scientific reports

Check for updates

OPEN

Integrative biomarker discovery and immune profiling for ulcerative colitis: a multi-methodological approach

Lai Jiang[1,2,4,5], Shengke Zhang[2,5], Chenglu Jiang[2,5], Haiqing Chen[2], Jinbang Huang[2], Jinyan Yang[2], Hao Chi[2]✉, Qibiao Wu[1,4]✉ & Guanhu Yang[1,3]✉

Background We aimed to pinpoint biomarkers, create a diagnostic model for ulcerative colitis (UC), and delve into its immune features to better understand this autoimmune condition. **Methods** The sequencing data for both the UC and the control groups were obtained from GEO, including both bulk and single-cell data. Using GSE87466 as training group, we applied differential analysis, WGCNA, PPI, LASSO, RF, and SVM-RFE for biomarker selection. A neural network shaped our diagnostic model, corroborated by GSE92415 as the validation cohort with ROC assessment. Immune cell profiling was conducted using CIBERSORT. **Results** 53 disease-associated genes were screened. Enrichment analysis highlighted roles in complement cascades and cell adhesion. Eight biomarkers were finally identified through multiple machine learning and PPI: B4GALNT2, PDZK1IP1, FAM195A, REG4, MTMR11, FLJ35024, CD55, and CD44. The diagnostic model had AUCs of 0.984 (training group) and 0.957 (validation group). UC tissues revealed heightened plasma cells, CD8 T cells, and other immune cells. Two unique UC immune patterns emerged, with certain T and NK cells central to immune modulation. **Conclusion** We identified eight biomarkers of UC by various methods, constructed a diagnostic model through neural networks, and explored the immune complexity of the disease, which contributes to the diagnosis and treatment of UC.

Keywords Ulcerative colitis, Machine learning, Single cell sequencing, Diagnostic models, Immune infiltration, Therapeutic targets

Keywords ulcerative colitis, machine learning, neural network, diagnostic models, immune infiltration.

Ulcerative colitis is a persistent inflammatory disorder affecting the gastrointestinal tract, wherein there is a consistent and extensive erosion of the intestinal mucosa, accompanied by persistent inflammation and the development of ulcerative lesions[1]. The global prevalence of UC has shown a concerning upward trend, with a notable increase observed in developing nations[2]. Typical symptoms of UC include diarrhea, rectal bleeding, abdominal pain, and the urgency to defecate. In severe cases, systemic symptoms such as anemia, weight loss, and loss of appetite may occur[3,4]. Due to the non-specific initial symptoms of UC, it can often be misdiagnosed as other gastrointestinal disorders. Moreover, the lack of standardized and precise clinical diagnostic criteria for UC often results in delayed diagnosis until the disease has advanced to intermediate or advanced stages[5,6]. Therefore, gaining a comprehensive understanding of the pathogenesis of UC and developing accurate diagnostic models are crucial for its early detection, which can prevent the progression of the disease and reduce the unnecessary utilization of medical resources[7]. UC is a multifaceted and intricate disorder characterized by a multifactorial etiology, and its underlying pathogenesis remains elusive[8,9].

[1]Faculty of Chinese Medicine, State Key Laboratory of Quality Research in Chinese Medicine, and University Hospital, Macau University of Science and Technology, Macau, Macao SAR, China. [2]Clinical Medical College, Southwest Medical University, Luzhou 646000, China. [3]Department of Specialty Medicine, Ohio University, Athens, OH 45701, USA. [4]Chinese Medicine Guangdong Laboratory (Hengqin Laboratory), Guangdong-Macao In-Depth Cooperation Zone in Hengqin, Zhuhai 519000, China. [5]Lai Jiang, Shengke Zhang and Chenglu Jiang contributed equally to this work. ✉email: chihao7511@163.com; qbwu@must.edu.mo; guanhuyang@gmail.com

Current treatment strategies for UC primarily involve the use of aminosalicylates and corticosteroids to alleviate inflammation. Additionally, immunosuppressants (such as azathioprine and methotrexate), biologics (such as anti-TNFα antibodies), and small molecule drugs like JAK inhibitors are widely utilized in clinical practice[1]. JAK inhibitors, for example tofacitinib, target specific intracellular signaling pathways and offer a targeted treatment approach that may help control symptoms in patients who are unresponsive to conventional therapies[8,9]. Despite the variety of treatment options available, managing ulcerative colitis remains challenging, including issues such as drug efficacy diminishing over time and varied patient responses to treatments, and these methods do not cure UC. Therefore, further research into the pathogenesis of the disease and identification of disease marker genes, particularly the role of immune cells in the progression of the disease, is crucial for developing more effective treatment strategies and potential cures. Such research will help reveal new therapeutic targets and promote the development of personalized treatment strategies, thereby better meeting the needs of diverse patients[10].

Recently, the application of transcriptome sequencing studies has made significant advances in elucidating the complex molecular mechanisms of various diseases and in identifying potential diagnostic biomarkers[11]. Weighted correlation network analysis (WGCNA) has shown promising applications in the study of gastrointestinal diseases, providing powerful support for understanding their molecular mechanisms[12,13]. In our present study, we performed an extensive analysis to elucidate the molecular mechanisms underlying UC pathogenesis and the functional contributions of immune cells, employing an integrated approach that combines UC microarray datasets and scRNA-seq. Our main objective was to advance our comprehension of UC pathogenesis, with the ultimate goal of establishing a theoretical foundation for early diagnosis and subsequent therapeutic interventions.

Materials and methods

Data preparation

Two microarray datasets (GSE87466, GSE92415) and one single-cell transcriptome sequencing dataset (GSE182270) for UC were obtained from the Gene Expression Omnibus (GEO) database. GSE87466 was used as the training dataset for the study, comprising 87 UC samples and 21 normal samples. GSE92415 served as the testing dataset; from this dataset, we selected 53 UC samples and 21 normal samples. These 53 samples were chosen from a total of 162 UC patients who had not received any drug treatment. The annotation for both microarray datasets was conducted using the Affymetrix HT HG-U133 + PM Array Plate (GPL13158). We normalized the expression data for subsequent analysis. We obtained 3 UC samples and 3 normal control samples from GSE182270. The single-cell transcript data was merged, and batch effects were removed, followed by quality control and normalization using the "Seurat" R package[14].

Identification of disease-associated genes and functional enrichment analysis

We conducted differential gene expression analysis on the training dataset (GSE87466) using the 'limma' R package, setting the differential expression threshold at an absolute $\log FC > 1$ and $P.Val < 0.05$, thereby identifying genes that are differentially expressed between UC samples and normal control samples[15]. Using the 'WGCNA' R package, we constructed a gene co-expression network based on gene expression data to identify biologically relevant gene modules and to screen for disease-related genes[16,17]. Differential analysis primarily aims to identify genes whose expression levels significantly differ under various conditions, while WGCNA builds a co-expression network of gene expression data and identifies modules within the network, linking them to specific phenotypes or disease characteristics. By combining these two methods to identify disease-related genes and obtaining intersecting genes from both approaches, we leverage the advantages of integrative analysis. This strategy effectively reduces false positives, enhances the biological relevance of candidate genes, and provides a more reliable basis for subsequent research. We performed functional enrichment analysis of disease-associated genes, which included Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analyses. This analysis was carried out using the "clusterProfiler" R package[18].

Feature genes screening

In our study, we employed two distinct methodologies to identify disease biomarkers: Protein-Protein Interaction (PPI) network analysis and machine learning algorithms. PPI network analysis focuses on the biological connectivity and interactions among proteins, which can reveal key players and pathways involved in disease mechanisms. On the other hand, machine learning algorithms analyze complex data patterns to predict disease markers based on feature importance and classification accuracy[19]. By integrating these two approaches, we enhance the robustness and accuracy of our biomarker identification. The complementary nature of these methodologies allows for a more comprehensive analysis, mitigating the limitations inherent in using either approach in isolation. PPI network analysis is based on the interaction relationship between proteins to build the network, which can reflect the protein interactions within the cell and the connection between functional modules, and then the topology of the network can be analyzed to filter out the hub genes in the network. This is done by using the "STRINGdb", "igraph" package in R[20]. The Lasso algorithm is a linear regression algorithm based on regularization to filter feature genes in high-dimensional data. We used the Lasso model from the "glmnet" R package to filter the gene expression data, retaining genes with non-zero coefficients[21]. RF is an integrated learning algorithm that performs feature selection by constructing multiple decision trees. We use the RF model from the "randomForest" R package for feature selection. The SVM-RFE algorithm is a support vector machine based recursive feature elimination algorithm that performs feature selection by continuously removing the most detrimental features for classification. We used the SVM-RFE model from the "caret" R package for feature selection. We finally used the hub genes from the PPI analysis and the intersection genes from the three machine learning algorithms as our feature genes.

Diagnostic model building and evaluation

The Neural Network is a computational model that simulates the human nervous system. Neural networks learn complex mapping relationships between inputs and outputs by changing the weights of the connections between nodes[22]. We utilized the "neuralnet" and "NeuralNetTools" R packages to construct a neural network model using the feature genes as input features and trained the model on the training set (GSE87466)[23]. During training, the connection weights between neurons were adjusted based on the training data to achieve accurate classification of input data. We utilized the "pROC" R package to generate ROC curves for the training group (GSE87466) and test group (GSE92415) in the neural network model[24]. The area under the ROC curve was calculated to assess the performance of the model constructed using the feature genes. We drew up the nomogram, which translates complex diagnostic models into an easy to understand and apply graphical form, helping doctors and patients to better understand and apply the results of diagnostic models and thus make better decisions. To assess the predictive accuracy and clinical utility of the diagnostic model, calibration plots and decision curve analysis (DCA) curves were generated for the nomogram employing the 'rms' and 'rmda' R packages. These analytical tools provided valuable insights into the performance and applicability of the diagnostic model. To evaluate the classification performance of each feature gene, we plotted ROC curves using the "pROC" R package.

