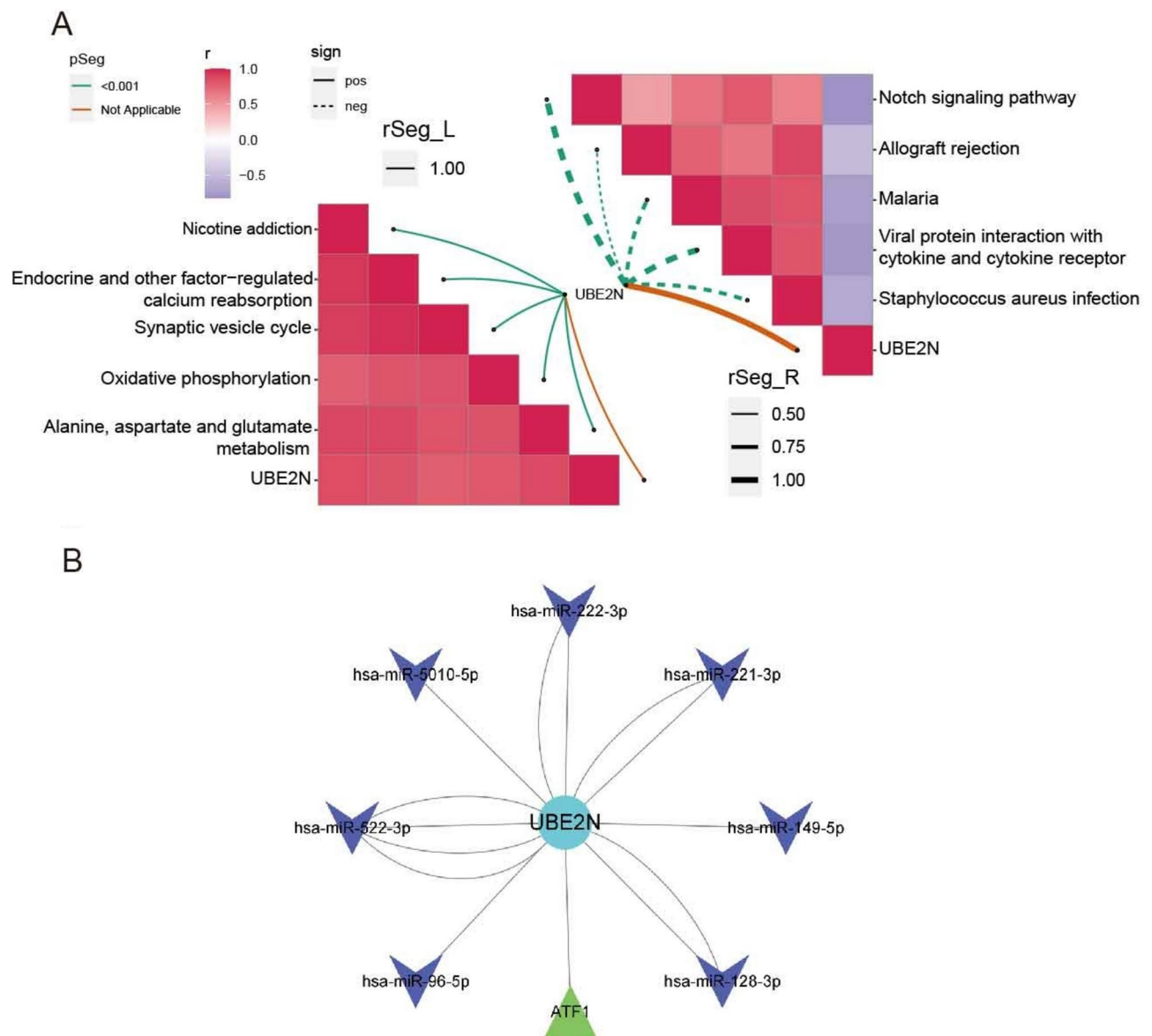


Fig. 5. Analysis of potential regulatory mechanisms of UBE2N. **(A)** Butterfly diagram showing the first five upregulated and downregulated pathways in the GSEA. **(B)** Regulatory network of UBE2N cells. The blue graphs represent microRNAs associated with UBE2N; ATF1 as the transcription factor (green triangle) is associated with UBE2N. GSEA: Gene Set Enrichment Analysis.

vesicle signaling pathways (Fig. 6D). Subsequently, the genes in the cluster 27 module overlapped with 13 AD biomarkers, and seven genes (ATP6V1E1, CCKBR, DYNC1I1, NRN1, SV2B, SYT1, and TUBB2A) were highly correlated with UBE2N expression (Fig. 6E).

Validating and interpreting machine learning models

To determine the best machine learning model for predicting AD, we used 281 samples for the training set (124 normal and 157 AD samples) and 78 samples for the validation set (32 normal and 46 AD samples). Seven machine learning models including eXtreme Gradient Boosting (XGBoost), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (Logistic), Gaussian Naive Bayes (GNB), and Multilayer Perceptron (MLP) were built to predict the results using model genes (UBE2N, ATP6V1E1, CCKBR, DYNC1I1, NRN1, SV2B, SYT1, and TUBB2A). By calculating the performance (accuracy, sensitivity, specificity, Positive predictive value, Negative predictive value, F1 score, Kappa score) of multiple machine learning models in the training cohort versus the validation cohort, we found that XGBoost model demonstrated strong performance in the test set, achieving an accuracy of 0.69, a sensitivity of 0.96, a negative predictive value of 0.96 and an F1 score of 0.79 (Fig. S7A,B). The XGBoost model had an AUC value of 0.77 and a PR value of 0.81 in the test cohort (Fig. 7A,B). Next, we investigated how the model predicted AD onset

