Journal of Clinical Laboratory Analysis

WILEY

RESEARCH ARTICLE OPEN ACCESS

Multi-Omics and Machine Learning Integration Identifies Key Endothelial Modules in Carotid Artery Stenosis

Jiayi Wu[1] | Chunguang Guo[2] | Linfeng Zhang[3] | Gaopo Cai[3] | Yi Liu[2] | Kai Tang[2]

[1]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Department of Pathology, Peking University Cancer Hospital and Institute, Beijing, China | [2]Department of Cardiovascular Surgery, State Key Laboratory of Cardiovascular Disease, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences, Peking Union Medical College, Fuwai Hospital, Beijing, China | [3]Department of Endovascular Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China

Correspondence: Kai Tang (tangkai_fuwai@163.com)

Received: 23 December 2025 | **Revised:** 8 April 2026 | **Accepted:** 13 May 2026

Keywords: biomarkers | carotid artery stenosis | endothelial cell | machine learning | single-cell RNA sequencing

ABSTRACT

Background: Carotid artery stenosis (CAS) is a major cause of ischemic stroke, yet reliable molecular biomarkers for early identification remain limited.

Methods: We integrated single-cell RNA sequencing (scRNA-seq), bulk transcriptomics, and in-house multi-omics data, applying WGCNA and machine learning to identify endothelial cell-derived diagnostic biomarkers, validated across independent GEO and ZZ cohorts at single-cell, transcriptomic, and proteomic levels.

Results: scRNA-seq identified endothelial enrichment in CAS and yielded 836 markers. Integrative analysis (DEGs + WGCNA) defined 80 candidates, with NRP1 and XAF1 selected by machine learning. Both were consistently upregulated and validated across multi-omics datasets, showing strong diagnostic performance.

Conclusion: NRP1 and XAF1 represent novel endothelial cell-derived biomarkers with potential utility for early CAS screening and clinical diagnosis.

1 | Introduction

Carotid artery stenosis (CAS), a pathological narrowing of the carotid arteries caused primarily by atherosclerotic plaque accumulation, is a major contributor to ischemic stroke worldwide. Stroke remains the second leading cause of death and the third leading cause of disability-adjusted life years globally [1, 2]. Among stroke subtypes, CAS accounts for approximately 10%–20% of ischemic events, particularly in aging populations [3]. According to recent epidemiological modeling, over 57 million individuals aged 30–79 years were estimated to be affected by CAS globally in 2020, with the highest burden observed in the Western Pacific region, including China [4]. Despite its clinical significance, CAS is often asymptomatic in early stages. Current diagnostic methods, such as carotid artery ultrasound or CT angiography, mainly focus on the degree of lumen stenosis and have limited understanding of potential molecular or cellular changes. This limitation hinders early risk stratification and targeted intervention and also indicates the current lack of carotid plaque biomarkers.

In recent years, bulk transcriptomic analyses have enabled global gene expression profiling in vascular tissues and contributed to the identification of differentially expressed genes (DEGs) in CAS [5]. However, bulk RNA sequencing represents an average signal across diverse cell types within a sample, masking cell-specific regulatory signatures critical to disease progression. The advent of single-cell RNA sequencing (scRNA-seq) has transformed our ability to characterize complex tissues like atherosclerotic plaques with unprecedented resolution [6, 7]. By dissecting gene expression patterns at the single-cell level, scRNA-seq allows for the identification of rare or

disease-driving cell populations, mapping of intercellular communication networks, and reconstruction of cellular trajectories within dynamic pathological environments [8, 9]. In the context of atherosclerosis, scRNA-seq has revealed previously unappreciated heterogeneity in endothelial cells, smooth muscle cell phenotypic transitions, and immune cell activation states that contribute to plaque instability [10].

Despite these advances, several challenges persist in CAS research. First, studies often rely on small sample sizes or limited tissue types, restricting generalizability. Second, there is a lack of integrative analytic frameworks that bridge single-cell resolution with bulk cohort-level validation. Third, few studies have harnessed machine learning techniques to prioritize key molecular features from high-dimensional data in a clinically translatable manner. Recent efforts combining weighted gene co-expression network analysis (WGCNA) and machine learning have shown promise in identifying disease-relevant gene modules and refining biomarker panels for complex cardiovascular disorders [11, 12]. However, such approaches remain underexplored in CAS, particularly in relation to endothelial cell–driven molecular pathology and immune involvement.

In this study, we integrated single-cell RNA sequencing, WGCNA, and machine learning to identify reliable biomarkers associated with carotid plaque formation. Endothelial cell–specific genes were identified from scRNA-seq data and intersected with DEGs and co-expression modules from bulk transcriptomic profiles. Random forest and LASSO regression were used to prioritize key genes, which were validated in an external dataset. Immune infiltration analysis further explored their potential immunological relevance. This integrative approach offers novel insights into CAS pathogenesis and highlights potential diagnostic markers.

2 | Materials and Methods

2.1 | Data Acquisition and Preprocessing

Transcriptomic datasets related to carotid artery stenosis (CAS) and normal arterial tissues were obtained from the Gene Expression Omnibus (GEO) database. GSE43292 includes 32 CAS plaque samples (CS group) and 32 normal carotid artery samples (Con group) [13]. GSE28829 dataset consists of 16 CAS samples and 13 healthy control samples [14]. In addition, to further investigate the findings at the single-cell level, we incorporated the GSE159677 dataset [15], which contains carotid tissue samples from three patients with CAS and three control individuals.

In addition, one in-house bulk RNA sequencing cohort (ZZ-Cohort) was generated from carotid plaque specimens collected during carotid endarterectomy at the Department of Vascular Surgery, The First Affiliated Hospital of Zhengzhou University. All specimens were promptly preserved and processed for RNA sequencing using standardized protocols. To complement this dataset, an in-house single-cell RNA sequencing (scRNA-seq) dataset was also established, including seven carotid plaque samples (six from CAS patients and one from a control individual) collected at the same institution. All datasets underwent standardized preprocessing and normalization. For the GEO datasets, raw expression matrices were log2-transformed and quantile-normalized. Batch effects across datasets were corrected using the "ComBat" function in the sva R package. For the single-cell data, low-quality cells with <200 detected genes, $>5\%$ mitochondrial gene content, or $>20{,}000$ total counts were removed. Normalization, scaling, and dimensionality reduction were performed using Seurat Package. The detailed information of the patient can be found in Table S1.

2.2 | Single-Cell RNA-Seq Data Processing

We analyzed single-cell RNA sequencing data from the GSE159677 dataset, which includes carotid tissue samples from three patients with carotid plaques and three healthy controls. Data preprocessing and dimensionality reduction were performed using the *Seurat* R package [16]. Cells with fewer than 500 total counts (nCount_RNA <500) or mitochondrial gene content greater than 20% were excluded to ensure data quality. Principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) were used for dimensionality reduction and visualization. According to previous studies, cell type annotation was manually performed based on canonical marker gene expression profiles, including PECAM1, VWF, CDH5 for endothelial cells; ACTA2, MYH11 for smooth muscle cells; CD68, CD14 for macrophages; CD3E, CD4 for T cells; CD19, CD79A for B cells; COL1A1, DCN for fibroblasts; and TPSAB1, TPSB2 for mast cells.

2.3 | Identification of Endothelial Cell Markers

To investigate the potential contributions of different cell types to carotid plaque formation, we compared the relative proportions of each cell type between CAS patients and healthy controls. Cell composition differences were visualized using boxplots generated with the *ggplot2* R package. Cell types that exhibited significantly increased proportions in the CAS group were selected for further analysis. Subsequently, marker genes were identified using the FindMarkers function in Seurat, with selection criteria set to an adjusted p-value <0.05 and an absolute log2 fold change >0.25 [17, 18].