Immune infiltration characteristics in UC

The "CIBERSORT" R package was utilized to quantify the extent of immune cell infiltration within the samples (GSE87466), enabling us to compare and evaluate the significant disparities in immune cell infiltration profiles between the UC and normal control groups. Furthermore, we performed Spearman analysis to correlate eight feature genes with immune cell infiltration.

Identification of disease subtypes and immune infiltration

Consensus clustering is a widely employed computational method utilized to determine the optimal number of unsupervised clusters within a given dataset. We utilized the R package "ConsensusClusterPlus" to classify UC patients in GSE87466 and GSE292415 into distinct clusters[25]. In order to ascertain the most appropriate number of clusters, several analytical techniques were employed, including consensus matrix plots, consensus cumulative distribution function (CDF) plots, assessment of relative changes in the area under the CDF curve, and trace plots. Principal components analysis (PCA) was conducted on the disease typing results to visually depict the dissimilarities and similarities among the clusters. The "CIBERSORT" R package was used to calculate immune infiltration levels between disease subtypes and identify immune cell types with significant differences.

Single-cell data processing

We utilized the 'Seurat' R package for the cell filtering process, applying a set of exclusion criteria that included: (1) cells expressing fewer than 200 genes, (2) cells expressing more than 2500 genes, and (3) cells with mitochondrial gene content exceeding 5%. Dimensionality reduction of the gene expression matrix was performed using the PCA method, selecting the top 20 principal components. The reduced data were then clustered using Seurat's FindClusters function, resulting in 18 distinct cell clusters. Cluster results were visualized using the t-distributed stochastic neighbor embedding (tSNE) method. We annotated the cell types for further analysis using cell marker information available on the CellMarker website (https://biocc.hrbmu.edu.cn/CellMarker/). For subsequent analyses, histograms of cell proportions in the samples and feature plots were visualized using the ggplot2 package.

Analysis of cellular communication networks

We used the CellChat package for cell communication analysis.CellChat is designed to identify and quantify intercellular communication networks. It utilizes a database of known ligand-receptor interactions to infer communication signals between cell types. We screened the same number of normal group cells and tumor group cells to infer cell signaling networks separately and finally integrated them for comparison.

Statistical analysis

All statistical analyses were performed using R version 4.2.2, 64-bit, along with relevant auxiliary packages. Student's t-test was employed for comparing normally distributed variables between two groups, whereas the Wilcoxon test was utilized for non-normally distributed variables to assess significant differences between the groups. Spearman correlation analysis was used to examine the correlation between variables. For all statistical tests conducted, the significance level was predetermined at $P < 0.05$, indicating a threshold for determining statistical significance.

Results

Integration of differential analysis and WGCNA for gene selection

Figure 1 shows the flowchart of this study. The analysis was conducted using data from GSE87466, which comprised a total of 87 UC and 21 normal control tissues. All expression values were normalized (Fig. 2A,B), and disease-associated genes were screened using a combined approach of differential analysis and WGCNA. We employed the "limma" package to identify differentially expressed genes (DEGs) with an adjusted p-value (adj.P.Val) < 0.05 and log fold change (logFC) > 1 or < -1. Within the GSE87466 dataset, a comprehensive analysis revealed a collection of 91 DEGs, consisting of 55 upregulated genes and 36 downregulated genes. To provide a visual representation of the expression patterns of these DEGs across individual samples, a heatmap was generated (Fig. 2C). The DEGs were further visualized using volcano plot and divergent bar plot. (Figs. 2D,E). Following the identification of DEGs through the differential analysis, we proceeded to subject these DEGs to Gene Set Enrichment Analysis (GSEA). The pathways significantly enriched in the UC group

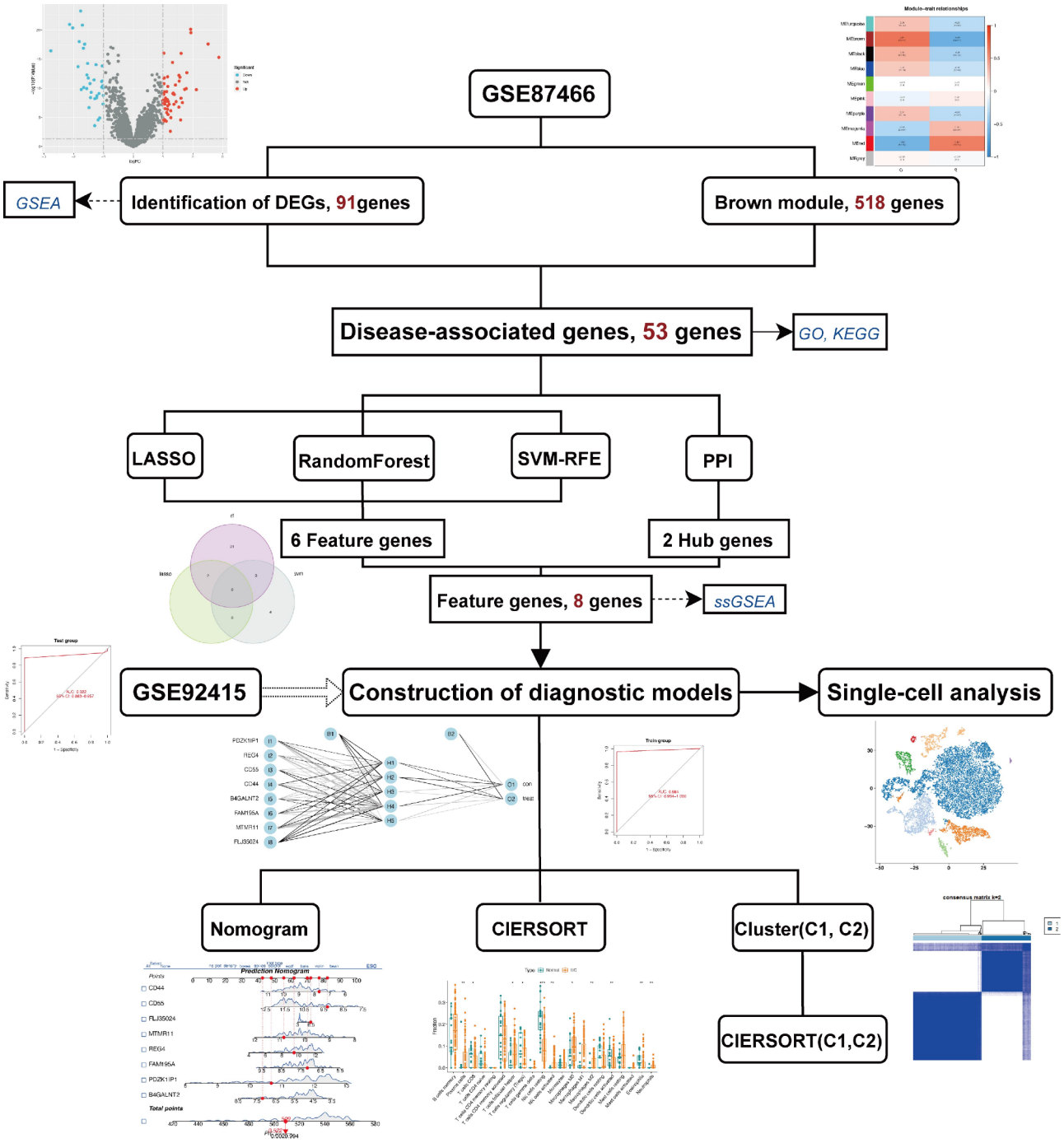

Fig. 1. Flow chart of this study.

were predominantly associated with amoebiasis, complement and coagulation cascades, leishmaniasis, and staphylococcus aureus infection, as revealed by the GSEA analysis (Fig. 2F,G). To establish a Weighted Gene Co-Expression Network for the GSE87466 dataset, we employed the "WGCNA" package, a powerful computational tool widely employed for gene co-expression analysis. As illustrated in Fig. 3A, we set the soft threshold to 6, ensuring an R^2 value greater than 0.9 and a high mean connectivity. After merging highly correlated modules, a total of 10 modules were identified for further investigation. The resulting modules were then displayed together under the clustering tree, showcasing the merged and primed modules. (Fig. 3B). The robustness of module delineation was validated through transcriptional correlation analysis within modules, which revealed no significant inter-module linkage. This suggests that the identified modules are distinct and not strongly correlated with each other (Fig. 3C). Correlations and significance were calculated between the modules and the disease and control groups, respectively. Among them, the brown module showed the highest correlation and was selected as the key module (Fig. 3D). In the MM (module membership) versus GS (gene significance)