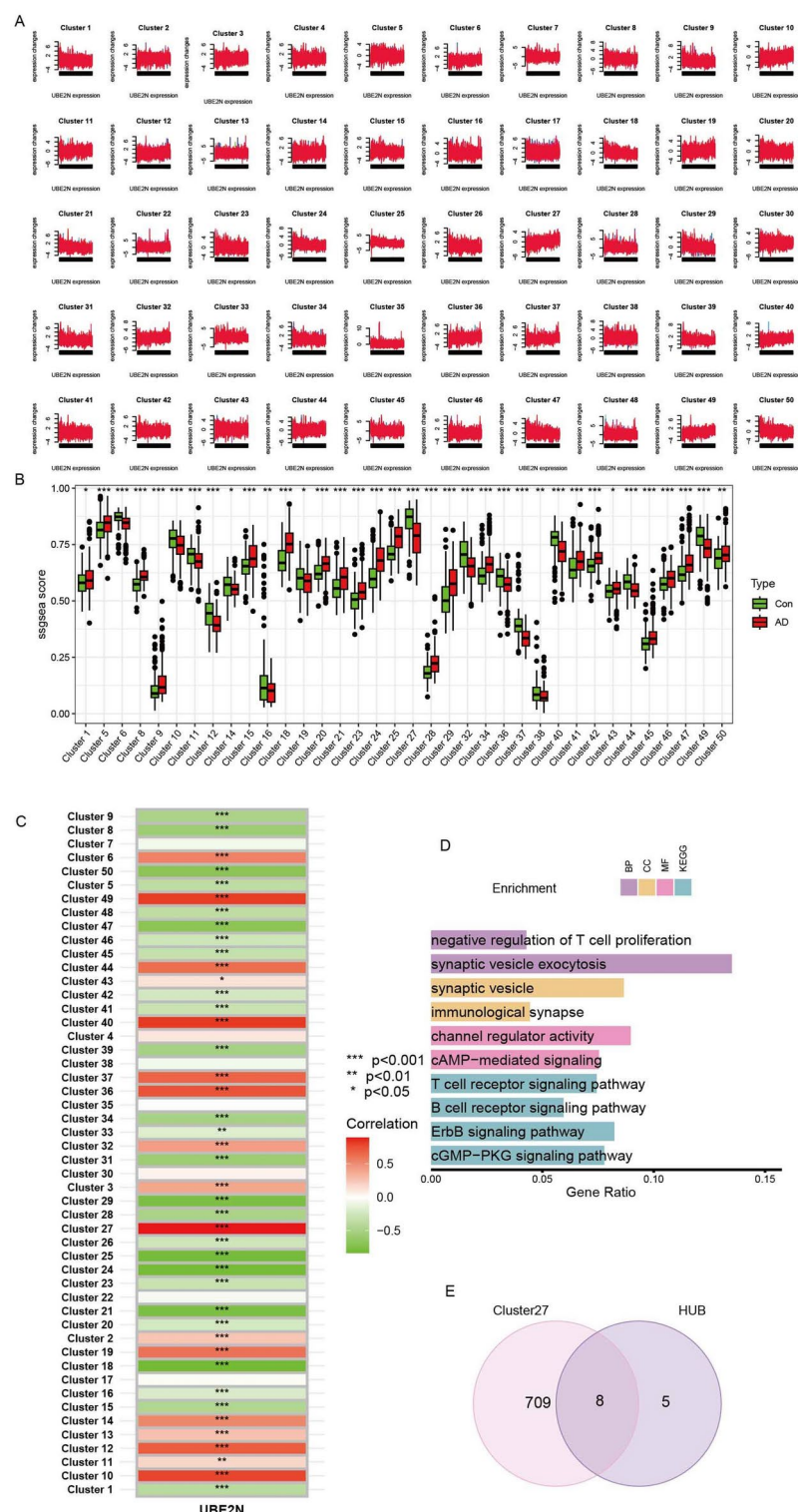


Fig. 6. Cluster Analysis of MFUZZ Expression Patterns. **(A)** Clustering results for MFUZZ expression patterns. **(B)** ssGSEA scores and expression characteristics of the clustering modules between the control groups (green column) and AD (red column). **(C)** Correlation between clustering modules and UBE2N. Red indicates a positive correlation with the cluster and green indicates a negative correlation with the cluster. The darker the color, the stronger is the correlation. **(D)** KEGG enrichment analysis of genes in cluster 27^{64,65}: the darker the color, the more genes enriched in the pathway, and the longer the column, the smaller the *p*-value. **(E)** Core genes co-expressed with UBE2N in AD cells. Eight core genes were obtained by crossing the Cluster27 and AD datasets. GSEA: Gene Set Enrichment Analysis.

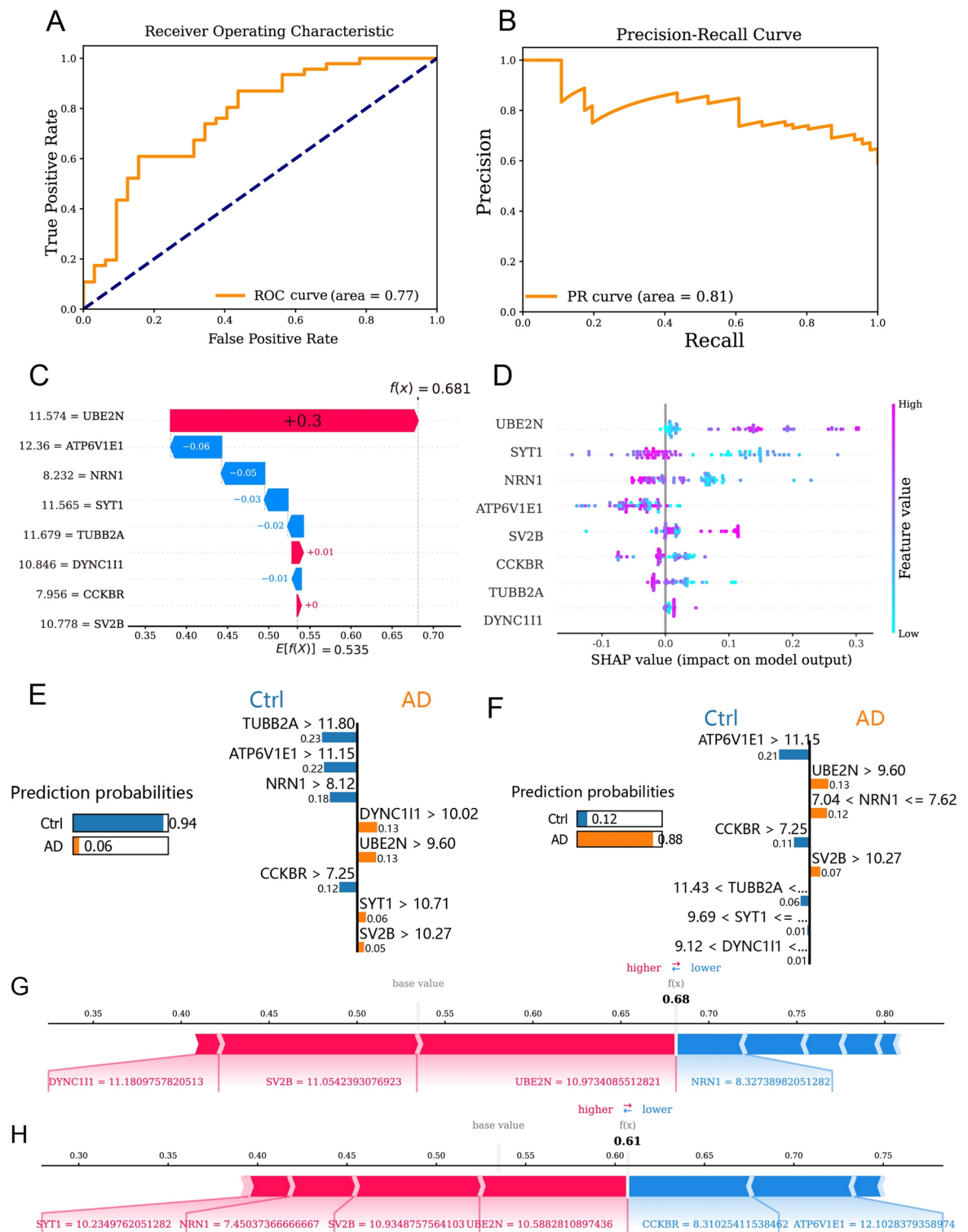


Fig. 7. Validating and interpreting machine learning models. **(A)** Specific values of AUC in the XGBoost model. **(B)** Specific values of P-R in the XGBoost model. **(C)** The importance matrix shows the contribution of each feature gene in the XGBoost model. **(D)** The SHAP summary plot shows the contribution of each feature to the XGBoost model. **(D, E)** The left side of the figure shows the predicted results of the LIME. The ten variables with the greatest impact on normal or AD onset are listed on the right side. The length of each feature bar indicates the importance of the corresponding feature in the prediction. **(G, H)** Predicted probabilities of normal and AD onset based on SHAP force maps, respectively. Red and blue bars represent increased and decreased likelihood of AD, respectively.

through Shapley Additive exPlanation (SHAP) values and elucidated the impact of each feature variable on the predictive model. Importance ranking of the feature variables based on the SHAP summary plot of the XGBoost model showed that UBE2N contributed most to the XGBoost model (Fig. 7C). In addition, we employed SHAP dependency analysis to characterize how a single feature variable affects the results of the XGBoost prediction model (Fig. 7D). The SHAP values of the feature variables were positively correlated with the accuracy of the model. The accuracy of the model can be improved by a positive value of UBE2N but reduced by a negative value of ATP6V1E1. The normal and AD samples in the testing cohort were further depicted using SHAP force diagram and Interpretable Model-agnostic Explanations (LIME) algorithm. The prediction outcomes for the normal sample (randomly selected from the test set) by XGBoost model showed a 6% probability of classifying this particular sample as normal and a 94% probability of classifying it as AD (Fig. 7E,G). At the same time, the XGBoost model predicted an AD sample (randomly selected from the test set) with an 88% probability of being AD and a 12% probability as normal (Fig. 7F,H). These results are consistent with the actual status of the samples. After excluding the feature gene UBE2N from the module, the performance of the XGBoost model obviously declined, with the ROC value decreasing to 0.72, which is a reduction of 0.05 from its initial performance. Additionally, the P-R value also dropped by 0.02 (Fig. S8).

Consensus clustering identifies two AD subclasses related to UBE2N

Utilizing the “ConsensusClusterPlus” software package, we conducted a consensus cluster analysis involving UBE2N and its related genes: tubulin beta 2A class IIa (TUBB2A), synaptic vesicle glycoprotein 2B (SV2B), neuritin 1 (NRN1), cholecystokinin B receptor (CCKBR), dynein cytoplasmic 1 intermediate chain 1 (DYNC1I1), ATPase H + transporting V1 subunit E1 (ATP6V1E1), and synaptotagmin 1 (SYT1). Following a comprehensive analysis, $k=2$ was determined to be the optimal cluster number. Consequently, AD samples were stratified into two distinct subclasses, designated as C1 and C2, as illustrated in the cumulative distribution function (CDF) plot (Fig. 8A). Principal component analysis (PCA), uniform manifold approximation and projection (UMAP), t-distributed stochastic neighbor embedding (TSNE) demonstrated significant differences between these two subclasses (Fig. 8B). A variance heatmap analysis indicated that the expression levels of UBE2N and the seven related genes were significantly lower in the C1 group compared to the C2 group (Fig. 8C). Furthermore, we employed the Gene Set Variation Analysis (GSVA) method to examine the functional implications of differential gene regulation between the subclasses. The results showed that the differential bases were enriched in several signaling pathways, including apoptosis, the mammalian target of rapamycin (mTOR) signaling pathway, and the T cell/B cell receptor signaling pathway (Fig. 8D).