2.4 | Enrichment Analysis

To explore the biological functions associated with endothelial cell–specific genes, we performed functional enrichment analysis. This approach is commonly used in transcriptomic studies to infer the functional relevance of a set of genes. Two widely adopted enrichment strategies were applied: Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. GO enrichment provides insights into the molecular functions, biological processes, and cellular components related to the input gene set, as defined by the Gene Ontology Consortium. KEGG analysis integrates genomic and pathway information to predict the involvement of genes in signaling cascades and metabolic pathways. These analyses help elucidate the potential roles of differentially expressed genes in disease mechanisms. Enrichment significance was assessed based on

adjusted *p*-values, and results were visualized using color gradients to indicate statistical significance.

2.5 | Co-Expression Network Analysis

To identify gene modules associated with carotid plaque formation, we performed weighted gene co-expression network analysis (WGCNA) on the GSE43292 dataset using the WGCNA R package [19]. WGCNA is a systems biology method used to describe correlation patterns among genes across multiple samples, grouping genes with similar expression profiles into distinct modules. In this study, we first calculated the median absolute deviation (MAD) for each gene and selected the top 5000 most variable genes for network construction to reduce noise and improve computational efficiency. An unsigned co-expression network was built by computing pairwise Pearson correlations between genes, followed by transformation into an adjacency matrix and topological overlap matrix (TOM). A soft-thresholding power (β) was chosen to ensure scale-free topology. Hierarchical clustering was applied to the TOM, and modules were defined using a dynamic tree cut algorithm. Each module was assigned a unique color label. Module–trait relationships were assessed by correlating the module eigengenes with clinical phenotypes (e.g., CS vs. Con group status). The module showing the strongest positive or negative correlation with plaque phenotype was selected for downstream analysis.

2.6 | Differential Gene Expression Analysis

Differential gene expression analysis between CAS and control samples was conducted using the *limma* package on the GSE43292 dataset. Raw expression data were background-corrected, log2-transformed, and normalized prior to linear modeling. Statistical significance was assessed using empirical Bayes moderated *t*-statistics, with *p*-values adjusted by the Benjamini–Hochberg method to control the false discovery rate (FDR). Genes with a p value < 0.05 and $|\log2 \text{ fold change}| > 0.585$ were considered differentially expressed. A volcano plot was generated using the ggplot2 package to visualize the distribution of significant genes.

2.7 | Machine Learning

To identify key molecular features associated with carotid plaque formation, we first derived the intersection of endothelial cell markers, differentially expressed genes (DEGs), and genes from WGCNA modules most significantly correlated with the CAS phenotype. The resulting gene set was subjected to machine learning–based feature selection. A random forest classifier was implemented to assess the relative importance of each gene in distinguishing CAS from control samples. The random forest model was constructed using the randomForest R package, with the number of trees set to 500 and the number of variables tried at each split (mtry) determined by the default value (square root of the number of features). Feature importance was ranked by mean decrease in Gini impurity, and the top-ranking genes were selected for subsequent LASSO analysis. LASSO regression was performed using the glmnet R package with 10-fold cross-validation. The optimal regularization parameter (lambda) was selected as lambda.min, corresponding to the value that minimized the cross-validation mean squared error [20]. Through the combination of the above algorithms, the genes that are most likely to be related to the formation process of carotid artery plaques were identified.

2.8 | Diagnostic Model Construction and Validation

To construct a diagnostic model for CAS, the selected key genes were incorporated into a logistic regression framework using the rms package. A nomogram was generated to provide a visual representation of the predictive model, assigning risk scores to individual gene expression values to estimate the probability of CAS. Model performance was evaluated using receiver operating characteristic (ROC) curve analysis, and the area under the curve (AUC) was calculated with the pROC package to assess discriminatory ability. Calibration of the model was assessed using bootstrap resampling ($B = 1000$) to compare predicted probabilities against observed outcomes. Internal validation was conducted within the GSE43292 dataset. To evaluate the generalizability of the model, external validation was performed using an independent dataset, GSE28829. ROC curves and gene expression comparisons were also generated in the validation cohort to verify diagnostic robustness across datasets. The detailed information of the patient can be found in Table S1.

2.9 | Immune Infiltration Analysis

To investigate the immunological landscape associated with CAS, we performed immune cell infiltration analysis using the single-sample gene set enrichment analysis (ssGSEA) algorithm [21]. The analysis was conducted with the GSVA R package, based on predefined gene sets representing 28 immune cell types, as previously described by Bindea et al. [22] and curated from the ImmPort database, including various subsets of T cells, B cells, dendritic cells, macrophages, and natural killer (NK) cells. Relative enrichment scores were computed for each immune cell type in each sample from the GSE43292 dataset. Differences in immune cell infiltration levels between CAS and control groups were visualized using boxplots generated with the ggplot2 package. To explore the immunological relevance of the candidate biomarkers, Spearman correlation analysis was conducted to assess the association between expression levels of selected genes and immune infiltration scores. Correlation heatmaps were generated to highlight significant gene-immune cell interactions.

2.10 | Validation of Key Genes in the ZZ Cohorts

To verify the robustness of the identified biomarkers, the expression of NRP1 and XAF1 was validated in the in-house ZZ cohorts using multi-omics data. Their expression patterns across different cell types were first examined in the ZZ single-cell RNA sequencing dataset (seven carotid plaque samples, including six CAS and one control). The scRNA-seq data were processed using standard quality control,

normalization, and clustering procedures, and gene expression levels were compared between groups. Subsequently, the transcriptional expression of both genes was assessed in the ZZ bulk RNA-seq cohort containing carotid plaques from CAS and control individuals, with differential analysis performed using the limma package. Finally, protein expression levels were validated using proteomic data from five paired CAS and control plaque samples, analyzed by label-free quantification (LFQ) and compared using a Student's *t*-test. Consistent trends across single-cell, transcriptomic, and proteomic levels confirmed the reliability of NRP1 and XAF1 expression alterations in CAS.

2.11 | Statistical Analysis

Statistical significance was defined as a two-tailed p-value < 0.05 or a false discovery rate (FDR) < 0.05. For descriptive statistics, normally distributed continuous variables were summarized as mean ± standard deviation, while non-normally distributed variables were reported as median (range). Group comparisons were performed using the Wilcoxon rank-sum test. Pearson or Spearman correlation analysis was applied based on data distribution. ROC curves were used to evaluate diagnostic performance, and the area under the curve (AUC) was calculated. All statistical analyses and data visualizations were conducted using R software (v4.3.2).

3 | Results

3.1 | Single-Cell Sequencing Analysis

Firstly, we removed batch effects from the single-cell dataset. As shown in Figure 1A, data uniformity was markedly improved after Harmony correction, and batch effects were significantly reduced. Subsequently, cells with a high proportion of mitochondrial gene expression (> 20%) were filtered out during dimensionality reduction and preprocessing (Figure S1). A total of 47,472 cells were retained, including 11,214 cells from the CAS group and 36,258 cells from the control group. The lower cell yield in carotid plaque tissues may be attributed to the abundant presence of lipids, thrombi, and necrotic components, which impair single-cell dissociation and capture compared to normal vascular tissue. We performed clustering analysis on the retained cells and identified 26 distinct cell clusters (Figure 1B). It is well established that different cell types exhibit specific transcriptomic expression profiles under physiological conditions, with certain genes being highly or uniquely expressed in particular cell types. Based on previous studies and known marker genes, we annotated the 26 clusters into seven major cell types (Figure 1C): T cells (18,464), endothelial cells (6354), fibroblasts (3788), macrophages (10,104), smooth muscle cells (6366), B cells (1857), and mast cells (539). A small number of cells could not be annotated due to the absence of specific marker gene expression. To further explore the key cell populations involved in carotid plaque formation, we compared the proportions of these seven cell types between the two groups. Violin plots revealed that endothelial cells showed the most significant difference in abundance between CAS and control samples (Figure 1D). We also generated bar plots to illustrate the relative proportions of each cell type between the two groups. Finally, the FindMarkers function was used to identify marker genes of endothelial cells, resulting in 836 endothelial-specific genes for subsequent analyses.