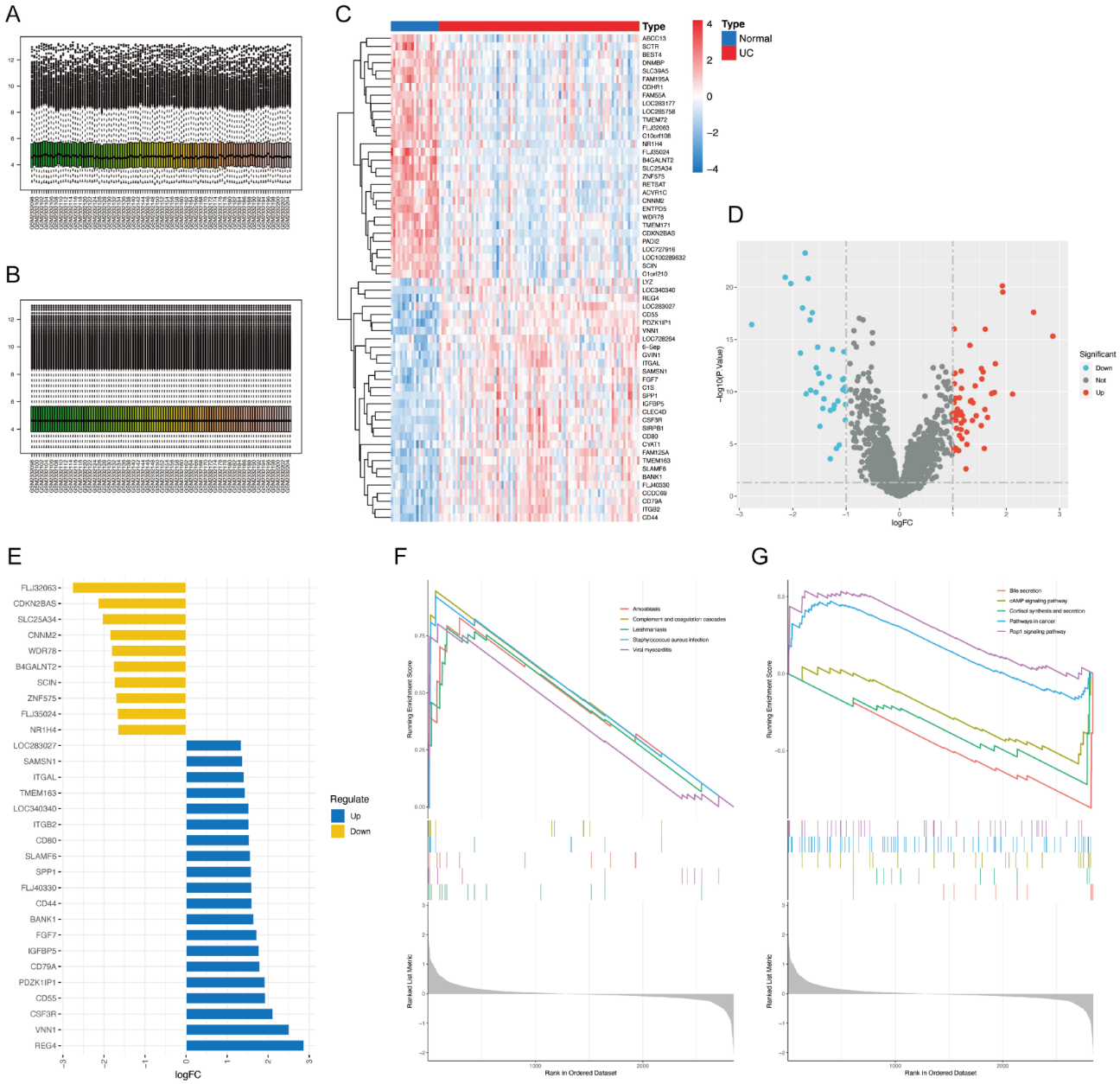

Fig. 2. Identification and enrichment analysis of differential genes between UC and normal control. (**A**) Raw data of the training set without batch correction. (**B**) Expression matrix of the training set after batch correction to remove batch effects. (**C**) Heat map of differential genes between UC and normal control. The blue module represents normal and the red module represents UC, and the red color in the heat map represents up-regulated gene expression and blue color represents down-regulated gene expression. (**D**) Volcano plot of differential genes. Expression is up-regulated when $|\log FC|>0$, $p<0.05$ and gene expression is down-regulated when $|\log FC|<0$, $p<0.05$. Differential genes were screened by $|\log FC|>1$, $p<0.05$, blue indicates down-regulation and red indicates up-regulation. (**E**) Dispersion bar graph of differential genes. Blue represents up-regulated genes, yellow represents down-regulated genes, and the length of the bar represents the value of logFC. (**F**) GSEA enrichment analysis of differential genes and selection of the five pathways with the strongest enrichment significance for display. (**G**) GSEA enrichment analysis of differential genes and selection of the five pathways with the weakest enrichment significance for display.

scatter plot of the brown module, a positive correlation was observed between MM and GS, indicating that these highly disease-related genes play a pivotal role in the key module (Fig. 3E). Clinically meaningful modules were identified, and further examination was conducted on all the genes in the brown module. Finally, we took the intersection of the genes identified by differential analysis and all the genes in the brown module of WGCNA to obtain disease-associated genes (Fig. 3F).

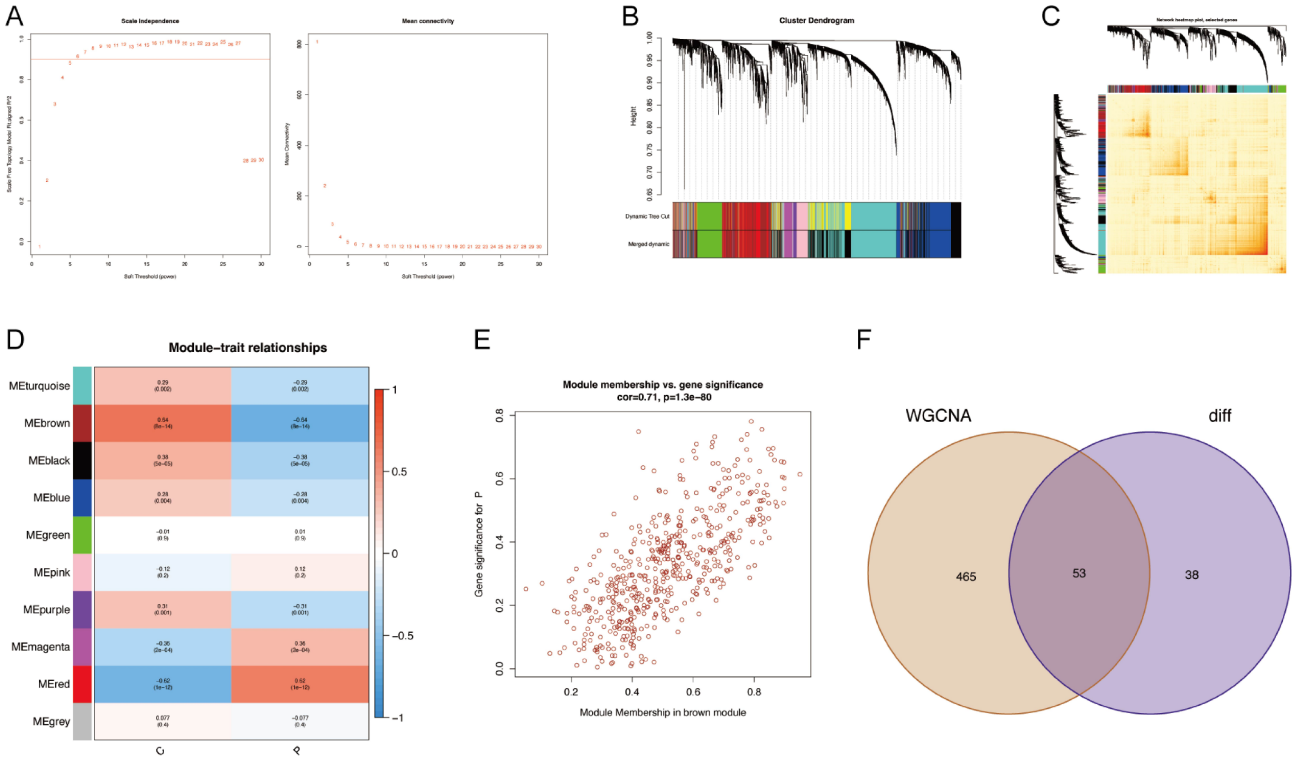

Fig. 3. Construction of gene co-expression network. (**A**) The left panel is set when the scale-free topological fit index $R^2 = 0.9$, and the best soft threshold $\beta = 6$ is chosen to obtain the best average connectivity of the co-expression network on the right panel. (**B**) Sample clustering tree with the modules before and after merging, Dynamic Tree Cut for the original modules and Merged dynamic for the result after merging the strongly associated modules. (**C**) Heat map of correlation between merged modules and modules. (**D**) Heat map of correlations between modules and clinical traits. Red indicates positive correlation, blue indicates negative correlation, and the darker the color, the stronger the correlation. The numbers in parentheses are the p-values of correlations between modules and traits to test whether they are statistically significant with each other. The numbers above the brackets indicate the magnitude of correlation between modules and traits. (**E**) Scatter plot between brown module affiliation and UC gene significance with a correlation between each other of cor = 0.71, $p < 1.3e{-}80$. (F) Wayne diagram of 91 intersecting genes obtained from DEGs and co-expressed genes with the strongest association with UC as disease associated genes.

Functional enrichment analysis

The GO enrichment analysis results revealed that disease-associated genes were significantly enriched in biological processes such as positive regulation of leukocyte cell-cell adhesion, positive regulation of cell-cell adhesion, regulation of cell-cell adhesion, adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains, regulation of leukocyte cell-cell adhesion, positive regulation of T cell activation, and secretory granule membrane, indicating their potential involvement in these pathways (Fig. 4A). The KEGG enrichment analysis results demonstrated that disease-associated genes were significantly enriched in pathways such as complement and coagulation cascades, viral myocarditis, ECM-receptor interaction, staphylococcus aureus infection, hematopoietic cell lineage, and Toll-like receptor signaling pathway (Fig. 4B). Additionally, we conducted enrichment analysis of disease-related genes using the Metascape website, which revealed that these genes were primarily involved in biological processes such as positive regulation of immune response, positive regulation of leukocyte cell-cell adhesion, regulation of peptidyl-tyrosine phosphorylation, neutrophil degranulation, complement system, and innate immune response (Fig. 4C). The relationship between enriched terms is illustrated in Fig. 4D. In summary, the functional enrichment analysis reveals that the disease-associated genes are predominantly involved in complement and coagulation cascades, regulation of adhesion between various cells, and immune response, as indicated by the enrichment results from the bioinformatics analysis.