Immune cell infiltration analysis

To evaluate the infiltration status of immune cells in the control and AD groups, we compared the expression levels of 28 types of immune cells. Compared with that in the control group, the infiltration rate of T cells in the AD group was higher, including natural killer, gamma delta, central memory CD8⁺, factor memory CD8⁺, and central memory CD4⁺ T cells. However, the infiltration rate of factor memory CD4⁺ T cells was lower in the AD group than in the control group (Fig. 9A). Subsequently, correlations between the eight genes and immune cells were analyzed. UBE2N and DYNC1I1 positively correlated with activated CD4⁺ T cells, SYT1 negatively correlated with activated CD8⁺ T cells, and CCKBR negatively correlated with immature dendritic cells (Fig. 9B). UBE2N had the highest correlation with effector memory CD4⁺ T cells ($r=0.7$) (Fig. 9C, D).

Comprehensive analysis of single-cell transcriptome data from cerebrospinal fluid (CSF) of patients with AD and healthy controls

To further investigate the role of UBE2N in the immune cells of AD, we conducted a comprehensive analysis of 59 CSF samples, which included 45 healthy controls, 8 patients with MCI, and 6 AD patients. Following the filtering of single-cell sequencing data, we obtained 70,391 high-quality cells. Using the highly variable gene-based uniform manifold approximation and projection (UMAP) algorithm of the Seurat package, these cells were distributed into 18 distinct clusters labeled by 23 immune markers. Excluding the indeterminate clusters (13th and 17th clusters), we identified the following 10 cell types for subsequent analysis: B Cells, CD4⁺ T Cells, CD4⁺/CD8⁺ T Cells, CD8⁺ T Cells, CD14⁺/CD68^{hi} Monocytes, CD14⁺/CD68^{lo} Monocytes, dendritic cells (DCs), Natural Killer (NK) Cells, Plasma Cells, and Regulatory T (Treg) Cells (Fig. 10A–C). Quantification of immune cell proportions in each subpopulation revealed that the numbers of CD4⁺ T Cells, CD4⁺/CD8⁺ T Cells, and CD8⁺ T Cells were lower in MCI and AD patients compared with healthy individuals. Conversely, the number of Treg cells was higher in MCI and AD (Fig. 10D). To validate the accuracy of the classified cell types, we analyzed the top five differentially expressed genes within each subpopulation (Fig. 10E). UBE2N was found to be highly expressed in CD4⁺ T cells and DCs. Moreover, the expression of UBE2N in CD4⁺ T cells of MCI and AD patients was significantly reduced compared with healthy controls (Fig. 10F,G).

Expression levels of UBE2N were decreased in the brain of an AD mouse model

Quantitative real-time PCR (qRT-PCR) results revealed that the gene expression levels of UBE2N, ATP6V1E1, CCKBR, SV2B, and TUBB2A were significantly reduced in the cerebral cortex of Tau^{P301S} mice, whereas the expression levels of DYNC1I1, NRN1, and SYT1 remained unchanged compared to the control group (Fig. 11A). Additionally, the protein levels of UBE2N decreased significantly in both the cerebral cortex and hippocampus of Tau^{P301S} mice (Fig. 11B,C). However, no significant change in UBE2N expression was observed in the cerebral cortex of APP/PS1 mice (Fig. S6). Immunofluorescence studies demonstrated that UBE2N co-localized with NeuN in the hippocampus and temporal cortex of Tau^{P301S} mice. Compared to C57BL/6 mice, Tau^{P301S} mice exhibited a marked reduction in UBE2N fluorescence intensity in both the hippocampus and cortex (Fig. 11D,E).

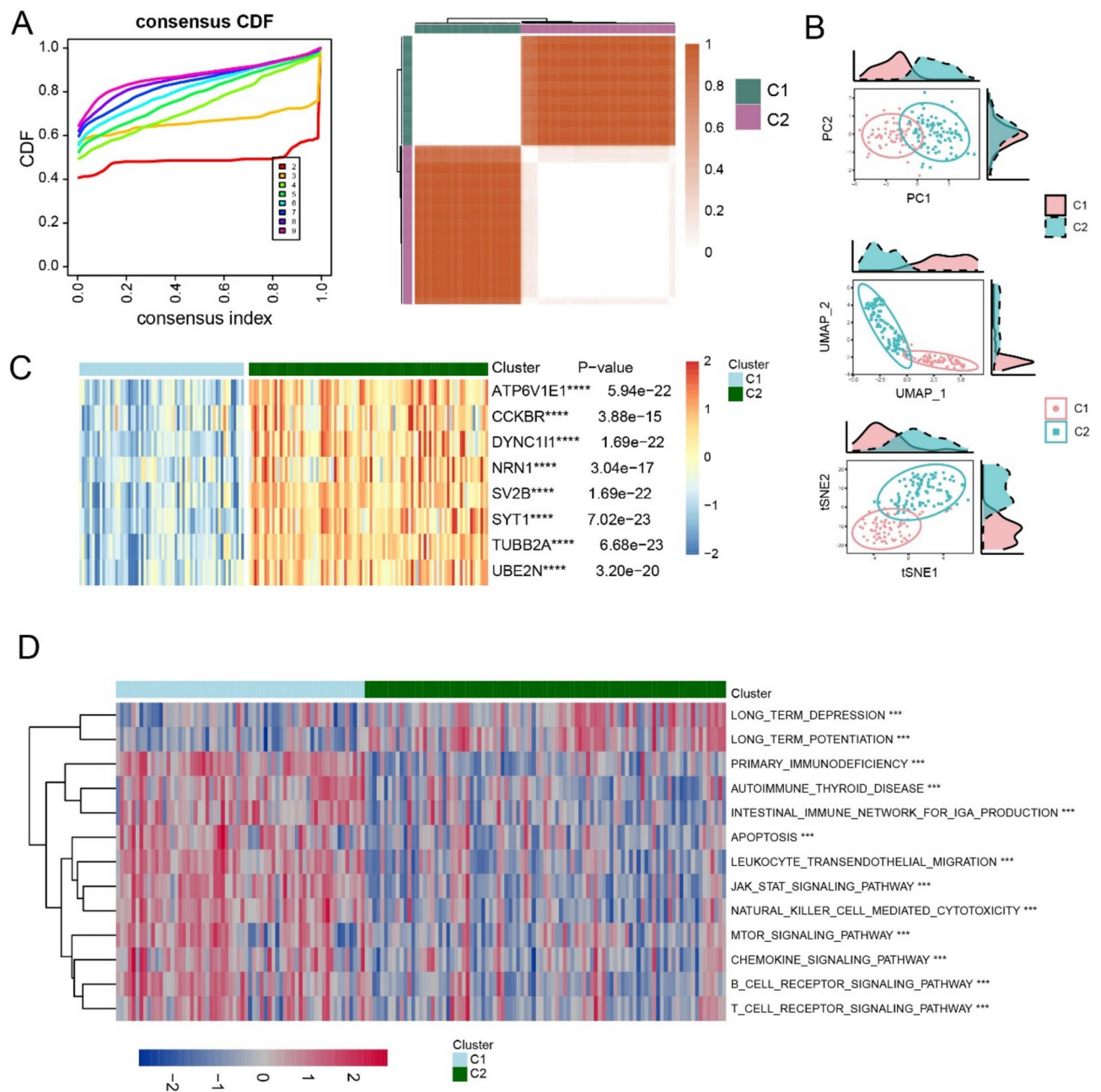


Fig. 8. Consensus clustering identifying two AD subclasses. **(A)** Consensus Clustering CDF for $k=2-9$, and consensus clustering for $k=2$. **(B)** tSNE, PCA, and UMAP analyses between the two subgroups. **(C)** Analysis of differences in gene expression between the two subgroups. **(D)** Pathway activity between two different subgroups based on the ssGSEA algorithm. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns, no statistical significance.

Discussion

AD is a neurodegenerative disease that impairs cognitive function, mainly affecting brain regions related to learning and memory, such as the temporal lobe and hippocampus²². In the current study, we performed a comprehensive and in-depth analysis of gene expression profile in temporal lobe to identify AD-specific genes, resulting in the identification of 13 HUB genes. Among them, UBE2N was validated in the cerebral cortex and hippocampus of Tau^{P301S} mice and emerged as the most robust biomarker for AD in our study.