3.2 | Functional Enrichment of Endothelial Cell Marker Genes

To investigate the biological functions associated with endothelial cell–specific genes, GO and KEGG enrichment analyses were performed on the 836 marker genes. GO analysis revealed significant enrichment in terms related to vascular development, angiogenesis, endothelial cell migration, and adhesion (Figure S2A). KEGG pathway analysis further identified key signaling pathways, including fluid shear stress and atherosclerosis, PI3K-Akt signaling, focal adhesion, and leukocyte transendothelial migration (Figure S2B), suggesting that these endothelial genes are functionally involved in vascular remodeling and inflammation during plaque progression.

3.3 | Brown and Black Modules Were Identified as Key Modules

WGCNA is widely used to identify gene co-expression networks and uncover gene modules that may play critical roles in disease pathogenesis. Transcriptomic data from 64 patients in the GSE43292 dataset were analyzed, encompassing a total of 21,580 genes. Genes were ranked by median absolute deviation (MAD), and the top 5000 most variable genes were selected for network construction. To ensure scale-free topology and robust gene connectivity, a soft-thresholding power of 12 was chosen as the optimal β value (Figure 2A). Network modules were then identified using a dynamic tree cut algorithm. Modules with eigengene similarity greater than 75% were subsequently merged, resulting in a final set of 29 gene modules (Figure 2B). To visualize the topological structure of the network, a TOM heatmap was generated based on a subset of 400 genes. As shown in Figure 2C, genes with high topological overlap are clustered together and displayed as red blocks along the diagonal, indicating strong intra-module connectivity. Importantly, we assessed the correlation between each module and clinical traits. As shown in Figure 2D, both the brown and black modules showed significant correlations with the presence of carotid artery stenosis. The brown module contained 1626 genes, while the black module consisted of 2027 genes, and both were selected for downstream analysis as key trait-associated modules.

3.4 | Identification of Consensus Genes

Substantial differences in gene expression between CAS patients and healthy controls may underlie the formation and progression of carotid plaques. Differential expression analysis was performed using the limma package, identifying 488 upregulated and 370 downregulated genes in the CAS group compared to controls (Figure 3A). To further narrow down genes potentially involved in plaque development, we intersected three gene sets: endothelial cell–specific markers, genes from trait-associated WGCNA modules (brown and black), and

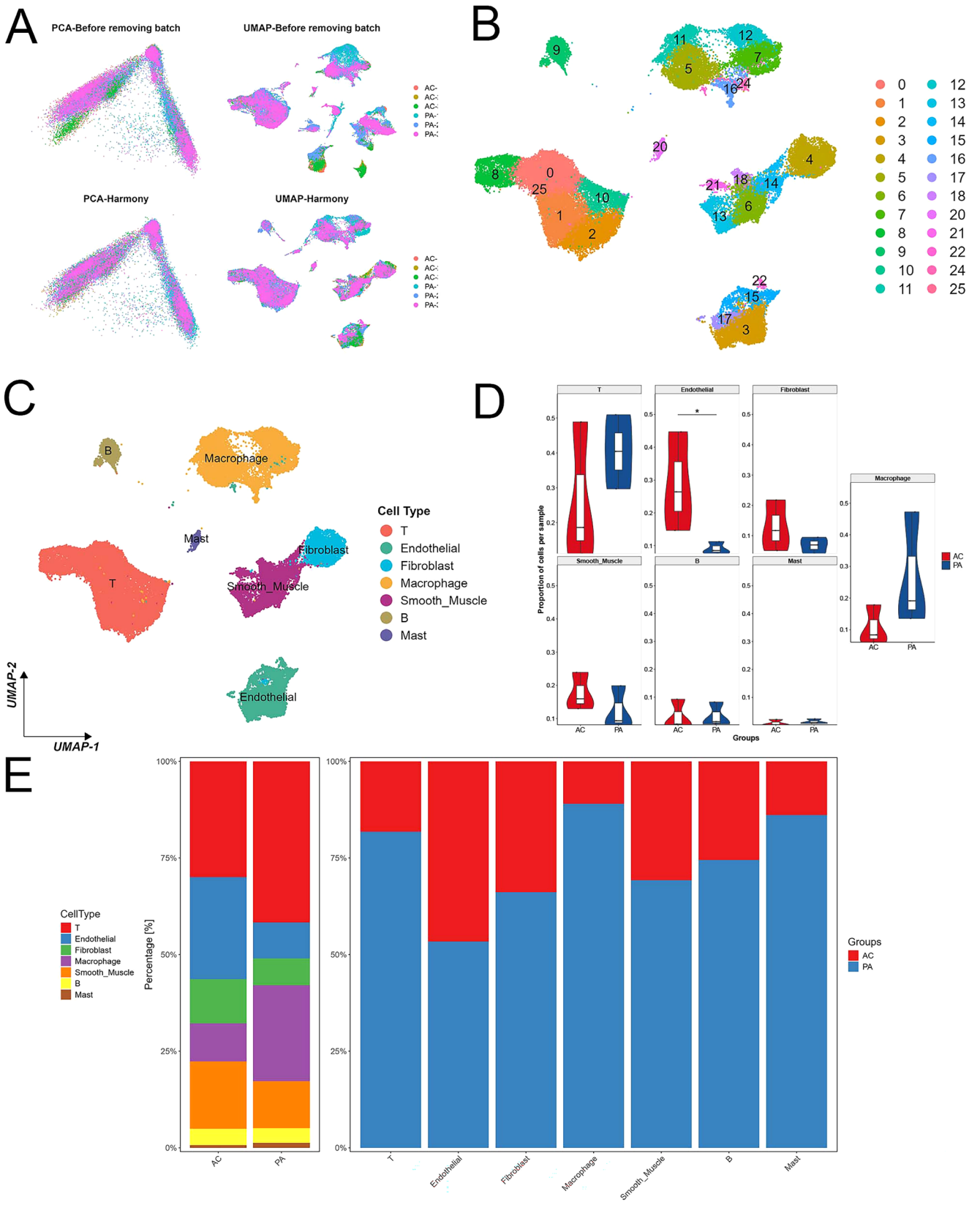

FIGURE 1 | Identification of endothelial cell markers using single-cell RNA-seq. (A) Batch effect correction via Harmony integration. (B, C) Cell clustering and annotation showing seven cell types. (D) Boxplot comparing cell-type proportions between CAS and control samples. (E) Bar plot showing cell composition per sample group.

differentially expressed genes (DEGs). As shown in the Venn diagram (Figure 3B), this intersection yielded 80 consensus genes for downstream analysis. Although all 80 genes may be associated with carotid plaque formation, their relative importance likely varies. To identify the most influential candidate genes, we applied a two-step machine learning approach. First,