Selection of feature genes

Three machine learning algorithms and PPI network analysis were utilized in combination to identify the optimal feature genes for the diagnosis of UC. The LASSO regression algorithm identified 11 genes, while the RF algorithm identified 32 genes, and the SVM-RFE algorithm identified 16 genes (Fig. 5A-D). We utilized Venn diagrams to identify six genes that showed overlap in the results obtained from the three machine learning algorithms mentioned above (Fig. 5E). Additionally, we performed PPI network analysis and identified two key genes in the constructed network (Fig. 5F, G). Lastly, we utilized the overlapping genes from the three machine

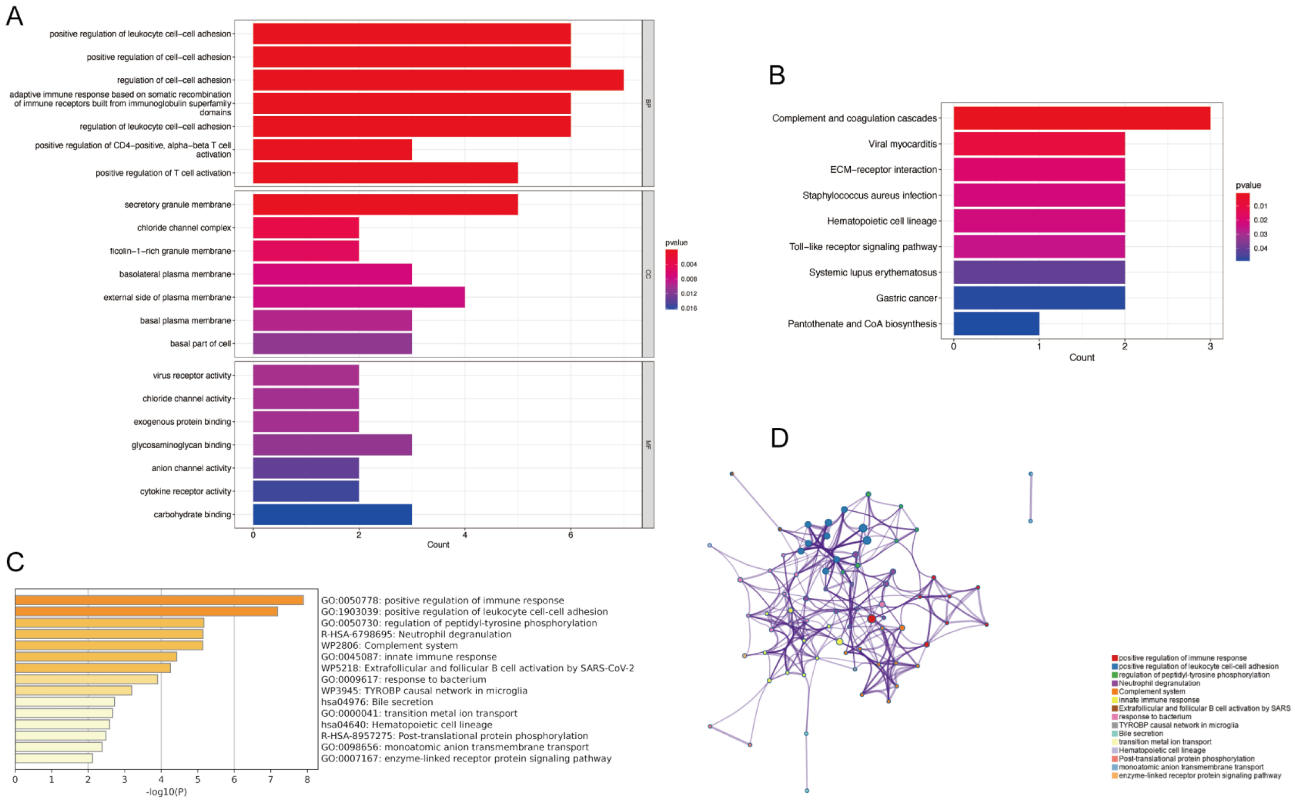

Fig. 4. Functional enrichment analysis. (**A**) Histogram of GO enrichment analysis of disease associated genes. (**B**) Histogram of KEGG enrichment analysis of disease associated genes. (**C**) Histogram of Matascape enrichment analysis results. (**D**) Node Network of Matascape enrichment analysis results.

learning algorithms, in combination with the hub genes identified from the PPI network, namely B4GALNT2, PDZK1IP1, FAM195A, REG4, MTMR11, FLJ35024, CD55, and CD44, as feature genes for subsequent analyses.

Validation of feature gene expression

We further investigated the expression of the selected feature genes in GSE87466 using a box plot, and validated the findings using an additional dataset, GSE92415 (Fig. 6A,B). Gene correlations were also examined, as shown in Fig. 5H.

Analysis of the feature genes using GSEA

To better understand the role of feature genes in UC, we analyzed them using single-gene GSEA enrichment analysis. The results of the single-gene GSEA enrichment analysis revealed that all the selected feature genes, except for REG4, were significantly enriched in the complement and coagulation cascades. REG4 was found to be enriched in the Adherens junction pathway (Fig. 7A–H).

Modeling and testing of diagnostic performance using neural network and generation of diagnostic column line graph

We employed a neural network model to effectively model and test the feature genes, leveraging its robust non-linear modeling capabilities to uncover intricate patterns and associations in gene expression data and evaluate the importance of our screened signature genes. (Fig. 8A). The ROC curve of the neural network model was generated using the training datasets (GSE87466) and test datasets (GSE92415), and the performance of the model was comprehensively evaluated by calculating the AUC to assess its overall predictive accuracy (Fig. 8B,C). Then we utilized the Rms package to construct UC diagnostic column line graph models for the feature genes, including B4GALNT2, PDZK1IP1, FAM195A, REG4, MTMR11, FLJ35024, CD55, and CD44 (Fig. 8D). The calibration curves demonstrated that the disparity between the actual and predicted risks of UC was very minimal, suggesting that the column line graph model for UC is highly accurate(Fig. 8E). The Decision Curve Analysis (DCA) was performed to evaluate the impact of utilizing the model for diagnosis at various prediction probability thresholds, providing further assessment of the performance and utility of the model (Fig. 8F). To further validate the diagnostic value of B4GALNT2, PDZK1IP1, FAM195A, REG4, MTMR11, FLJ35024, CD55, and CD44, we conducted receiver operating characteristic (ROC) analysis. B4GALNT2 (AUC:0.952), PDZK1IP1 (AUC:0.969), FAM195A (AUC:0.907), REG4 (AUC:0.929), MTMR11 (AUC:0.996), FLJ35024 (AUC:0.949), CD55 (AUC:0.958), CD44 (AUC:0.957) (Fig. 8G).

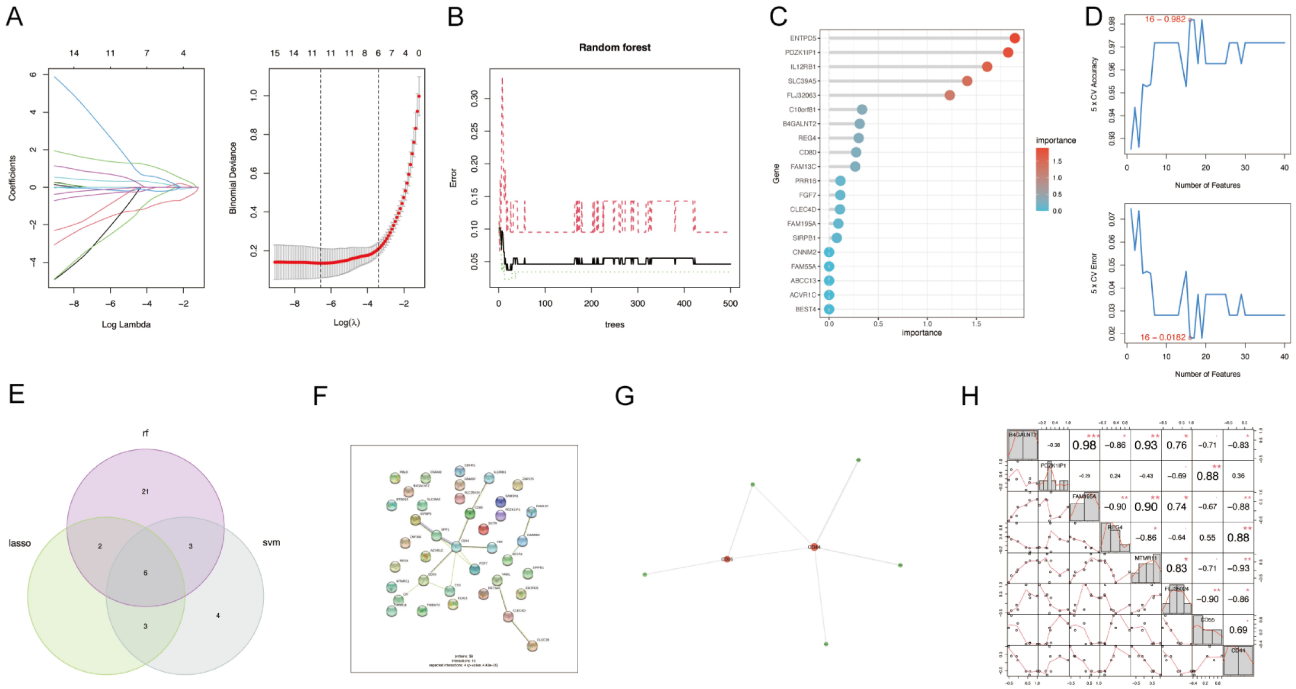

Fig. 5. Machine learning screening of feature genes. (**A**) LASSO algorithm for feature gene screening. (**B**) Combination of number of decision trees and error rate of RF algorithm. (**C**) RF calculates the top 20 genes for gene importance for ranking. (**D**) support vector machine recursive feature elimination (SVM-RFE) algorithm to screen biologic feature genes, and the point with the lowest accuracy and error rate is used as the number of feature genes screened by SVM-RFE. (**E**) Wayne diagram to obtain the feature genes screened by three machine learning algorithms. (**G**) PPI network of disease associated genes, the circles indicate protein nodes, and the lines between each other indicate the existence of interrelationship between two proteins. (**F**) Interacting networks of hub genes in PPI networks. (H) Correlation line graph between 8 disease signature genes.