By analyzing microarray data from the temporal lobes of AD patients in the GEO database, we identified 239 DEGs, consisting of 86 upregulated and 153 downregulated genes. Overlapping the key modular genes obtained by WGCNA with DEGs yielded 109 key DEGs. These DEGs were mainly enriched in GABAergic synaptic, B-cell receptor, and synaptic vesicle cycle signaling pathways, all of which are critical pathological changes in the pathogenesis of AD²³. Subsequently, a machine learning algorithm was used to identify 13 HUB genes. Cross-referencing immune genes in the InnateDB database revealed the UBE2N gene. Furthermore, gene correlation

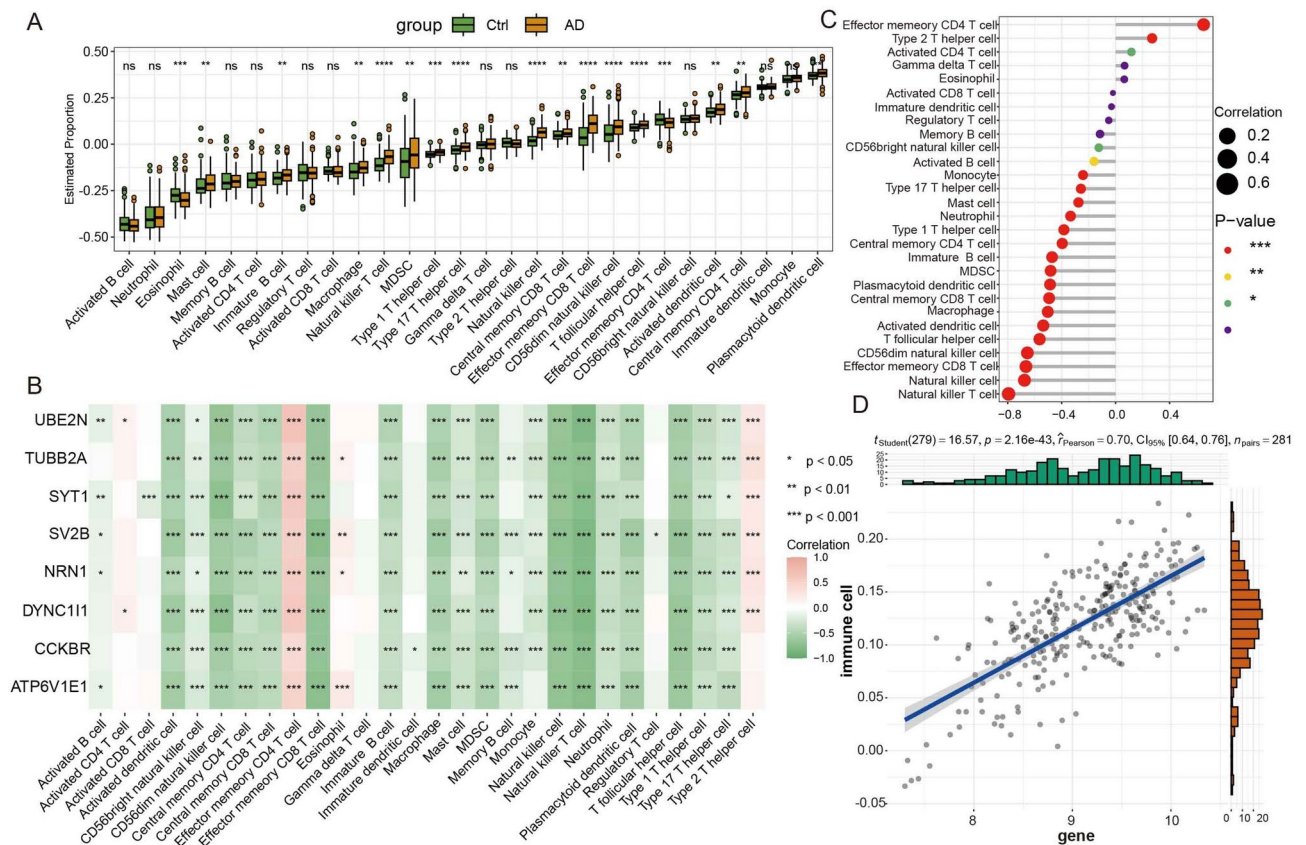


Fig. 9. Immune cell infiltration analysis. **(A)** Comparison of 28 immune cell types; green and orange represent the control and AD groups, respectively. **(B)** Correlation between characteristic genes and immunity. **(C)** Correlation between UBE2N and the 28 types of immune cells. **(D)** Correlation between UBE2N and effector memory CD4⁺ T cells.

analysis demonstrated that ATP6V1E1, CCKBR, SV2B, DYNC111, NRN1, SYT1, and TUBB2A were positively correlated with UBE2N expression. The AUC areas under the ROC curves are all greater than 0.7, indicating that the constructed model could accurately predict the onset of AD²⁴.

UBE2N plays an important role in several neurodegenerative diseases. Overexpression of UBE2N increases the aggregation of mutant Huntington's proteins²⁵. In contrast, knockdown of the E2 enzymes UBE2N, UBE2L3, UBE2D2 and UBE2D3 (UBE2D2/3) significantly reduces the autophagic clearance of depolarized mitochondria; furthermore, UBE2N, UBE2L3, and UBE2D2/3 synergistically promote Parkin-mediated mitochondrial autophagy²⁶. Single-cell analysis revealed reduced expression of UBE2N in cerebrospinal fluid CD4⁺ T cells from patients with AD. Notably, the abnormal downregulation of UBE2N causes in vivo immunosuppressive dysfunction of regulatory T cells, leading to abnormal activation of T cells and induction of various inflammatory responses²⁷. Therefore, we analyzed the infiltration of 28 immune cells and found that the infiltration rate of T cells was significantly higher in AD brains, suggesting that the balance of T cells may be dysregulated in AD. One reason could be the decrease in tight junction molecules in the vascular endothelium during AD progression, leading to an increase in the permeability of the blood–brain barrier. Owing to increased chemokines for T cells in the brain of patients with AD, these changes together promote T cell infiltration^{28,29}. The accumulation of activated T cells has been demonstrated to induce neuronal death and exacerbate neuroinflammation³⁰. Activated T cells can also promote the release of TNF- α , IL-1, and IL-6 pro-inflammatory factors from peripheral blood mononuclear cells (PBMC), exacerbating the inflammatory response³¹. In addition, a large number of CD8⁺ T cells are found in the hippocampus of AD patients, and tau-specific CD4⁺ T cells are widely distributed in the peripheral blood of AD patients, suggesting that T cells may be closely associated with AD progression, particularly in terms of tau pathology^{32,33}. Therefore, we speculate that UBE2N may affect AD pathology by regulating T cells, which needs to be confirmed by further experiments.

Furthermore, MFUZZ cluster analysis showed that the cluster comprising the 27th modular gene had the highest correlation with UBE2N, and the functions of the modular genes were mainly related to immunity and synapses, as expected. Notably, Cluster 27 overlapped with 8 of the 13 identified hub genes (UBE2N, ATP6V1E1, CCKBR, DYNC111, NRN1, SV2B, SYT1, and TUBB2A), 7 of which were positively associated with UBE2N. Our qRT-PCR and Western blot results showed that UBE2N expression was significantly reduced in the cerebral cortex of Tau^{P301S} mice, but not in APP/PS1 mice, indicating that UBE2N may play a specific role in certain models of AD, particularly tau-related pathways in AD pathogenesis.