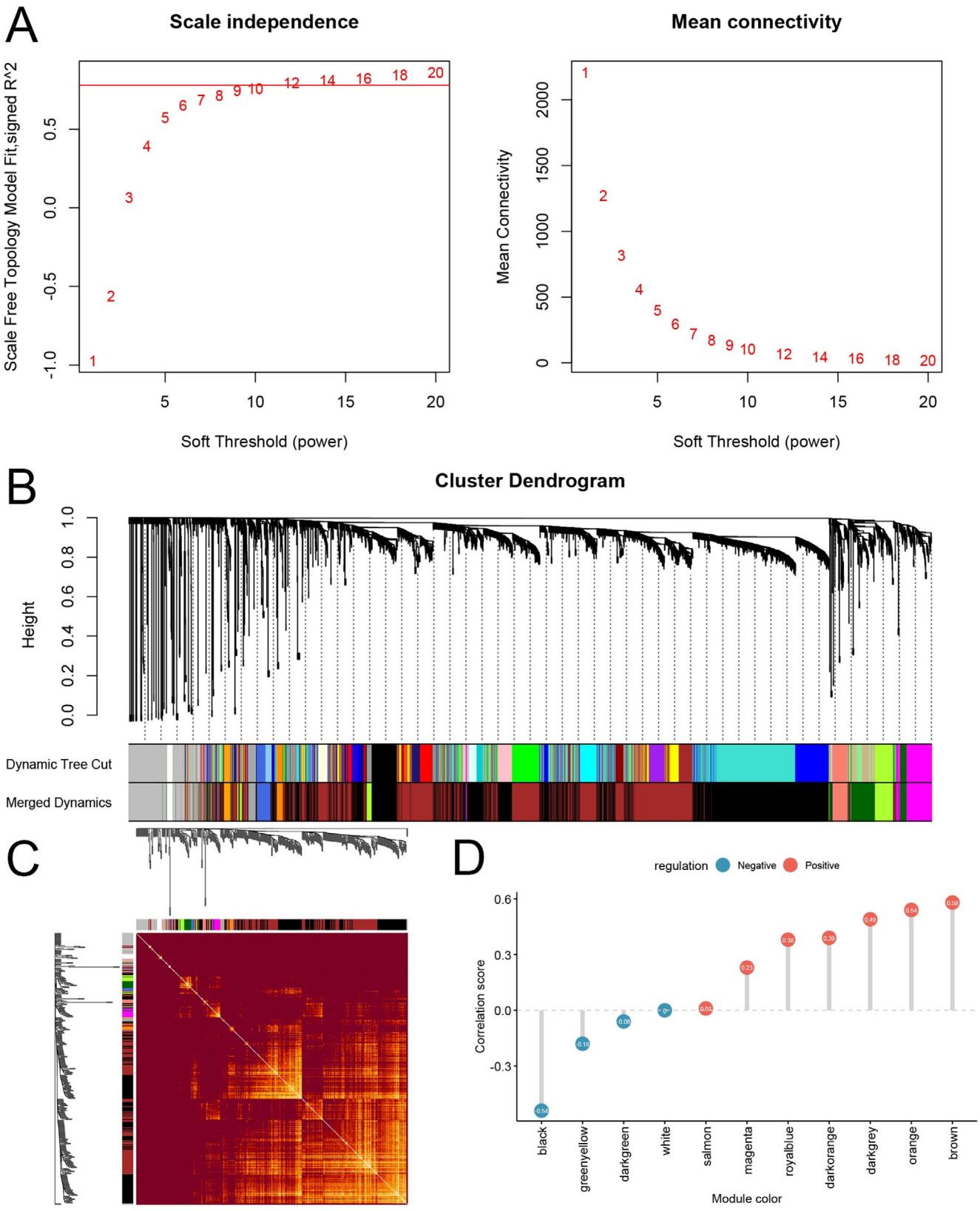

FIGURE 2 | WGCNA identifies modules associated with CAS. (A) Scale-free topology fit and soft threshold ($\beta = 12$). (B) Hierarchical clustering and dynamic tree cutting of gene modules. (C) TOM matrix heatmap. (D) Module-trait correlation heatmap.

a random forest classifier was used to rank gene importance, resulting in the selection of 10 key genes, including NRP1, PLCG2, CAVIN3, PLXND1, and others (Figure 3C,D). These genes were then subjected to LASSO regression, which further refined the list and identified NRP1 and XAF1 as the most critical genes associated with CAS (Figure 3E,F). These two genes were thus considered potential core biomarkers for the diagnosis and progression of carotid artery stenosis.

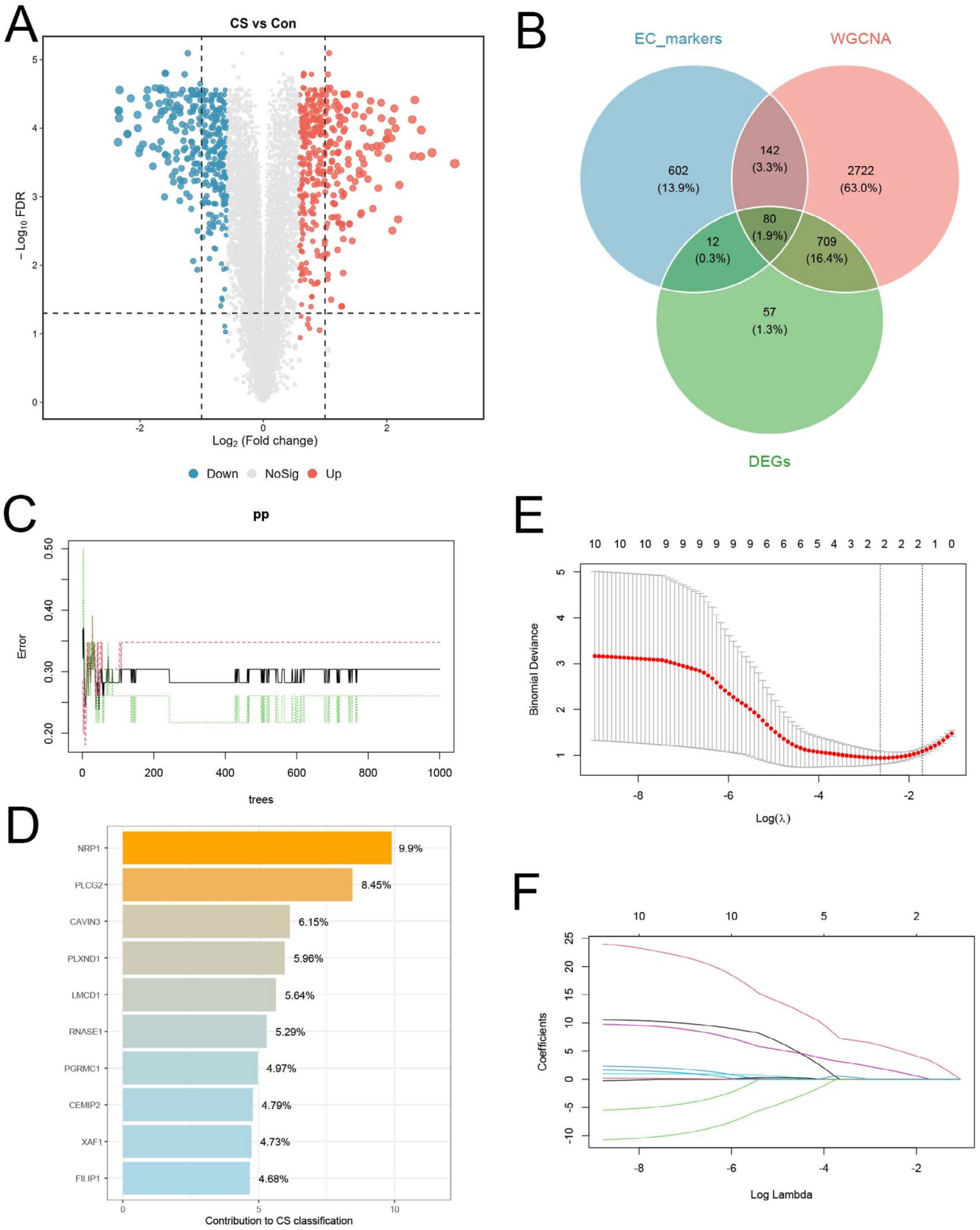

FIGURE 3 | Machine learning identifies diagnostic biomarkers. (A) Volcano plot of DEGs between CAS and controls. (B) Venn diagram of DEGs, WGCNA, and endothelial marker genes. (C, D) Random Forest ranking and importance scores for top 10 genes. (E, F) LASSO cross-validation and coefficient path.