Immunological infiltration in the UC group and healthy controls

Immune cells play an important role in the development and progression of UC. To gain insights into the relative proportion of different immune cell types that undergo changes in the disease and control groups, we utilized the CIBERSORT package for immune cell infiltration analysis. After excluding non-statistically significant immune cell infiltration, our results revealed that plasma cells, T cells CD8, T cells follicular helper, NK cells activated, macrophages M0, macrophages M2, and neutrophils were found to be higher in the disease group compared to the control group (Fig. 9A). Our analysis revealed that FAM195A was significantly and positively correlated with NK cells resting, T cells CD4 memory activated, macrophages M0, and T cells CD4 naive. On the other hand, FAM195A was significantly and negatively correlated with NK cells activated, plasma cells, and mast cells resting. MTMR11 was significantly positively correlated with NK cells resting and T cells CD4 naive, while it was negatively correlated with NK cells activated and plasma cells. CD44 showed a significant positive correlation with mast cells activated, NK cells activated, and plasma cells, while it was negatively correlated with NK cells resting and T cells CD4 naive. B4GALNT2 showed a significant positive correlation with T cells CD4 naive, and a significant negative correlation with NK cells activated and plasma cells (Fig. 9B). In conclusion, there are significant differences in immune infiltration between the disease group and the normal group in UC. Moreover, most of the characteristic genes, including FAM195A, MTMR11, CD44, and B4GALNT2, showed significant correlations with specific immune cell types.

Identification of subtypes of UC

We combined and removed batch effects from the UC group data in GSE87466 and GSE92415 to identify distinct subtypes of UC. UC samples were classified into subtypes using the consensus clustering method based on gene expression profiles. The optimal number of subtypes was determined as 2 using various criteria including consensus matrix plots, CDF plots, relative changes in regions under the CDF curve, and trace plots (Fig. 10A-C,E). Principal Component Analysis (PCA) revealed notable distinctions among the subtypes, which were labeled as C1 and C2 (Fig. 10F). The heatmap visualizes the differential gene expression across the two identified subtypes (Fig. 10D).

Different immunological characteristics of the two subtypes

As illustrated in Fig. 10G,H, our analysis of immune infiltration in UC revealed that levels of T cells CD8, T cells CD4 memory resting, T cells gamma delta, NK cells activated, monocytes, and macrophages M0 were higher in the C1 subtype compared to the C2 subtype, after excluding non-statistically significant immune cell infiltration.

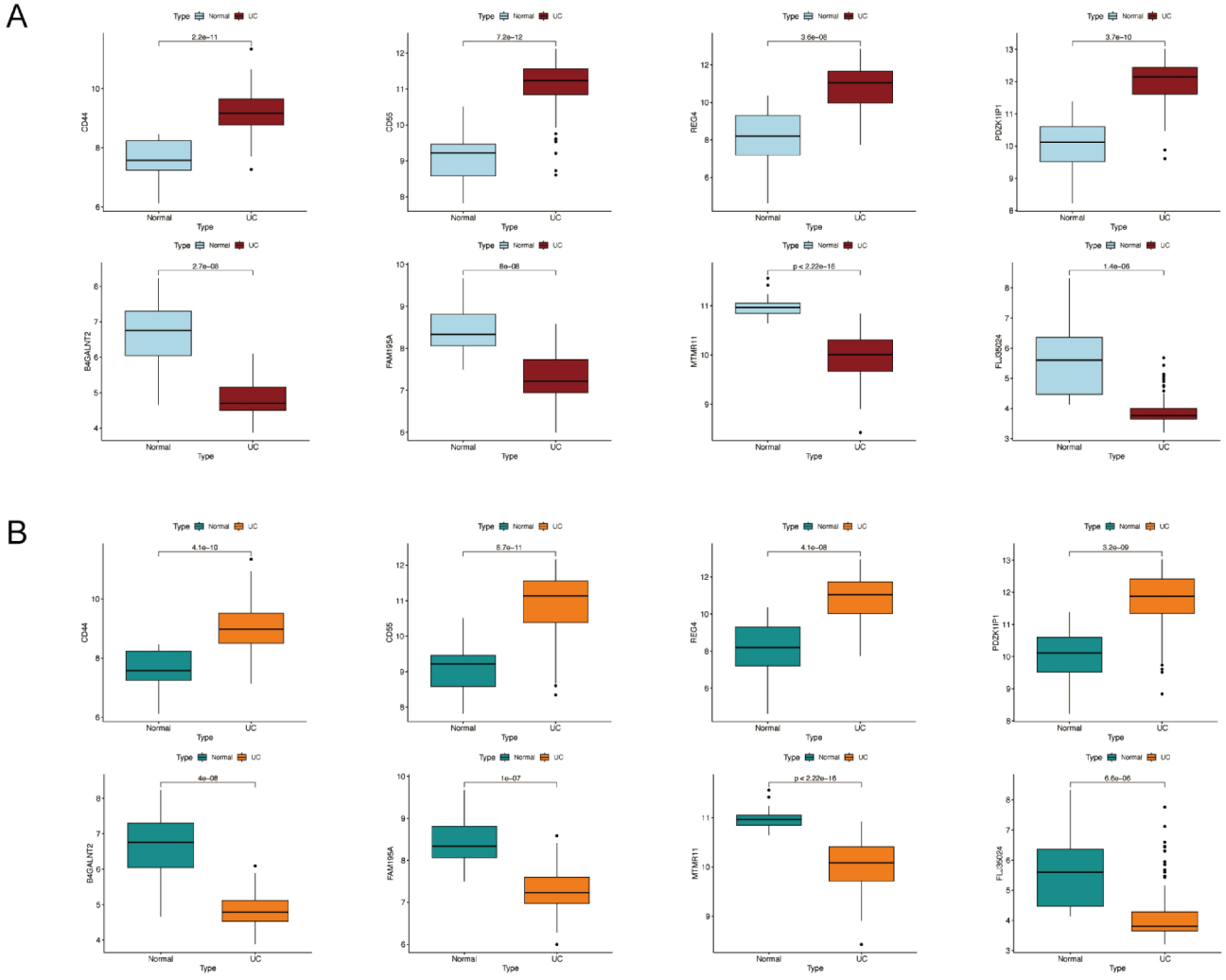

Fig. 6. Validation of feature genes expression. (**A**) Box line plot of differential expression of feature genes between UC and normal individuals in the training group GSE87466. The size range of the box line plot indicates the range of their gene expression, while the black line in the plot indicates the mean of their expression. The values between the two box-line plots are P values, and the differences are statistically significant when $P < 0.05$. (**B**) Box line plot of differential expression of characteristic genes between UC and normal individuals in the validation group GSE92415.

Compared to the C1 subtype, the C2 subtype exhibited higher levels of B cells memory, T cells follicular helper, NK cells resting, macrophages M2, dendritic cells activated, mast cells activated, and neutrophils.

Expression of diagnostic genes in single-cell data

From the GSE182270 dataset, we selected three UC samples (GSM5525955, GSM5525956, GSM5525957) and three healthy control samples (GSM5525960, GSM5525961, GSM5525962) for scRNA-seq analysis. Scatter plots of PCA-reduced data illustrated the distribution of cells from different samples, with the Elbow Plot indicating the importance of each principal component (Fig. 11A). We selected the top ten principal components for further dimension reduction and clustering, resulting in 18 cell clusters. These clusters were visualized using the t-SNE method (Fig. 11B). To accurately annotate these subclusters, we referred to marker genes reported in the literature and supplemented them from the CellMarker database. Based on these markers, we identified cell types. Histograms displayed the proportions of various cell types across the samples (Fig. 11C). Figure 11D shows the distribution and cell type information of cells from the normal group samples. We plotted the expression of diagnostic genes in normal group sample cells using feature plots; however, due to differences in sequencing technologies, the gene FLJ35024 was absent in the single cell sequencing data. Figure 11F shows the distribution and cell type information of cells from the UC group samples. Figure 11G displays the expression of diagnostic genes in UC group cells.