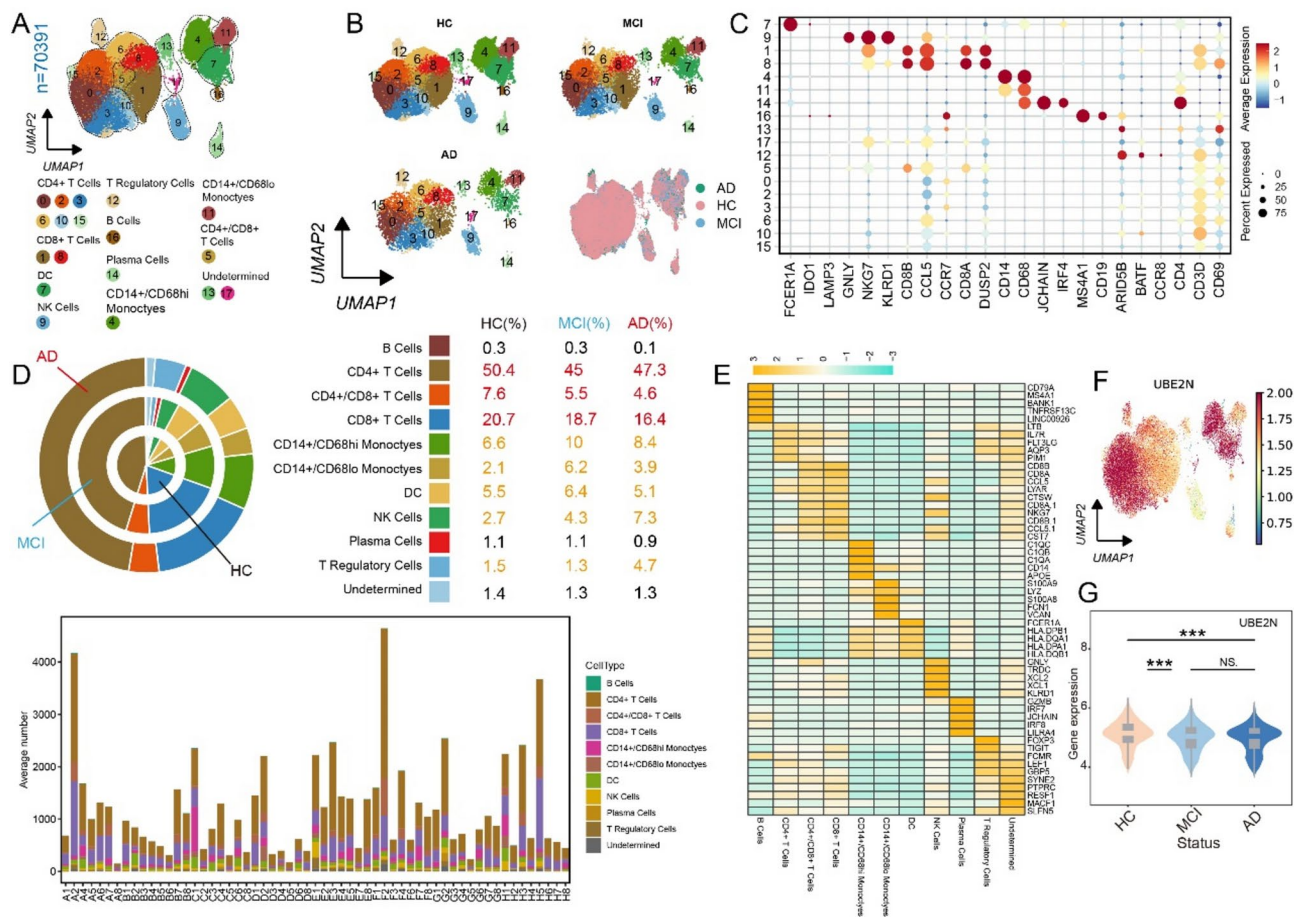


Fig. 10. Comprehensive analysis of single-cell transcriptome data from cerebrospinal fluid of AD patients and healthy samples. (A) UMAP projections of 70,391 CFS cells from patients with AD. (B) UMAP projections for different subgroups and after the removal of batch effects. (C) Annotation of individual cell clusters by using marker genes for different cell types. (D) Proportions of different cell types in each subgroup. (E) Heatmap of the first five DEGs in each cell population. (F) UMAP projections of the UBE2N expression distribution in each cell population. (G) Violin plots of differences in UBE2N expression in each subgroup. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns, no statistical significance.

In addition to UBE2N, we found that the mRNA levels of ATP6V1E1, CCKBR, SV2B, and TUBB2A were decreased in the Tau^{P301S} cortex. ATP6V1E1 is a large multi-subunit complex divided into a peripheral structural domain (V1) and a proton transmembrane translocation structural domain (V0) that is upregulated in early AD and downregulated in late-stage AD³⁴. ATP6V1E1 serves as a proton pump and mediates the acidification of endosomes, lysosomes, the Golgi and synaptic vesicles³⁵. Its dysfunction, therefore, disrupts PH homeostasis, affecting organelle acidification and, consequently, contributing to AD. ATP6V1E1 is reduced not only in the brain but also in the peripheral blood of patients with AD, suggesting that ATP6V1E1 may play an important role in the diagnosis and treatment of AD³⁶. Cholecystokinin (CCK) mediates its action through two G-protein-coupled receptors, CCKAR and CCKBR. Its absence leads to abnormalities in the cerebral cortex and corpus callosum development and further affects the migration of cortical interneurons³⁷. Synthetic CCK analogs can effectively reduce A β load in the brain and normalize the levels of protein kinase A (PKA), cAMP response element binding protein (CREB), brain-derived neurotrophic factor (BDNF) and TrkB receptors, thereby improving APP/PS1 mice cognition³⁸. Synaptic vesicle glycoprotein 2B (SV2B) is a synaptic protein involved in APP/A β metabolism³⁹. There is evidence that SV2B knockout protects against A β -induced memory deficits and ameliorates cholinergic system dysfunction caused by A β injection. However, another study found that A β levels were significantly elevated in the hippocampus of SV2B knockout mice than in those of WT mice⁴⁰. Here, we observed downregulation of SV2B in Tau^{P301S} mice, suggesting that SV2B may also be related to tau pathology. TUBB2A is a microtubule protein. Tau from the AD brains increased endogenous Tau in cortical neurons; furthermore, transcriptome sequencing results showed that TUBB2A is remarkably present in neurons⁴¹. Here, we first demonstrated that TUBB2A is reduced in the cerebral cortex of Tau^{P301S} mice, indicating that TUBB2A might play a role in the tau-related pathway in AD.

Our study integrates four AD datasets from the GEO database, which enhances the robustness and generalizability of the findings, but relying on these existing datasets might have inherent biases or limitations in terms of sample diversity and size. Our findings need to be validated in larger and more diverse cohorts.

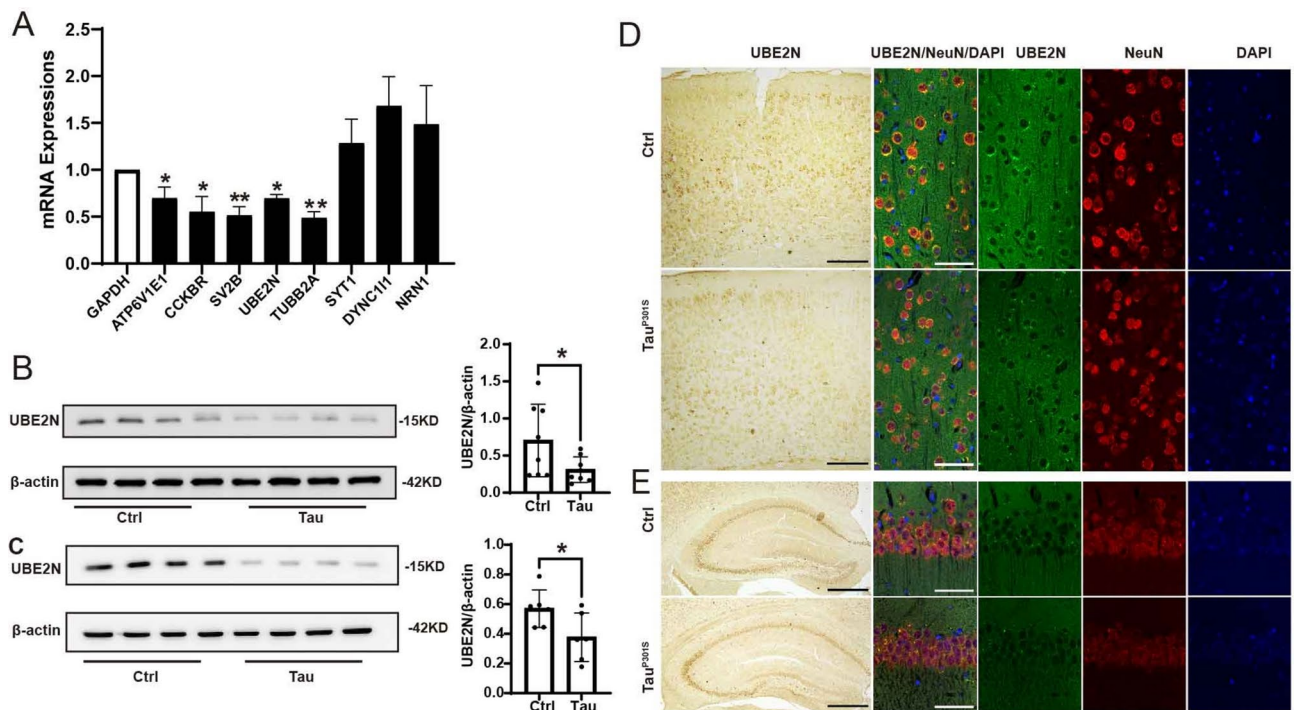


Fig. 11. Expression of UBE2N in the brains of $\text{Tau}^{\text{P301S}}$ mice. **(A)** Differential expression of *UBE2N* and related genes in the cortex of control ($n=8$) and $\text{Tau}^{\text{P301S}}$ mouse brains ($n=8$) was analyzed by RT-qPCR using GAPDH as a control. **(B)** UBE2N expression in the temporal cortex of the control ($n=8$) and $\text{Tau}^{\text{P301S}}$ mice ($n=8$). **(C)** UBE2N expression in the hippocampi of control ($n=8$) and $\text{Tau}^{\text{P301S}}$ mice ($n=8$). Positive staining for UBE2N was observed in the cortex **(D)** and hippocampus **(E)**, and it colocalized with red-labeled NeuN in the cortex and hippocampus in immunofluorescence images. The scale bar for immunohistochemistry images is 50 μm , and the scale bar for cortical and hippocampal immunofluorescence images is 200 μm and 500 μm , respectively. All data from at least three independent experiments are presented as mean $\pm$ SEM. * $p < 0.05$, ** $p < 0.01$ with respect to the control group.