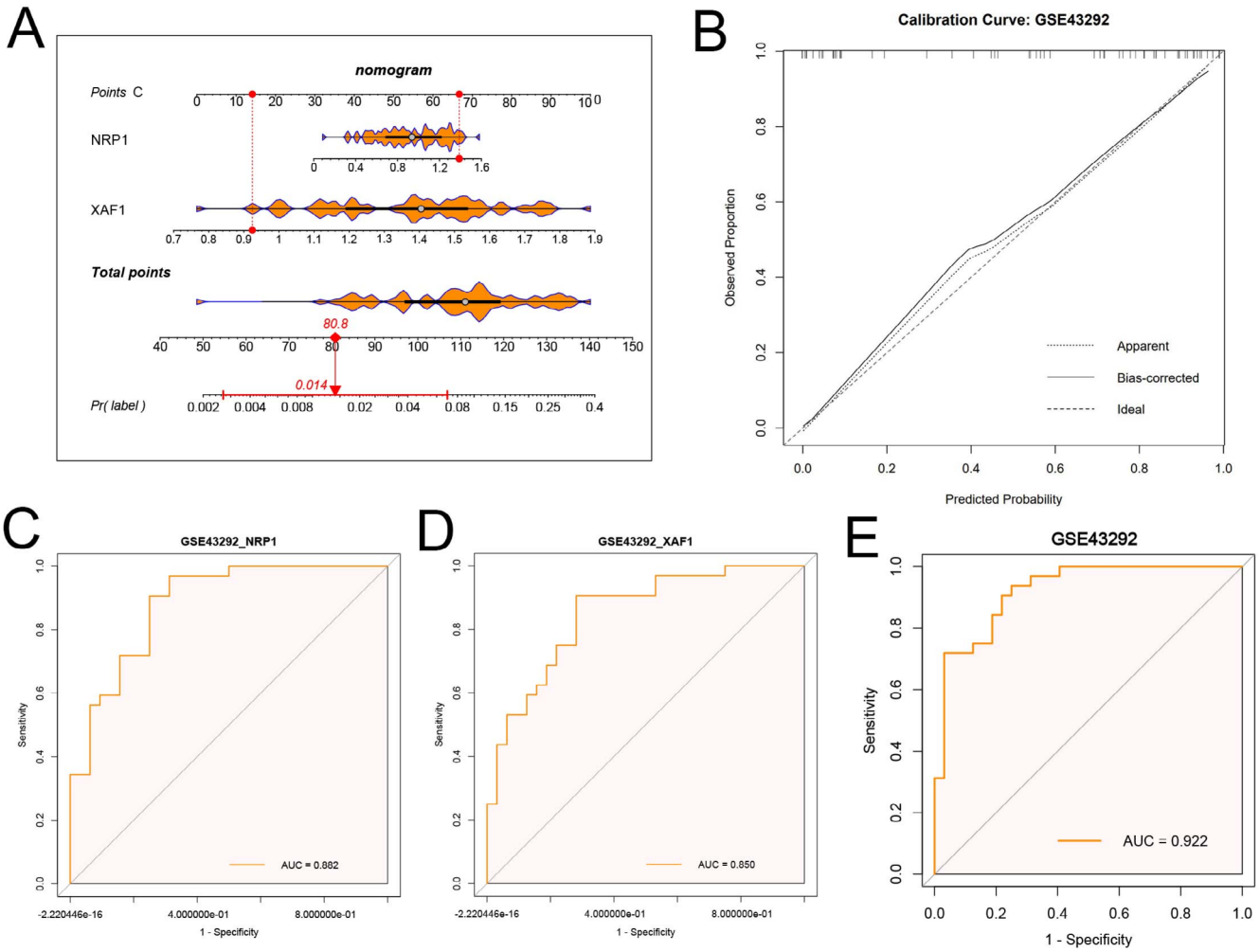

FIGURE 4 | Nomogram-based diagnostic model for CAS. (A) Nomogram constructed using NRP1 and XAF1. (B) Calibration curve shows agreement between predicted and observed probabilities. (C–E) ROC curves for NRP1, XAF1, and combined genes.

3.5 | Diagnostic Performance Evaluation of NRP1 and XAF1

Subsequently, the diagnostic performance of NRP1 and XAF1 for the target condition was systematically evaluated. The nomogram (Figure 4A) visually integrated the contributions of NRP1 (range: 0–1.6 points) and XAF1 (0.7–1.9 points) to generate a total score predictive of disease probability, with an optimal cutoff at 80.8 points yielding a sensitivity of 0.014. The calibration curve (Figure 4B) demonstrated good agreement between predicted and observed probabilities, with the bias-corrected curve closely aligning with the ideal line. The ROC curves further demonstrated the strong diagnostic performance of the selected genes. The area under the curve (AUC) for NRP1 and XAF1 was 0.882 (Figure 4C) and 0.850 (Figure 4D), respectively. Notably, the combined model achieved an even higher AUC of 0.922 (Figure 4E), indicating superior diagnostic accuracy for carotid artery stenosis.

3.6 | External Validation of Diagnostic Biomarkers

To evaluate the generalizability of the diagnostic model, NRP1 and XAF1 were validated in the independent GSE28829 dataset. ROC curve analysis showed that NRP1 retained strong diagnostic power with an AUC of 0.808 (Figure 5A), while XAF1 demonstrated moderate performance (AUC = 0.649, Figure 5B). The combined model achieved an AUC of 0.784 (Figure 5C), supporting its applicability in external cohorts. Expression level comparisons further confirmed the robustness of these biomarkers. In both GSE43292 and GSE28829, NRP1 expression was significantly elevated in CAS patients compared to controls (Figure 5D,F). XAF1 also showed a significant increase in the GSE43292 dataset (Figure 5E), while a trend toward higher expression was observed in GSE28829 (Figure 5G). These results validate the consistency of NRP1 and XAF1 expression patterns and diagnostic potential across multiple cohorts.

3.7 | Validation of NRP1 and XAF1 Expression in the ZZ Cohorts

Using the in-house ZZ cohorts, we validated the expression of NRP1 and XAF1 at the single-cell, transcriptomic, and proteomic levels. Single-cell RNA-seq analysis showed that both genes were mainly expressed in endothelial cells (Figure 6A). Violin plots demonstrated that NRP1 and XAF1 expression