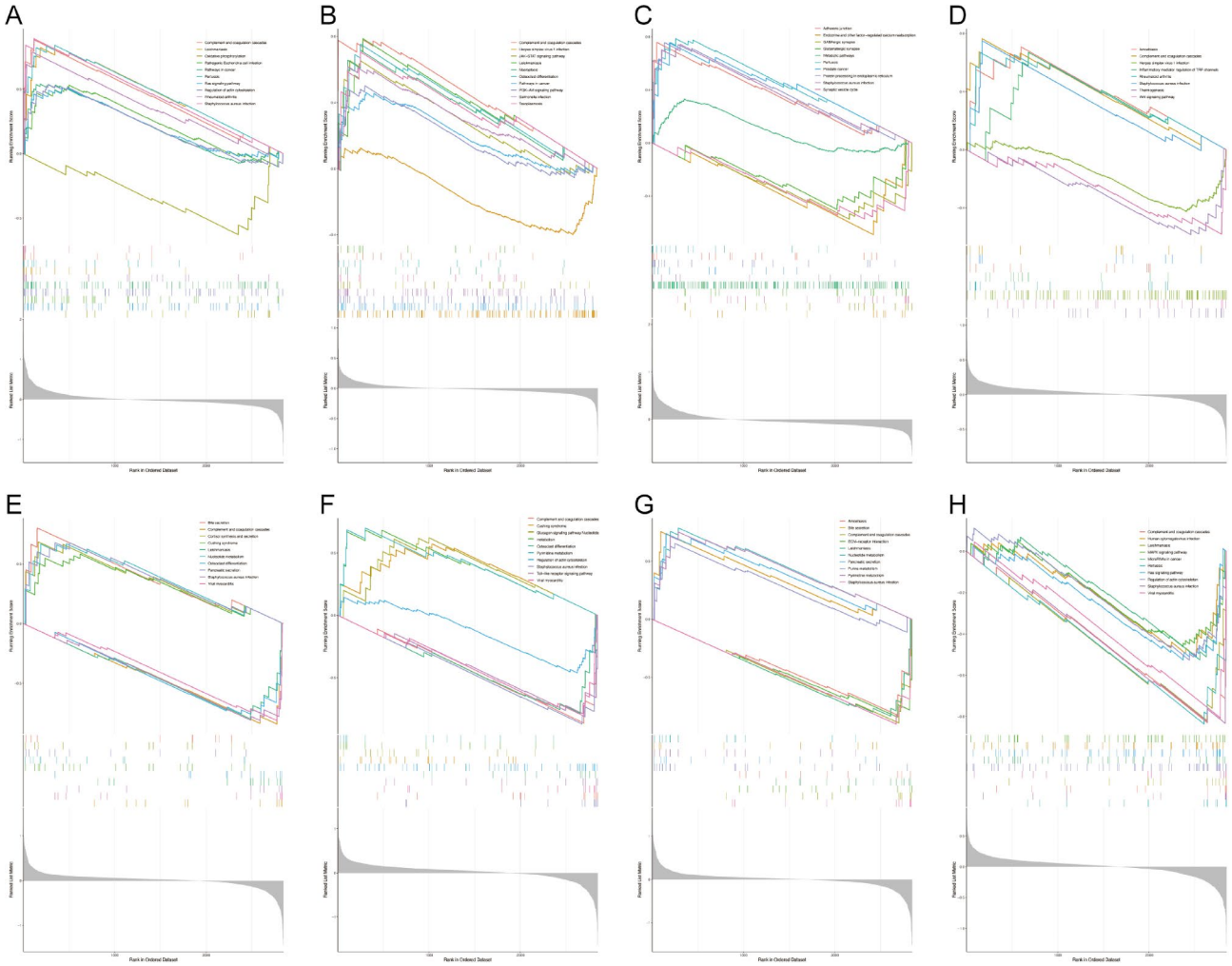

Fig. 7. Single gene GSEA analysis. (**A**–**H**) Single gene GSEA analysis of feature genes, respectively, and the top five pathways with the highest and lowest enrichment significance were selected for display.

Analysis of CD44 and CD55 in cellular communication

We found that two UC diagnostic genes, CD44 and CD55, have a more pronounced and widespread expression in the single-cell data. They differ greatly in their molecular functions and mechanisms of action, but they are both expressed on the cell surface and encode proteins that are cell surface proteins involved in the regulation of immune responses and protection of cells from damage. These commonalities reveal their importance in the study of cell-environment interactions, immune regulation, and disease development. We used the CellChat package to analyze the role of CD44 and CD55 in cellular communication, particularly in understanding cell-cell interactions, signaling mechanisms, and disease development. Compared with the normal group samples, interactions between cells were more frequent in the UC group, but the strength of communication was decreased (Fig. 12A). In UC, inflammation leads to a massive infiltration of immune cells and an increase in the number and density of cells, allowing for increased physical contact and interaction between cells. Despite the frequency of interactions, the effectiveness or strength of each interaction may be reduced due to impaired cellular function in the inflammatory state. In the normal and UC groups, the strength of communication between different cell types also changed, but fibroblasts consistently maintained an active signal output (Fig. 12B). Cells such as B cells, epithelial cells, and T cells underwent very pronounced changes. The cellular communication families with significant differences in intensity in the normal and UC groups were demonstrated by histograms (Fig. 12C). The violin plots demonstrated the expression of two genes, CD44 and CD55, in cells of the 9 middle cell types (Fig. 12D). In the cellular communication network, we extracted all ligand-receptor pairs containing CD44 and CD55.MIF is an important inflammatory factor with its key role in multiple immune responses and inflammatory reactions. Its primary receptors include CD74 and CD44. this ligand-receptor pair complex is active in cellular communication between epithelial cells and T cells, B cells, and myeloid cells (Fig. 12E). The interaction of LGALS9 with CD44 plays an important role in a variety of biological processes, particularly in immune regulation. In UC samples, the LGALS9-CD44 ligand receptor pair played an important role in cellular communication between myeloid and epithelial cells (Fig. 12F). Among them, the communication of myeloid cells with T cells through this ligand-receptor pair attracted our attention. It has been shown that LGALS9 can

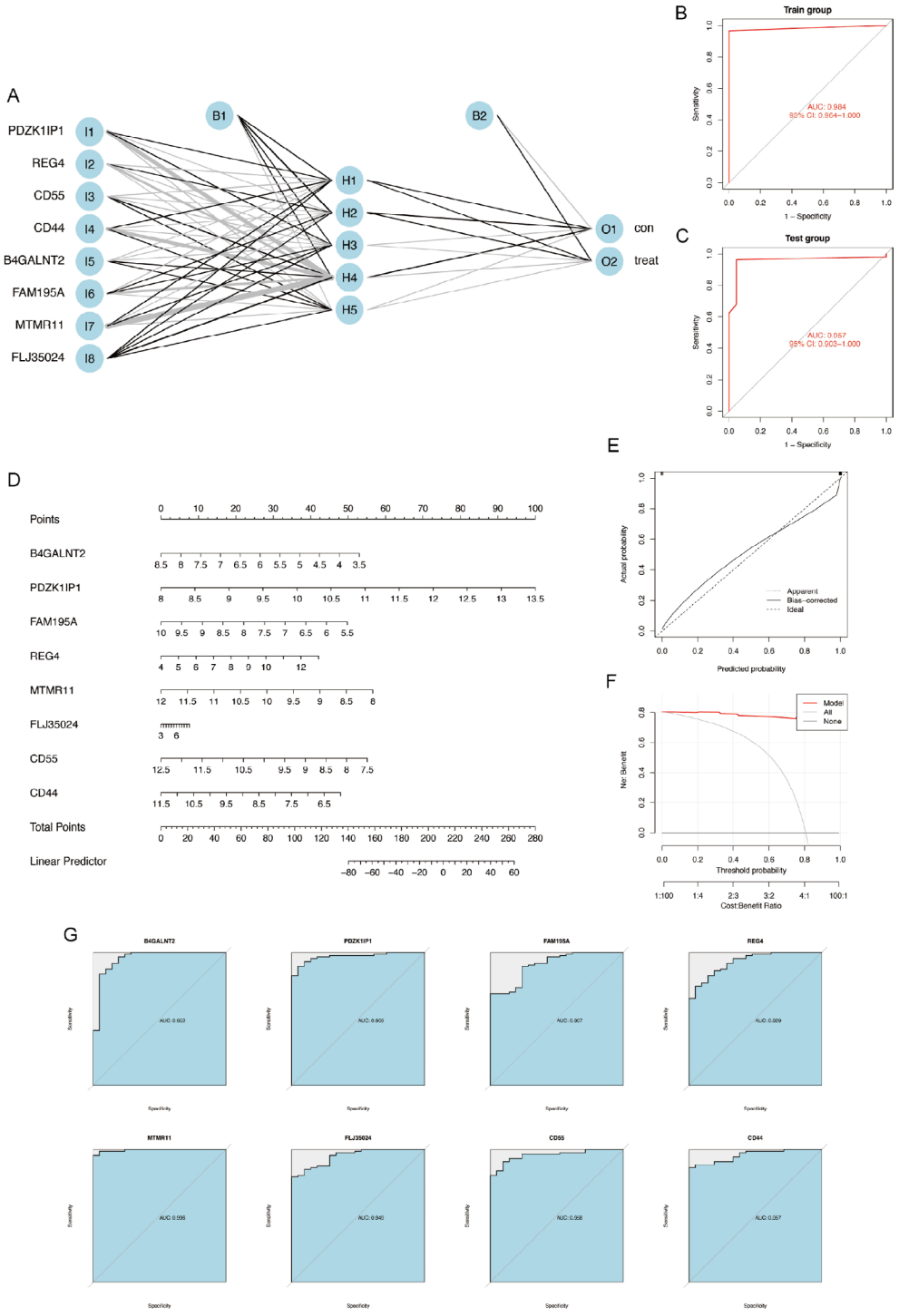

Fig. 8. Construction and validation of Artificial Neural Network (ANN) model and nomogram model. (**A**) ANN model constructed using feature genes, containing input layer, hidden layer and output layer. (**B**) Training set ROC curve with AUC = 0.984, 95% CI 0.964-1.000, used to illustrate whether the model has good prediction performance. (**C**) Validation set ROC curve, AUC = 0.957, 95% CI 0.903-1.000, used to demonstrate whether the stability and generalization of the model are good. (**D**) Nomogram plot of the characteristic gene construction, each element followed by a scoring scale. The scores of each element are summed to obtain a total score to predict the risk of disease. (**E**) Calibration curve for the evaluation of nomogram prediction performance. The higher the overlap between the solid and dashed lines and the closer the diagonal line, the better the performance. (**F**) Decision curve analysis (DCA), which compares the clinical benefit between the nomogram model and other diagnostic indicators. the higher the AUC, the higher the clinical benefit in the range of possible thresholds from 0 to 1. (**G**) The ROC curves of the eight feature genes in the training set were used to assess whether each gene had good disease predictive ability.