In future study, incorporating various heterogeneous biological data could enhance prediction accuracy. Due to advancements in high-throughput techniques, numerous heterogeneous biological datasets have emerged. These datasets have the potential to provide valuable insights into the underlying mechanisms of AD onset. Moreover, current analysis is based on cross-sectional data, which limits the ability to infer causal relationships. Longitudinal studies are needed to establish the temporal dynamics of UBE2N expression in AD progression. Machine learning models have powerful predictive capabilities; however, they are prone to overfitting when dealing with small datasets and sensitive to noise and outliers, which may potentially lead to degraded performance.

Among the various machine-learning techniques, Matrix factorization (MF) in recommender systems has received significant attention for predicting potential disease biomarkers. Previous studies predicted miRNAs as biomarkers for diseases using related computational models, such as IMIPMF and SMAP^{42–46}. In this regard, further efforts to integrate these methods to discover novel AD biomarkers could improve the prediction accuracy. While our study links UBE2N to certain pathways, the precise functional mechanisms through which UBE2N influences AD pathology remain to be fully elucidated, such as the role of UBE2N in CD4^+ T cell function in AD.

In conclusion, we identified an immune-related gene, UBE2N, as a biomarker for AD by combining WGCNA with machine learning approaches. UBE2N may play a role in synaptic vesicle function, neuroinflammation, and T-cell infiltration, thereby influencing AD pathologies, particularly tau-related. Our findings could not only improve the understanding of disease pathogenesis but also contribute to the treatment and detection of AD.

Materials and methods

Data acquisition and pre-processing

Four microarray datasets (GSE5281, GSE84422 and GSE132903, and GSE109887) related to AD were obtained from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>). The “SVA” package in R (4.2.1) was used to remove batch effects among data sets. Information about these datasets is provided in Table 1⁴⁷. The gene expression differences were analyzed by the “Limma” software package based on the screening criteria “Adjusted $p < 0.05$ and $|\log\text{FC}| \geq 0.5$ ”⁴⁸. The volcanic and thermal maps were created by the “ggplot2” software package and the “pheatmap” software package, respectively.

GEO datasets	Platform	Sample normal	Sample AD	Publication years	Regions
GSE5281	GPL570	12	16	2006	USA
GSE84422	GPL96	14	44	2016	USA
GSE132903	GPL10558	98	97	2019	USA
GSE109887	GPL10904	32	46	2019	Germany

Table 1. Source and sample size of data used in the article.

Weighted gene co-expression network analysis (WGCNA) to screen target genes

WGCNA was performed to identify co-expression modules using the R package of “WGCNA” (version 1.72.1). The top 25% of genes with the highest variance were applied for subsequent WGCNA analyses to guarantee the accuracy of the quality results by checking the missing values and clustering the samples. The “soft” threshold power (β) was calculated to construct a biologically meaningful scale-free topological network. In addition, a topological overlap matrix was constructed based on the adjacency matrix, and a dynamic tree-cutting algorithm was used to merge similar modules. Additionally, gene saliency (GS), module affiliation (MM), and correlation coefficients between the gene modules and clinical features were calculated to visualize the characteristic gene network. Finally, the potential gene targets for Alzheimer’s disease were obtained by the intersection of DEGs and genes within the significant gene module¹⁴.

Protein–protein interaction (PPI) networks construction

The STRING database (<https://string-db.org/>) was used to construct PPI networks for the AD potential target by setting a confidence level (0.7), followed by Cytascape (3.8.2) software to view this graph. Based on this network, the genes were selected as biomarker genes in the pathological process of patients with AD for subsequent screening.

Screening of AD markers by the machine learning algorithm

Machine learning algorithms were utilized to analyze the central genes in the PPI network and identify characteristic markers of AD. Initially, Lasso regression analysis was conducted using the “glmnet” (4.1.6) R package for variable selection. This method introduced an L1 regularization term to reduce model variables while maintaining predictive accuracy. Cross-validation was performed to determine the optimal value of lambda for regularization parameter setting, yielding 19 important genes⁴⁹. Next, Support Vector Machine Recursive Feature Elimination (SVM-RFE) analysis was carried out using the R package “e1071” (version 1.7.13). It was configured with a linear kernel function and the penalty parameter C was adjusted to optimize the trade-off between bias and variance of the model. We identified 34 genes that exhibited significant discriminatory power through that analysis, representing the most informative features in our model⁵⁰. Finally, Random Forest (RF) analysis was performed using the ‘randomForest’ (version 4.7–1.1) R package. The model’s performance was optimized by adjusting the number of trees and other relevant parameters. Genes were evaluated based on their importance scores, and only those with scores greater than two were retained, as they significantly contributed to the predictive ability of our model. Ultimately, a set of significant genes above a predefined threshold was filtered out⁵¹. The genes identified by combining three methods were considered potential biomarkers of AD.

Building and interpreting machine learning models

Based on feature genes, we utilized the Scikit-learn (version 1.3.2) Python package to construct seven machine learning models: eXtreme Gradient Boosting (XGBoost), RF, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (Logistic), Gaussian Naive Bayes (GNB), and Multilayer Perceptron (MLP). XGBoost is an ensemble learning method built on the gradient boosting framework, known for its exceptional proficiency in handling large datasets and its remarkable speed and performance⁵². RF algorithm is a comprehensive learning approach that enhances prediction performance by constructing multiple decision trees and subsequent voting or averaging. Its advantages encompass robustness against noise, effective feature selection, insensitivity to missing values, and efficient training⁵³. KNN is an instance-based learning method that makes predictions by identifying the nearest neighbors among samples. Its strengths include simplicity, intuitive understanding, and the absence of training requirements⁵⁴. SVM is a robust classifier that discriminates between classes by identifying optimally separated hyperplanes. It excels in handling nonlinear problems through the use of kernel functions efficacy, utilization of kernel functions, and capacity to address nonlinear problems⁵⁵. Logistic is a linear model utilized for binary classification problems, wherein the results of linear regression are mapped to the interval [0, 1] via a sigmoid function. Its advantages include simplicity and interpretability. GNB is a classifier grounded on Bayes’ theorem, assuming independence among features. It offers computational efficiency and simplifies the assumption of conditional independence among features⁵⁶. MLP is a feed-forward neural network that processes information by traversing multiple layers of interconnected neurons. It is notable for its capacity for nonlinear modeling and inherent flexibility⁵⁷. Subsequently, we employed the Shapley Additive exPlanation (SHAP) values to analyze the significance of individual features and demonstrate their impact on the final machine learning model. Additionally, we used Local Interpretable Model-agnostic Explanations (LIME) algorithm to align and interpret the predictions of the optimal machine learning model⁵⁸.

Curve analysis of receiver operating characteristics (ROCs)

The “Corrplot” (0.92) R software package was used to analyze the correlation of AD biomarkers screened through machine learning. We used the “pROC” (1.18.0) R software package to create ROC curves and calculated the area under the curve (AUC) to evaluate the clinical diagnostic value of biomarkers⁵⁹.

Diagnostic column line graph construction and validation

The immune gene dataset is from the InnateDB database (<http://www.innatedb.com>), including 1696 immune genes⁶⁰. Those immune genes and identified potential biomarkers of AD were intersected to screen the immunity-related biomarkers. We used the “RMS” (6.5.0) R software package to construct a column line graph model to predict the incidence rate of AD⁶¹. Calibration curves were used to evaluate the accuracy of the column line graph model, and decision curves and clinical impact curves were employed for clinical utility evaluation⁶².

Enrichment and regulatory mechanism analysis of UBE2N

We performed GSEA analysis on the selected immune biomarker-UBE2N and utilized the Enrichr database (<https://maayanlab.cloud/Enrichr/>) to analyze the transcription factor (TF) of UBE2N. The miRTarBase (<https://mirtarbase.cuhk.edu.cn/>), Starbase (<https://starbase.sysu.edu.cn/>), and TargetScan (www.targetscan.org) databases were used to predict the miRNAs that regulate UBE2N translation. Then the regulatory network diagram of UBE2N was constructed by Cytoscape (3.8.2).

Animals

Tau^{P301S} transgenic mice [B6C3-Tg (Prnp-MAPT*P301S) PS19 Vle/J] and APP/PS1 mice were originally purchased from the Jackson Laboratory (Bar Harbor, ME, United States), and C57BL/6 mice were obtained from Beijing HuaFuKang Bioscience Co., Ltd. (Beijing, China) for animal mating. In the offspring, Tau^{P301S} transgenic and wild-type mice were obtained in the same month through genotype identification. They (n = 8 in each group) were housed under a 12-h light/dark cycle of 8:00/20:00 and controlled temperature (24 ± 2 °C) and humidity (40–70%) for 9 months till sacrificed under isoflurane. All authors complied with the ARRIVE guidelines. All treatments and experimental procedures were performed in accordance with the National Institutes of Health guidelines and were approved by the Northeastern University Laboratory Animal Ethical Committee (EC-2023A012).