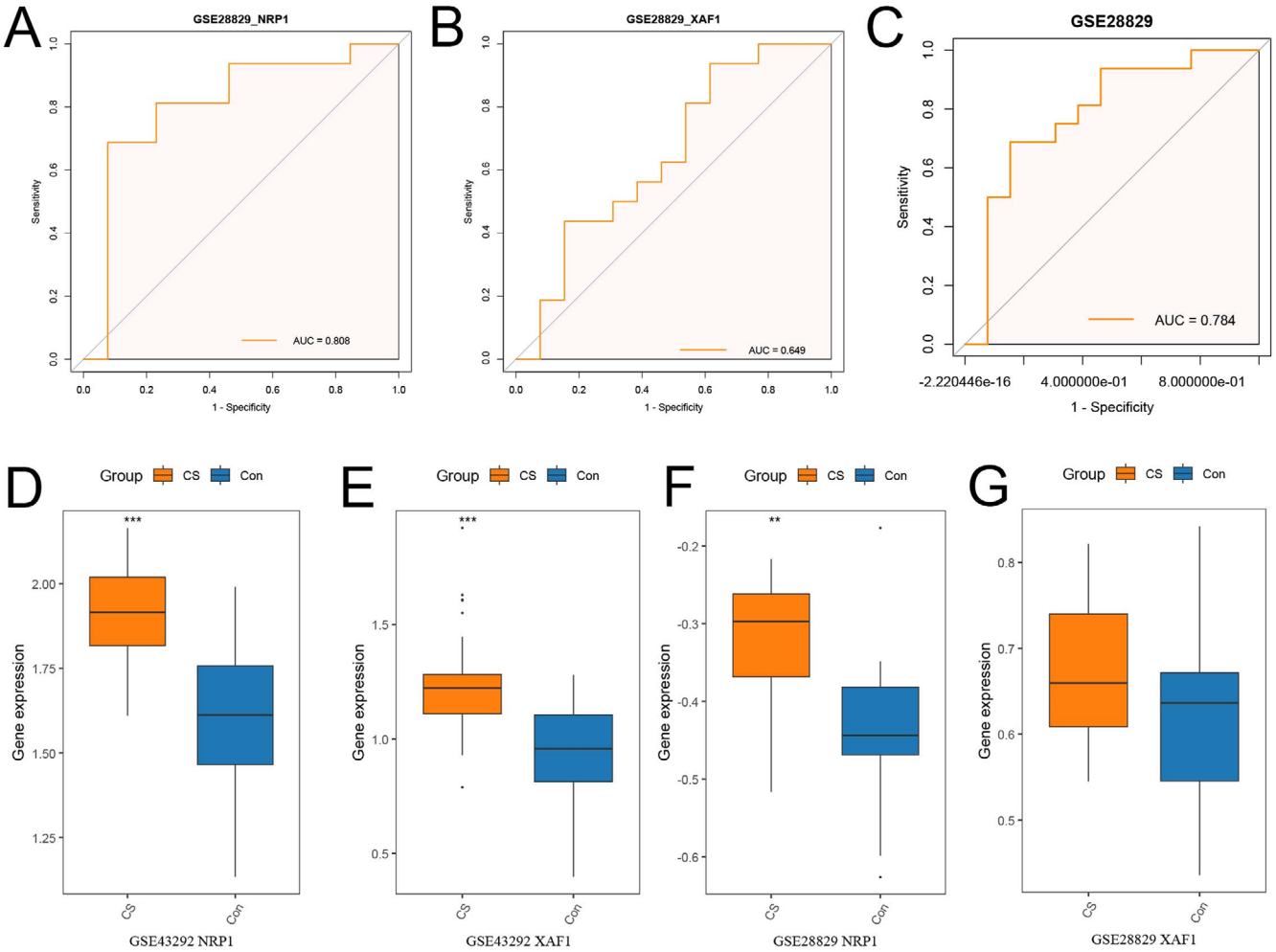

FIGURE 5 | External validation in GSE28829. (A–E) ROC curves for NRP1, XAF1, and combined genes in GSE28829. (D, E) Gene expression validation in GSE43292. (F, G) Expression validation in GSE28829.

in endothelial cells was significantly higher in the CAS group compared with controls (Figure 6B,C). Bulk RNA-seq data from carotid plaques further confirmed that both genes were upregulated in CAS tissues (Figure 6D). Consistent results were observed in the proteomic dataset, where NRP1 and XAF1 protein levels were increased in CAS plaques (Figure 6E,F). These findings demonstrate that NRP1 and XAF1 are consistently elevated in CAS across multi-omics levels.

3.8 | Immune Infiltration Analysis

To explore the immunological context of carotid artery stenosis, we performed ssGSEA to quantify immune cell infiltration. As shown in Figure S3A, the CAS group exhibited significantly higher infiltration levels of multiple immune cell types, including activated dendritic cells, macrophages, and various T cell subsets. Correlation analysis further revealed that NRP1 expression was positively associated with immunosuppressive cells such as MDSCs and neutrophils (Figure S3B), while XAF1 showed moderate correlations with regulatory and memory T cells (Figure S3C). These findings suggest that both genes may be associated with plaque progression via immune-related pathways, though causal mechanisms remain to be established.

4 | Discussion

Carotid artery stenosis (CAS) remains a leading cause of ischemic stroke, particularly in elderly populations [1]. However, the molecular mechanisms underlying plaque initiation and progression remain incompletely understood [23]. Traditional diagnostic approaches primarily focus on the degree of luminal narrowing and often fail to capture the cellular and molecular characteristics that contribute to disease progression. To address this challenge, we employed a systematic multi-layer analytical framework integrating single-cell transcriptomics, network-based gene prioritization, and machine learning. This study identified two endothelial cell-derived biomarkers, NRP1 and XAF1, with high diagnostic performance, which may facilitate mechanistic exploration of CAS and contribute to early screening strategies for CAS patients.

One of the most critical contributions of our study is the identification of endothelial cells as a key cellular population altered in CAS. Using scRNA-seq data from carotid tissues, we observed a significantly increased proportion of endothelial cells in the CAS group compared to controls. This finding aligns with emerging evidence suggesting that endothelial cell dysfunction plays a central role in atherogenesis and plaque instability

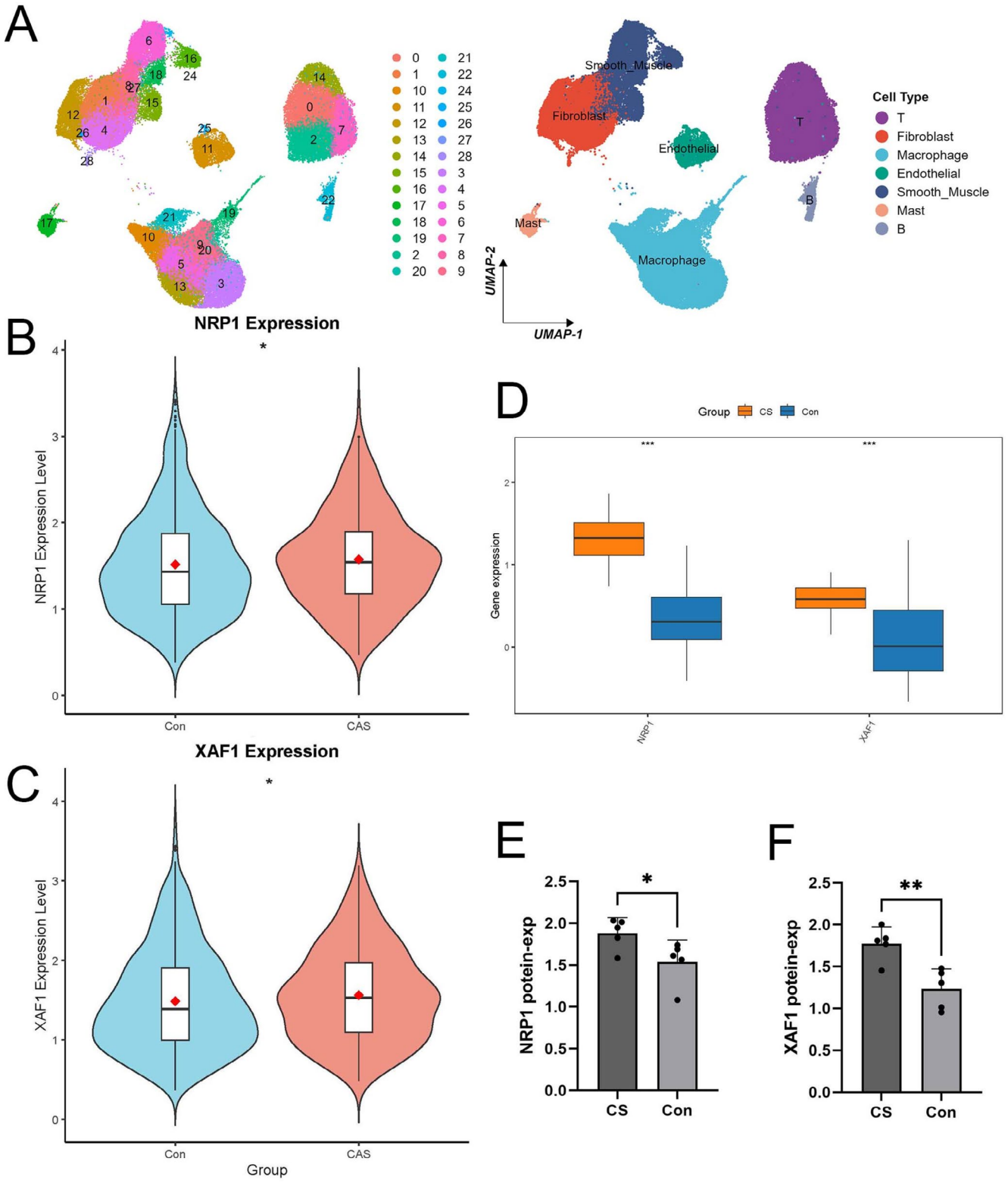

FIGURE 6 | Validation of NRP1 and XAF1 expression in the ZZ cohorts. UMAP visualization of the in-house single-cell dataset shows major cell types in carotid plaques (A); violin plots illustrate that NRP1 (B) and XAF1 (C) are significantly upregulated in endothelial cells in CAS compared with controls; bulk RNA-seq analysis confirms higher mRNA expression of NRP1 and XAF1 in CAS plaques (D); proteomic validation further demonstrates increased protein abundance of both genes in CAS tissues (E, F); $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