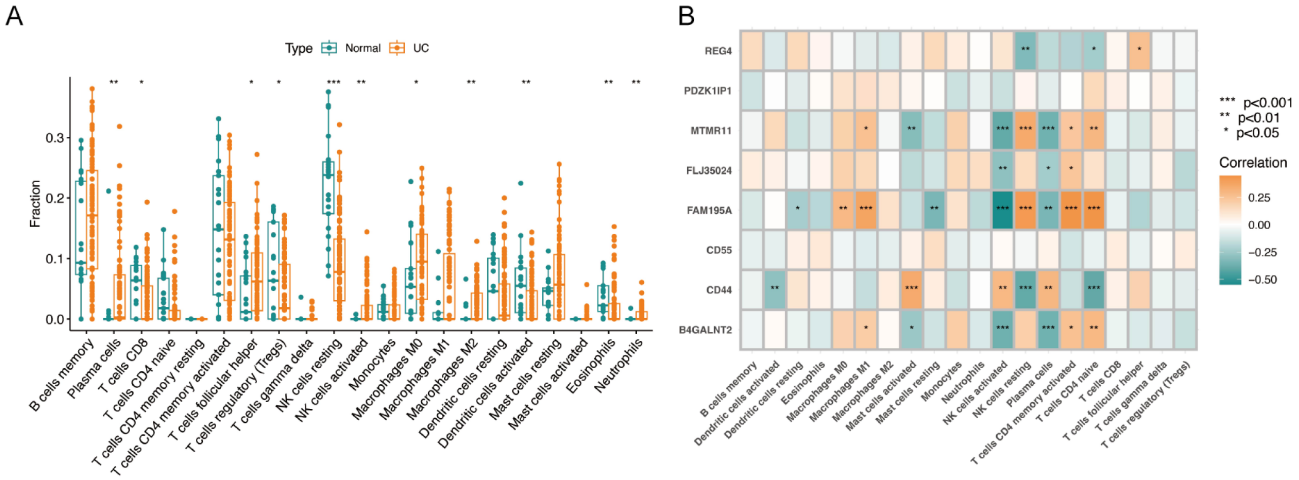

Fig. 9. Analysis of immune infiltration in UC group and normal control group. (A) The difference in immune infiltration between the UC and normal control groups with orange representing the UC group and green representing the normal control group.*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. (B) The correlation heatmap between 20 immune cells and eight feature genes with orange representing positive correlation and green representing negative correlation. The darker the color, the stronger the correlation. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

modulate the function of T cells and other immune cells by binding to CD44. The LGALS9-CD44 interaction inhibits T cell activation and proliferation, promotes immune tolerance, and reduces excessive immune responses. FN1 and CD44 ligand receptor pairs play important roles in cell adhesion, migration, and tissue repair. This ligand-receptor pair plays an important role in cellular communication between smooth muscle cells and fibroblasts (Fig. 12G). This is essential for the repair and regeneration of colonic epithelial cells after injury. COL1A2-CD44 also played an important role in smooth muscle cells and fibroblasts, which may be closely related to connective tissue formation and tissue repair (Fig. 12H). SELE, a member of the selectin family, is a cell adhesion molecule that is primarily expressed by activated endothelial cells. In immune and inflammatory responses, the SELE-CD44 ligand receptor modulates the migration and localization of immune cells and promotes cell-cell and cell-matrix interactions (Fig. 12I). The ADGRE5-CD55 ligand receptor pair is active in cellular communication between myeloid and T cells and may play a role in regulating immune responses and inflammation (Fig. 12J).

Discussion

The etiology of UC is multifactorial and not completely understood, with possible contributions from genetics, environmental factors, and immune dysregulation[1,8]. Currently, the clinical diagnosis of UC requires a comprehensive evaluation of clinical symptoms, medical history, laboratory tests, and imaging studies. However, there are certain limitations to this diagnostic approach. For example, obtaining tissue samples can be challenging, imaging studies may not always provide clear diagnostic evidence, and the lack of standardized and accurate diagnostic criteria can lead to delayed diagnosis or misdiagnosis of other gastrointestinal disorders[26]. Exploring the mechanisms of disease development at the genetic level, searching for new potential markers to construct diagnostic models, and analyzing the characteristics of immune cell infiltration in UC will provide a basis for early diagnosis of UC and the search for new therapeutic targets. The objective of this study is to develop highly sensitive diagnostic models by utilizing precise gene screening methods, and to investigate the immune cell infiltration characteristics of UC to provide novel insights for its treatment.

We initially identified a total of 53 disease-associated genes based on gene expression data from UC and normal controls, using two methods, differential analysis and WGCNA. Among these, eight genes (B4GALNT2, PDZK1IP1, FAM195A, REG4, MTMR11, FLJ35024, CD55, and CD44) were selected based on three machine learning algorithms (LASSO, RF, SVM-RFE) and PPI analysis. Using these eight genes as biomarkers for UC, we constructed a neural network model, mapped nomogram, and utilized an external validation set to evaluate the predictive performance of our diagnostic model. Additionally, we employed CIBERSORT to estimate the proportions of different immune cells in UC and normal control tissues and analyzed the correlations between our signature genes and various immune cells. We also utilized gene expression data to identify subtypes of the disease and assess the level of immune infiltration in different subtypes.

In the biomarker screening process, we utilized multiple methodologies, including conventional differential analysis, as well as emerging techniques such as WGCNA and machine learning. Differential analysis and WGCNA were employed to preliminarily screen for disease-associated genes. Genes identified through differential expression analysis may only be statistically associated with the disease state and not necessarily play a central role in biological functions or pathological processes. WGCNA assists in identifying genes that act as hubs within biological networks, often playing crucial roles in regulating disease-related biological pathways and processes. Intersecting the results from differential expression analysis and WGCNA helps in selecting genes that are significantly different in expression levels and also occupy central positions in disease-related biological

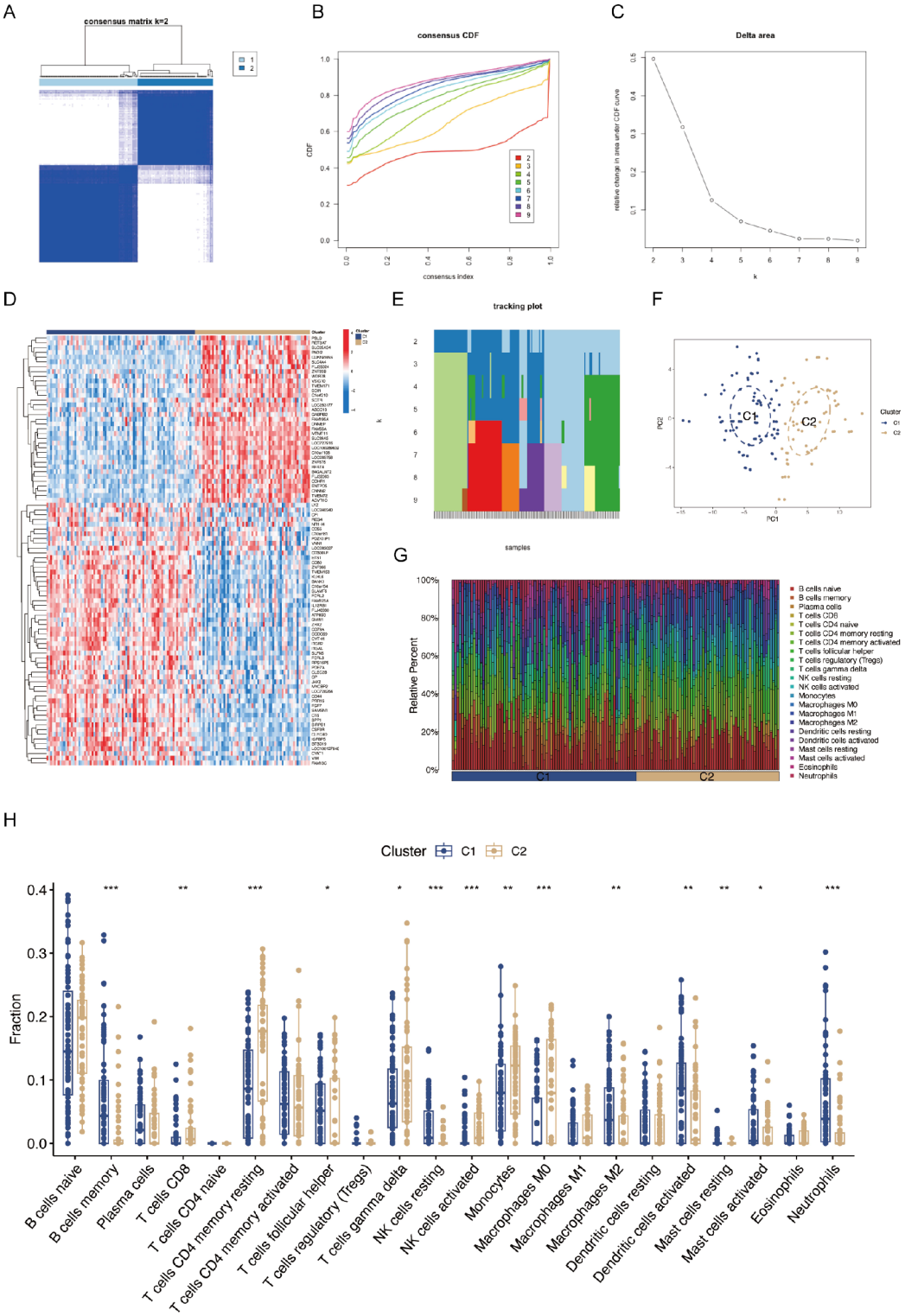

Fig. 10. Consensus clustering of UC samples and immune infiltration analysis between the subtypes of UC. (**A**) Consensus matrix plot, the cleaner the blank area between the blue modules indicates the more successful analysis. (**B**) CDF plot of consensus clustering, showing the relative change of consensus index from k = 2 to k = 9 with the change of CDF value, and the k value of the curve with the most stable change is the optimal fractal number. (**C**) Relative change in area under the CDF curve. (**D**) Heat map between C1 and C2 and genes, red indicates up-regulated gene expression and blue indicates down-regulated gene expression. (E) Trace plot of k = 2 to k = 9. (**F**) PCA plot of UC samples. The scatter plot allows visualization of the characteristic genes that classify UC into two subtypes, C1 and C2. (**G**) Stacked histogram of the infiltration ratio of 22 immune cells in each sample of the two subtypes C1 and C2. (**H**) Box plot of the difference in infiltration between C1 and C2 for the 22 immune cells. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