Immunohistochemistry

The mice in the two experimental groups (n = 8 in each group) were anesthetized, and half the brains were removed, fixed using 4% paraformaldehyde, and embedded in paraffin. Serial 5-μm coronal sections were incubated with blocking solution (5% bovine serum albumin and 1% normal goat serum) for 1 h and then incubated overnight at 4 °C with rabbit anti-UBE2N (1:200, Abcam). The following day, the sections were incubated with biotinylated goat anti-rabbit IgG (1:500) for 1 h at room temperature (RT), followed by incubation with the avidin–biotin–peroxidase complex (1:100) for 30 min at RT. After washing with phosphate-buffered saline, the sections were immersed in 3,3′-diaminobenzidine for development. One section was incubated with normal rabbit serum (1:100) for nonspecific staining and served as the negative control. Images of immunohistochemical staining were captured using a light microscope (DM4000B; Leica, Wetzlar, Germany).

Immunofluorescence staining and confocal laser scanning microscopy

The 5-μm coronal sections were preincubated with blocking buffer for 1 h and then with rabbit anti-UBE2N (1:200, Abcam) and mouse monoclonal anti-NeuN antibodies (1:200, Thermo Fisher Scientific) overnight at 4 °C. Alexa Fluor® 488 and Alex Fluor® 594-conjugated secondary antibodies were mixed together and treated to sections for 2 h and finally labeled using DAPI (1:500). After mounting with an anti-fade mounting medium, the cells were imaged using a laser scanning confocal microscope (Leica, TCS, SP8, Wetzlar, Germany).

Consensus clustering and principal component analysis

Consensus clustering (CC) analysis was carried out using the R package “ConsensusClusterPlus” (version 1.6.0)⁶³. The algorithmic model was configured to implement agglomerative “pam” clustering, utilizing a Spearman correlation distance metric and resampling 80% of the sample population for a total of 1000 iterations.

scRNA-seq data preprocessing and analysis

The human AD scRNA-seq dataset is from the GEO database (GSE200164), and the raw gene expression data were converted to Seurat objects using the Seurat package (v4.3.0.1) in R (v4.3.1), retaining cells with less than 10% of mitochondrial genes. These Seurat objects were subsequently converted to h5ad objects using the “seasay” package (v0.0.7) and samples were corrected and integrated using scvi (v0.6.8). Clustering was performed using the Leiden algorithm with a resolution parameter set to 0.8. Upon completion of the above steps, the h5ad object was converted to a Seurat object. The highly expressed genes in all cell subpopulations were then analyzed in R using the “FindAllMarkers” function with default parameters.

Western blot

The temporal cortex and hippocampus of the half-brains of C57BL/6 mice and Tau^{P301S} mice were lysed in RIPA buffer to extract the protein supernatant. Proteins (10 μg) were separated by 4–12% SDS-PAGE and transferred to polyvinylidene fluoride membranes (Millipore, Burlington, MA, USA). The membranes were incubated in 5% BSA solution at room temperature for 1 h. Subsequently, the membranes were incubated overnight at 4 °C with rabbit anti-UBE2N (1:2000, Abcam) and mouse anti-GAPDH (1:10000, A1978, Sigma, Burlington, MA, USA) antibodies. Finally, the membranes were washed and incubated with horseradish peroxidase (HRP)-conjugated

Gene	Forward	Reverse
UBE2N	CCGCACAGTCTCTGCTATCAA	AGTCCATGCTCTCGCTGTTT
ATP6V1E1	CTTGTAACAGCTGCTGGAGCC	AGGCCTCCTGGTCAATCTGGA
CCKBR	GATGGCTGCTACGTGCAACT	CGCACCACCCGCTTCTTAG
SV2B	GCGGCTGGCTGATAAACT	AGAGGAAGGCTCCATATCCCT
TUBB2A	TGCCCTCACCCAAGGTCTCTG	GGCAGGTGGTCACTCCACTCA
SYT1	CTGTCACCACTGTTGCGAC	GGCAATGGGATTTTATGCAGTTC
DYNC1I1	GTCGTCATGGAAGCAAAGCA	AAGGAGTAGAGCGGCTTGTT
NRN1	GCGGTGCAAATAGCTTACCTG	TGATGTTCTGCTTGTCTGCCA

Table 2. Primer sequences for qRT-PCR.

secondary antibodies for 2 h. Bands were detected using a chemiluminescence imaging analysis system (Tanon 5500, Shanghai, China) and enhanced chemiluminescence (ECL) kits (EMD Millipore, Burlington, MA, USA). Each experiment was repeated at least thrice.

Quantitative reverse transcription polymerase chain reaction (qRT-PCR)

Total RNA was extracted from the cortex of C57BL/6 and Tau^{P301S} mice using Total RNA KIT I (R6834-02, OMGEA, USA), and 500 ng of template RNA was reverse transcribed into cDNA using the GoScript™ Reverse Transcription System (Promega, A5001) according to the manufacturer's instructions. PCR reactions were performed with 20 ng of cDNA template in a 10-μL reaction mixture using the Bio-Rad CFX PCR system. The sequences of the genes encoding GAPDH and selected differential genes were obtained from the GenBank database, and specific primers were designed using Primer Premier 5.0 (Table 2). The mRNA expression was calculated using equation $2^{-\Delta\Delta CT}$.

Statistical analysis

Data are presented as the mean ± SEM. The student's t-test was used to analyze differences between groups, as appropriate. Analyses were performed using the ImageJ software and GraphPad Prism 9.0. Statistical significance was set at $P < 0.05$.

Data availability

The datasets extracted and analyzed in this study are available in the GEO database: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281>; <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84422>; <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132903>; <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109887>; <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE200164>. Please contact the corresponding authors for any other requests.

Code availability

The code is available on GitHub at the following link: <https://zenodo.org/records/14854728>.

Received: 28 January 2024; Accepted: 13 February 2025

Published online: 22 February 2025

References

- Scheltens, P. et al. Alzheimer's disease. *The Lancet* **397**, 1577–1590 (2021).
- Long, J. M. & Holtzman, D. M. Alzheimer disease: An update on pathobiology and treatment strategies. *Cell* **179**, 312–339 (2019).
- Armstrong, R. A. Risk factors for Alzheimer's disease. *Folia Neuropathol* **57**, 87–105. <https://doi.org/10.5114/fn.2019.85929> (2019).
- Cummings, J. et al. Drug development in Alzheimer's disease: The path to 2025. *Alzheimers Res. Ther.* **8**, 39 (2016).
- Mahase, E. Alzheimer's disease: Lecanemab gets full FDA approval and black box safety warning. *BMJ* **382**, 1580. <https://doi.org/10.1136/bmj.p1580> (2023).
- Dickson, S. P., Hennessey, S., Nicodemus Johnson, J., Knowlton, N. & Hendrix, S. B. Avoiding future controversies in the Alzheimer's disease space through understanding the aducanumab data and FDA review. *Alzheimers Res Ther* **15**, 98. <https://doi.org/10.1186/s13195-023-01238-1> (2023).
- Qureshi, I. A. et al. A randomized, single ascending dose study of intravenous BIIB092 in healthy participants. *Alzheimers Dement (N Y)* **4**, 746–755. <https://doi.org/10.1016/j.trci.2018.10.007> (2018).
- Cummings, J. The role of biomarkers in Alzheimer's disease drug development. *Rev. Biomarker Stud. Psychiat. Neurodegener. Disord.* 29–61 (2019).
- Gate, D. et al. Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. *Nature* **577**, 399–404 (2020).
- Monson, A., Nemirovsky, A. & Harpaz, I. CD4 T cells in immunity and immunotherapy of Alzheimer's disease. *Immunology* **139**, 438–446 (2013).
- Kim, K. et al. Therapeutic B-cell depletion reverses progression of Alzheimer's disease. *Nat. Commun.* **12**, 2185 (2021).
- Town, T. et al. Blocking TGF-β–Smad2/3 innate immune signaling mitigates Alzheimer-like pathology. *Nat. Med.* **14**, 681–687 (2008).
- Wang, T. et al. Integrated bioinformatic analysis reveals YWHAB as a novel diagnostic biomarker for idiopathic pulmonary arterial hypertension. *J. Cell. Physiol.* **234**, 6449–6462 (2019).
- Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
- Langfelder, P. & Horvath, S. Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* **46**, i11 (2012).
- Ozsolak, F. & Milos, P. M. RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