[24, 25]. Prior studies have shown that disturbed flow, inflammation, and oxidative stress can lead to phenotypic switching and dysfunction of endothelial cells, thereby promoting monocyte adhesion, leukocyte recruitment, and vascular remodeling [26, 27]. The enrichment of endothelial cells in the CAS group observed here not only reinforces their disease-associated relevance but also provides a rationale for focusing subsequent analyses on this population.

To extract mechanistically relevant genes from this cellular context, we performed differential gene expression analysis and WGCNA on bulk transcriptomic data. WGCNA enabled the identification of gene modules (brown and black) strongly associated with CAS, providing a network-based framework to prioritize genes beyond univariate statistics. The intersection of WGCNA-derived genes, DEGs, and endothelial cell-specific markers yielded a focused set of 80 consensus genes. This gene set likely represents transcriptional signals specific to endothelial dysfunction in the context of carotid plaque formation. Importantly, these consensus genes were further refined using machine learning algorithms [28, 29]. The application of random forest and LASSO regression allowed for both dimensionality reduction and feature importance ranking, resulting in the selection of NRP1 and XAF1 as the biomarkers.

NRP1 [30, 31] has been extensively characterized in vascular biology and is known to mediate VEGF signaling, endothelial proliferation, and angiogenesis. Recent studies also implicate NRP1 in immune regulation and macrophage recruitment within plaques, further highlighting its dual role in endothelial and immune cell crosstalk [32]. Our finding of elevated NRP1 expression in CAS samples, validated across multiple datasets, supports its potential as a diagnostic marker; whether it may serve as a therapeutic target requires further functional validation. In contrast, XAF1 (XIAP-associated factor 1) is a lesser-known regulator traditionally associated with pro-apoptotic signaling in cancer biology [33]. Its role in vascular pathology remains understudied. However, our data suggest that XAF1 expression may be associated with CAS, potentially through its involvement in endothelial cell apoptosis or immune cell engagement, though this requires further functional validation. The distinct but complementary functions of NRP1 and XAF1 provide a compelling rationale for their combined use as diagnostic biomarkers.

Another strength of our study lies in the robust validation of these biomarkers. Using independent external data (GSE28829), we confirmed the elevated expression of NRP1 and the modest but consistent upregulation of XAF1. The diagnostic models constructed from these genes achieved high AUCs in both the discovery and validation cohorts, emphasizing their reproducibility and potential clinical utility. In addition, nomogram construction and calibration curve analysis demonstrated excellent agreement between predicted and observed risk, suggesting that such models could be applicable in individualized risk stratification for CAS patients. Additionally, our study explored the immunological context of NRP1 and XAF1 expression. Immune infiltration analysis revealed that NRP1 correlated positively with immunosuppressive cell types such as MDSCs and neutrophils, while XAF1 showed associations with Treg and memory B cells. These findings suggest a possible association between these genes and immune modulation in the atherosclerotic microenvironment, though the underlying mechanisms warrant further investigation [34, 35]. The integration of immune profiling into gene-based diagnostics represents an emerging frontier that could inform both risk prediction and therapeutic response monitoring.

Despite these strengths, our study has several limitations. First, although machine learning enhances the robustness of biomarker selection, these methods are inherently sensitive to data quality and parameter settings. Second, although we validated our findings in an external cohort, prospective clinical studies are warranted to confirm the diagnostic performance and clinical relevance of NRP1 and XAF1 in real-world settings. Third, future functional studies, including in vitro cell-based assays and in vivo animal models, are warranted to mechanistically validate the causal roles of NRP1 and XAF1 in CAS pathogenesis and to establish the biological plausibility of our correlation-based findings. Future large-scale prospective studies are warranted to further validate and expand upon these findings.

Given the non-invasive nature of transcriptomic profiling from peripheral blood or plaque tissue, NRP1 and XAF1 hold potential as molecular adjuncts to current imaging-based diagnostics such as carotid ultrasound or CT angiography. Integration of these biomarkers into a multimodal screening framework—combining molecular signatures with structural imaging parameters—may enhance early risk stratification in asymptomatic CAS patients. Future prospective clinical studies are encouraged to evaluate such combined diagnostic strategies.

5 | Conclusion

In conclusion, our research has proposed a comprehensive multi-omics framework for the identification and validation of CAS-related biomarkers. By integrating single-cell and large-scale transcriptomic data with co-expression network analysis and machine learning, we identified NRP1 and XAF1 as key endothelial cell-associated genes in carotid artery stenosis, demonstrating strong diagnostic performance and potential immunological relevance. These findings will help decode the underlying mechanisms of carotid artery plaque formation and contribute to the early diagnosis and precise treatment of patients in clinical practice.

Author Contributions

Jiayi Wu, Chunguang Guo, and Kai Tang designed this work. Jiayi Wu, Chunguang Guo, Linfeng Zhang, Gaopo Cai, Yi Liu, and Kai Tang edited and revised the manuscript. All authors approved this manuscript.

Acknowledgments

I would like to show my greatest appreciation to the staff who offered continuing support and constant encouragement.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (3332022019) and the CAMS Innovation Fund for Medical Sciences (CIFMS, 2025-I2M-C&T-C-006).

Ethics Statement

The studies involving human participants were reviewed and approved by The Ethics Committee of The First Affiliated Hospital of Zhengzhou University (Ethical Number: 2023-KY-1133-002). The studies protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Collaborators GBDS, "Global, Regional, and National Burden of Stroke and Its Risk Factors, 1990-2019: A Systematic Analysis for the Global Burden of Disease Study 2019," *Lancet Neurology* 20, no. 10 (2021): 795–820.

2. G. A. Donnan, M. Fisher, M. Macleod, and S. M. Davis, "Stroke," *Lancet* 371, no. 9624 (2008): 1612–1623.

3. M. de Weerd, J. P. Greving, B. Hedblad, et al., "Prevalence of Asymptomatic Carotid Artery Stenosis in the General Population: An Individual Participant Data Meta-Analysis," *Stroke* 41, no. 6 (2010): 1294–1297.

4. P. Song, Z. Fang, H. Wang, et al., "Global and Regional Prevalence, Burden, and Risk Factors for Carotid Atherosclerosis: A Systematic Review, Meta-Analysis, and Modelling Study," *Lancet Global Health* 8, no. 5 (2020): e721–e729.

5. R. Malik, G. Chauhan, M. Traylor, et al., "Multiancestry Genome-Wide Association Study of 520,000 Subjects Identifies 32 Loci Associated With Stroke and Stroke Subtypes," *Nature Genetics* 50, no. 4 (2018): 524–537.

6. A. E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, "Single-Cell RNA-Seq: Advances and Future Challenges," *Nucleic Acids Research* 42, no. 14 (2014): 8845–8860.

7. F. Tang, C. Barbacioru, Y. Wang, et al., "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell," *Nature Methods* 6, no. 5 (2009): 377–382.

8. M. A. C. Depuydt, K. H. M. Prange, L. Slenders, et al., "Microanatomy of the Human Atherosclerotic Plaque by Single-Cell Transcriptomics," *Circulation Research* 127, no. 11 (2020): 1437–1455.

9. G. Zhao, H. Lu, Z. Chang, et al., "Single-Cell RNA Sequencing Reveals the Cellular Heterogeneity of Aneurysmal Infrarenal Abdominal Aorta," *Cardiovascular Research* 117, no. 5 (2021): 1402–1416.

10. P. Kong, Z. Y. Cui, X. F. Huang, D. D. Zhang, R. J. Guo, and M. Han, "Inflammation and Atherosclerosis: Signaling Pathways and Therapeutic Intervention," *Signal Transduction and Targeted Therapy* 7, no. 1 (2022): 131.

11. B. H. Northoff, A. Herbst, C. Wenk, et al., "Circular RNAs Increase During Vascular Cell Differentiation and Are Biomarkers for Vascular Disease," *Cardiovascular Research* 121, no. 3 (2025): 405–423.

12. Y. Xue, Y. Peng, L. Jin, et al., "Macrophage KDM2A Promotes Atherosclerosis via Regulating FYN and Inducing Inflammatory Response," *International Journal of Biological Sciences* 21, no. 6 (2025): 2780–2805.

13. H. Ayari and G. Bricca, "Identification of Two Genes Potentially Associated in Iron-Heme Homeostasis in Human Carotid Plaque Using Microarray Analysis," *Journal of Biosciences* 38, no. 2 (2013): 311–315.

14. Y. Doring, H. D. Manthey, M. Drechsler, et al., "Auto-Antigenic Protein-DNA Complexes Stimulate Plasmacytoid Dendritic Cells to Promote Atherosclerosis," *Circulation* 125, no. 13 (2012): 1673–1683.

15. T. Alsaigh, D. Evans, D. Frankel, and A. Torkamani, "Decoding the Transcriptome of Calcified Atherosclerotic Plaque at Single-Cell Resolution," *Communications Biology* 5, no. 1 (2022): 1084.

16. Y. Hao, S. Hao, E. Andersen-Nissen, et al., "Integrated Analysis of Multimodal Single-Cell Data," *Cell* 184, no. 13 (2021): 3573–3587.

17. T. X. A. Pham, A. Panda, H. Kagawa, et al., "Modeling Human Extraembryonic Mesoderm Cells Using Naive Pluripotent Stem Cells," *Cell Stem Cell* 29, no. 9 (2022): 1346–1365.

18. Y. Fan, Y. Zheng, Y. Zhang, et al., "ARID5A Orchestrates Cardiac Aging and Inflammation Through MAVS mRNA Stabilization," *Nature Cardiovascular Research* 4, no. 5 (2025): 602–623.

19. P. Langfelder and S. Horvath, "WGCNA: An R Package for Weighted Correlation Network Analysis," *BMC Bioinformatics* 9 (2008): 559.

20. P. Danaher, P. Wang, and D. M. Witten, "The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes," *Journal of the Royal Statistical Society Series B* 76, no. 2 (2014): 373–397.

21. D. A. Barbie, P. Tamayo, J. S. Boehm, et al., "Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1," *Nature* 462, no. 7269 (2009): 108–112.

22. G. Bindea, B. Mlecnik, M. Tosolini, et al., "Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer," *Immunity* 39, no. 4 (2013): 782–795.

23. M. Brisset, P. Boutouyrie, F. Pico, et al., "Large-Vessel Correlates of Cerebral Small-Vessel Disease," *Neurology* 80, no. 7 (2013): 662–669.

24. D. Urban, J. Poss, M. Bohm, and U. Laufs, "Targeting the Proprotein Convertase Subtilisin/Kexin Type 9 for the Treatment of Dyslipidemia and Atherosclerosis," *Journal of the American College of Cardiology* 62, no. 16 (2013): 1401–1408.

25. L. Tang, J. N. Cheng, Y. Long, et al., "PCB 118-Induced Endothelial Cell Apoptosis Is Partially Mediated by Excessive ROS Production," *Toxicology Mechanisms and Methods* 27, no. 5 (2017): 394–399.

26. L. Zhao, E. Lee, A. M. Zukas, et al., "CD44 Expressed on Both Bone Marrow-Derived and Non-Bone Marrow-Derived Cells Promotes Atherogenesis in ApoE-Deficient Mice," *Arteriosclerosis, Thrombosis, and Vascular Biology* 28, no. 7 (2008): 1283–1289.

27. C. P. Marin Oyarzun and P. G. Heller, "Platelets as Mediators of Thromboinflammation in Chronic Myeloproliferative Neoplasms," *Frontiers in Immunology* 10 (2019): 1373.

28. C. He, Z. Li, J. Wang, Y. Huang, Y. Yin, and Z. Li, "Atherosclerotic Plaque Tissue Characterization: An OCT-Based Machine Learning Algorithm With Ex Vivo Validation," *Frontiers in Bioengineering and Biotechnology* 8 (2020): 749.

29. S. Latha, P. Muthu, K. W. Lai, A. Khalil, and S. Dhanalakshmi, "Performance Analysis of Machine Learning and Deep Learning Architectures on Early Stroke Detection Using Carotid Artery Ultrasound Images," *Frontiers in Aging Neuroscience* 13 (2021): 828214.

30. Z. He and M. Tessier-Lavigne, "Neuropilin Is a Receptor for the Axonal Chemorepellent Semaphorin III," *Cell* 90, no. 4 (1997): 739–751.

31. A. L. Kolodkin, D. V. Levengood, E. G. Rowe, Y. T. Tai, R. J. Giger, and D. D. Ginty, "Neuropilin Is a Semaphorin III Receptor," *Cell* 90, no. 4 (1997): 753–762.

32. S. Sarabipour, K. Kinghorn, K. M. Quigley, et al., "Trafficking Dynamics of VEGFR1, VEGFR2, and NRP1 in Human Endothelial Cells," *PLoS Computational Biology* 20, no. 2 (2024): e1011798.

33. J. Wang, Y. Peng, Y. W. Sun, et al., "All-Trans Retinoic Acid Induces XAF1 Expression Through an Interferon Regulatory Factor-1 Element in Colon Cancer," *Gastroenterology* 130, no. 3 (2006): 747–758.

34. A. G. Talepoor, H. Fouladseresht, S. Khosropanah, and M. Doroudchi, "Immune-Inflammation in Atherosclerosis: A New Twist in an Old Tale," *Endocrine, Metabolic & Immune Disorders Drug Targets* 20, no. 4 (2020): 525–545.

35. S. Eshghjoo, A. Jayaraman, Y. Sun, and R. C. Alaniz, "Microbiota-Mediated Immune Regulation in Atherosclerosis," *Molecules* 26, no. 1 (2021): 179.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Quality control of scRNA-seq data. (A, C) Before filtering: distribution of nFeature_RNA, nCount_RNA, and percent.mt. (B, D) After filtering: improved data quality with controlled mitochondrial content. **Figure S2:** Functional enrichment analysis of endothelial marker genes. (A) GO terms enriched among endothelial cell markers. (B) KEGG pathways enriched among endothelial cell markers. **Figure S3:** Immune cell infiltration and gene correlation. (A) ssGSEA analysis of 28 immune cell types between CAS and control groups. (B, C) Correlation plots between NRP1/XAF1 and immune infiltration levels. **Table S1:** Basic information of patients in the ZZ_Cohort. **Table S1:** Basic information of patients for single-cell sequencing samples. **Table S1:** Basic information of patients for protein sequencing samples.