17. Peiffer-Smadja, N. et al. Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. *Clin. Microbiol. Infect.* **26**, 584–595 (2020).
18. Chen, Y., Liao, R., Yao, Y., Wang, Q. & Fu, L. Machine learning to identify immune-related biomarkers of rheumatoid arthritis based on WGCNA network. *Clin. Rheumatol.* **41**, 1057–1068 (2022).
19. Li, J. et al. Identification of diagnostic genes for both Alzheimer's disease and metabolic syndrome by the machine learning algorithm. *Front. Immunol.* **13**, 1037318 (2022).
20. Zhao, X., Yao, H. & Li, X. Unearthing of key genes driving the pathogenesis of Alzheimer's disease via bioinformatics. *Front. Genet.* **12**, 641100 (2021).
21. Chai, K. et al. Application of weighted co-expression network analysis and machine learning to identify the pathological mechanism of Alzheimer's disease. *Front. Aging Neurosci.* **14**, 837770 (2022).
22. DeTure, M. A. & Dickson, D. W. The neuropathological diagnosis of Alzheimer's disease. *Mol. Neurodegener.* **14**, 1–18 (2019).
23. Liu, C. et al. Identification of immune cells and key genes associated with Alzheimer's disease. *Int. J. Med. Sci.* **19**, 112 (2022).
24. von Dadelszen, P. et al. Prediction of adverse maternal outcomes in pre-eclampsia: Development and validation of the fullPIERS model. *Lancet (London, England)* **377**, 219–227 (2011).
25. Yin, P. et al. Aged monkey brains reveal the role of ubiquitin-conjugating enzyme UBE2N in the synaptosomal accumulation of mutant huntingtin. *Hum. Mol. Genet.* **24**, 1350–1362. <https://doi.org/10.1093/hmg/ddu544> (2015).
26. Geisler, S., Vollmer, S., Golombek, S. & Kahle, P. J. The ubiquitin-conjugating enzymes UBE2N, UBE2L3 and UBE2D2/3 are essential for Parkin-dependent mitophagy. *J. Cell Sci.* **127**, 3280–3293. <https://doi.org/10.1242/jcs.146035> (2014).
27. Chang, J.-H. et al. Ubc13 maintains the suppressive function of regulatory T cells and prevents their conversion into effector-like T cells. *Nat. Immunol.* **13**, 481–490 (2012).
28. Carrano, A. et al. Amyloid Beta induces oxidative stress-mediated blood-brain barrier changes in capillary amyloid angiopathy. *Antioxid. Redox Signal.* **15**, 1167–1178 (2011).
29. Goldeck, D. et al. Enhanced chemokine receptor expression on leukocytes of patients with Alzheimer's disease. *PLoS ONE* **8**, e66664 (2013).
30. Giuliani, F., Goodyer, C. G., Antel, J. P. & Yong, V. W. Vulnerability of human neurons to T cell-mediated cytotoxicity. *J. Immunol.* **171**, 368–379 (2003).
31. Mietelska-Porowska, A. & Wojda, U. T lymphocytes and inflammatory mediators in the interplay between brain and blood in Alzheimer's disease: Potential pools of new biomarkers. *J. Immunol. Res.* **2017**, 4626540 (2017).
32. Lindestam-Arlehamn, C. S. et al. Widespread tau-specific CD4 T cell reactivity in the general population. *J. Immunol.* **203**, 84–92 (2019).
33. Merlini, M., Kirabali, T., Kulic, L., Nitsch, R. M. & Ferretti, M. T. Extravascular CD3+ T cells in brains of Alzheimer disease patients correlate with tau but not with amyloid pathology: An immunohistochemical study. *Neuro-Degenerat. Dis.* **18**, 49–56 (2018).
34. Chang, S.-H. et al. Proteomic profiling of brain cortex tissues in a Tau transgenic mouse model of Alzheimer's disease. *Biochem. Biophys. Res. Commun.* **430**, 670–675 (2013).
35. Paunescu, T. G., Jones, A. C., Tyszkowski, R. & Brown, D. V-ATPase expression in the mouse olfactory epithelium. *Am. J. Physiol.-Cell Physiol.* **295**, C923–C930 (2008).
36. Yu, H. et al. Integrated transcriptomics reveals the brain and blood biomarkers in Alzheimer's disease. *CNS Neurosci. Therapeut.* (2023).
37. Nishimura, S. et al. Functional synergy between cholecystokinin receptors CCKAR and CCKBR in mammalian brain development. *PLoS One* **10**, e0124295 (2015).
38. Zhang, Z. et al. Cholecystokinin signaling can rescue cognition and synaptic plasticity in the APP/PS1 mouse model of Alzheimer's disease. *Mol. Neurobiol.* 1–23 (2023).
39. Miyamoto, M. et al. Synaptic vesicle protein 2B negatively regulates the amyloidogenic processing of A β PP as a novel interaction partner of BACE1. *J. Alzheimers Dis.* **75**, 173–185 (2020).
40. Detrait, E., Maurice, T., Hanon, E., Leclercq, K. & Lamberty, Y. Lack of synaptic vesicle protein SV2B protects against amyloid- β 25–35-induced oxidative stress, cholinergic deficit and cognitive impairment in mice. *Behav. Brain Res.* **271**, 277–285 (2014).
41. Ficulle, E. et al. A human tau seeded neuronal cell model recapitulates molecular responses associated with Alzheimer's disease. *Sci. Rep.* **12**, 2673 (2022).
42. Ha, J., Park, C., Park, C. & Park, S. IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization. *J. Biomed. Inform.* **102**, 103358 (2020).
43. Ha, J. MDMF: Predicting miRNA-disease association based on matrix factorization with disease similarity constraint. *J. Personal. Med.* **12**, 885 (2022).
44. Ha, J. & Park, S. NCMD: Node2vec-based neural collaborative filtering for predicting miRNA-disease association. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 1257–1268 (2022).
45. Ha, J. SMAP: Similarity-based matrix factorization framework for inferring miRNA-disease association. *Knowl.-Based Syst.* **263**, 110295 (2023).
46. Chen, X., Li, T. H., Zhao, Y., Wang, C. C. & Zhu, C. C. Deep-brief network for predicting potential miRNA-disease associations. *Brief. Bioinform.* <https://doi.org/10.1093/bib/bbaa186> (2021).
47. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)* **28**, 882–883 (2012).
48. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl. Acids Res.* **43**, e47 (2015).
49. Wang, S. et al. An Eight-CircRNA assessment model for predicting biochemical recurrence in prostate cancer. *Front. Cell Dev. Biol.* **8**, 599494 (2020).
50. Sanz, H., Valim, C., Vegas, E., Oller, J. M. & Reverter, F. SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinform.* **19**, 432 (2018).
51. Yang, Y., Xu, L., Qiao, Y., Wang, T. & Zheng, Q. Construction of a neural network diagnostic model and investigation of immune infiltration characteristics for Crohn's disease. *Front. Genet.* **13**, 976578 (2022).
52. Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).
53. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
54. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13**, 21–27 (1967).
55. Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **10**, 988–999 (1999).
56. John, G. H. & Langley, P. Estimating continuous distributions in Bayesian classifiers. *arXiv preprint arXiv:1302.4964* (2013).
57. Haykin, S. *Neural Networks: A Comprehensive Foundation*. (Prentice Hall PTR, 1998).
58. Lundberg, S. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).
59. Cheng, S. et al. Novel blood test for early biomarkers of preeclampsia and Alzheimer's disease. *Sci. Rep.* **11**, 15934 (2021).
60. Breuer, K. et al. InnateDB: Systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucl. Acids Res.* **41**, D1228–1233 (2013).
61. Zhang, J.-A. et al. Development of an immune-related gene signature for prognosis in melanoma. *Front. Oncol.* **10**, 602555 (2020).
62. Liu, C. et al. Identification of a glycolysis- and lactate-related gene signature for predicting prognosis, immune microenvironment, and drug candidates in colon adenocarcinoma. *Front. Cell Dev. Biol.* **10**, 971992 (2022).

63. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
64. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30. <https://doi.org/10.1093/nar/28.1.27> (2000).
65. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucl. Acids Res.* **51**, D587–D592. <https://doi.org/10.1093/nar/gkac963> (2023).

Acknowledgements

This study was supported by Shenzhen Natural Science Foundation-The Stable Support Program (20220810144826003), the Research Start-up Fund for Young Investigators in Shenzhen University (QNJS0384), the Construction Project of Liaoning Provincial Key Laboratory, China (2022JH13/10200026), the Special Projects of the Central Government in Guidance of Local Science and Technology Development (2022JH6/100100025) and the National Natural Science Foundation of China (81771174, 81971015).

Author contributions

G.F. performed the experimental procedures and statistical analysis. M.Z. and H.H. assisted with data analysis. P.Z., X.Z., and T.W. generated and validated the mouse model and performed animal experiments. H.G. conceived the experiments and wrote the manuscript. H.X. conceived the experiments, supervised the project and revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-90578-z>.

Correspondence and requests for materials should be addressed to H.G. or H.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025