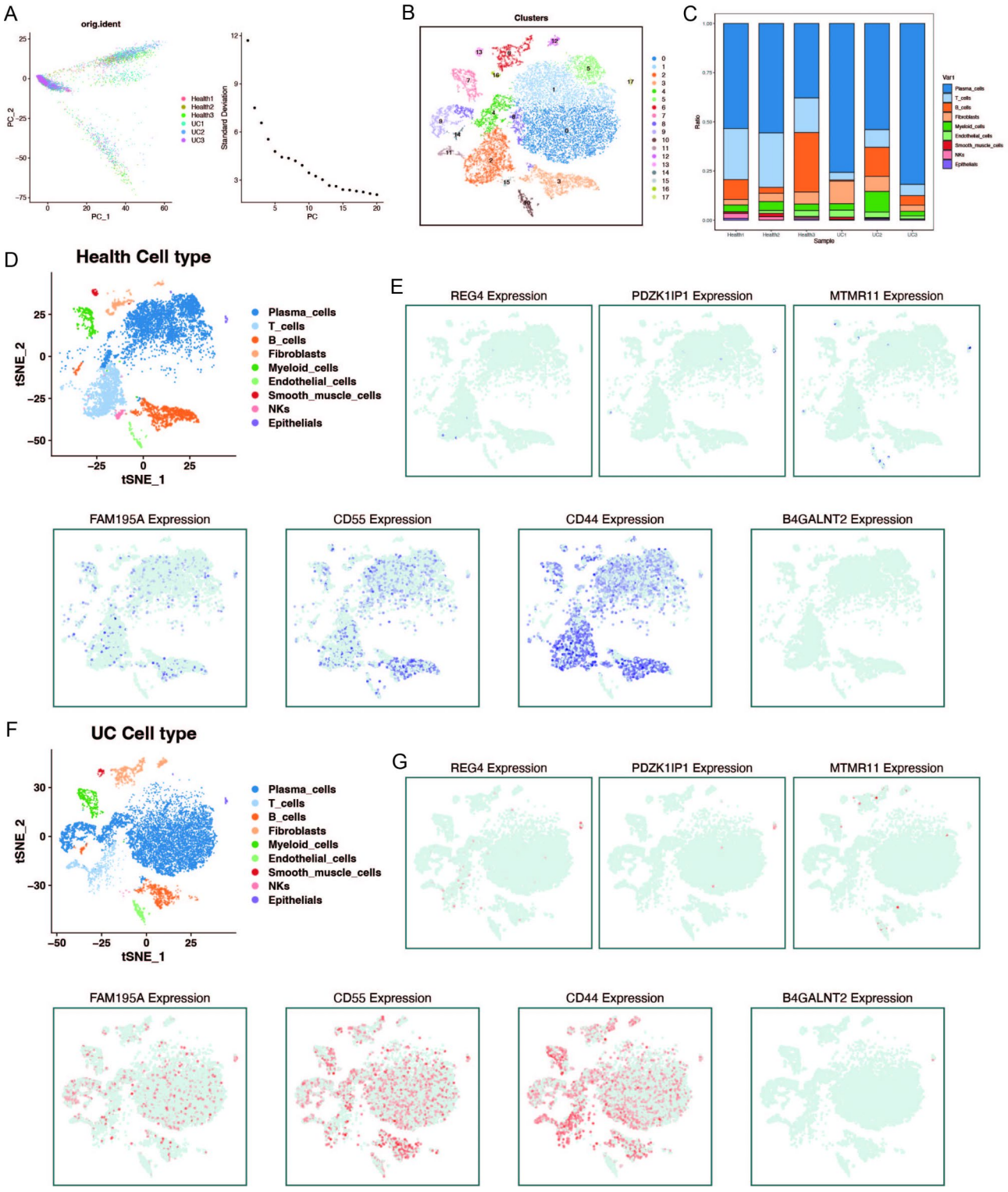

Fig. 11. Expression of diagnostic genes in single-cell data. (**A**) Left: PCA-based representation of cell distributions in the normal control group vs. UC group. Right: Contribution of the top 10 principal components. (**B**) t-SNE-based nonlinear dimensionality reduction clustering of all single-cell data, resulting in 18 cell clusters. (**D**) t-SNE distribution of cells in normal samples, with colors representing different cell types. (**E**) Feature plot illustrating the expression levels of diagnostic gene across cells in the normal group. (**F**) t-SNE distribution of cells in UC samples, with colors representing different cell types. (**G**) Feature plot illustrating the expression levels of diagnostic gene across cells in the UC group.

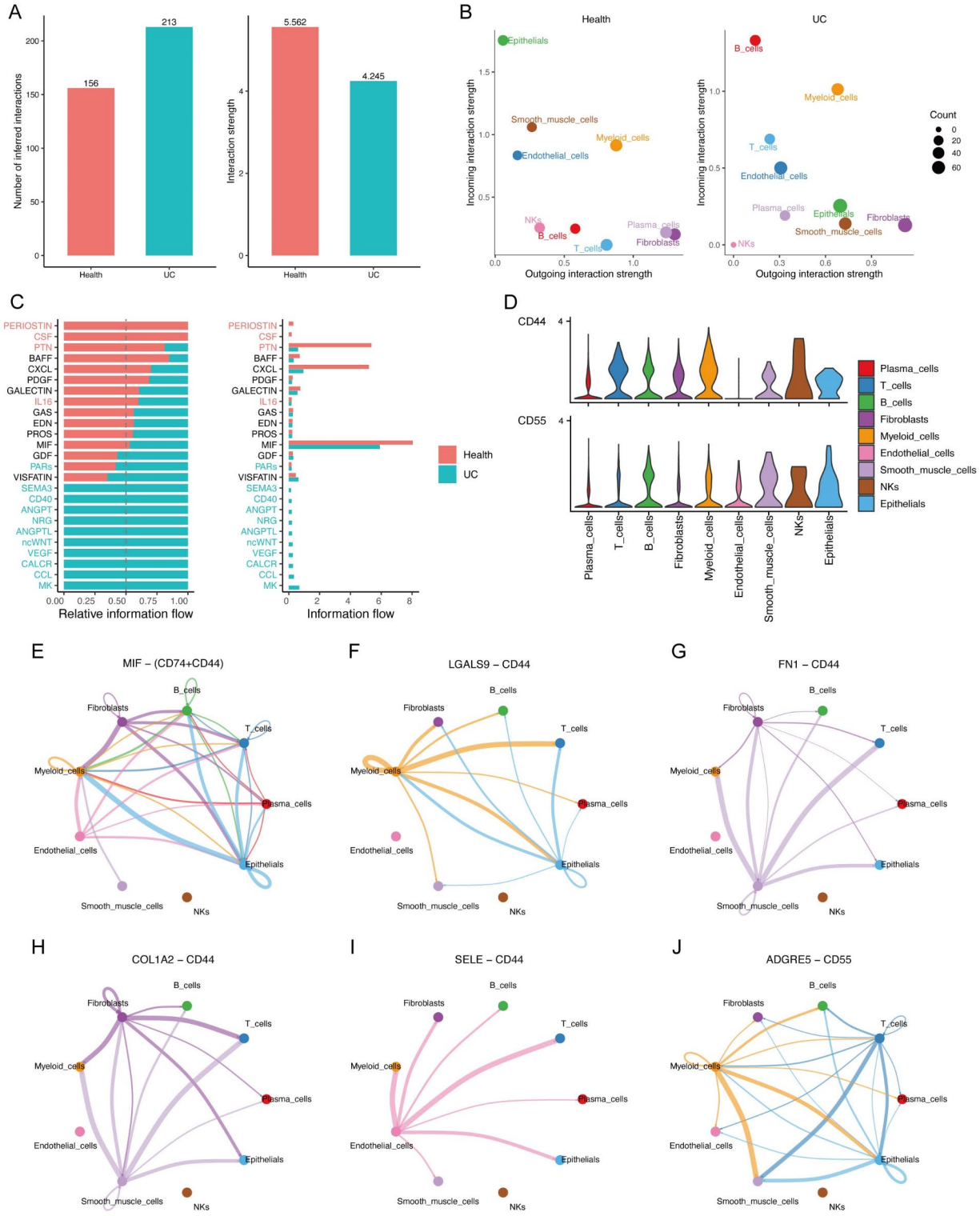

Fig. 12. Analysis of cellular communication networks. (**A**) Histogram of the number and intensity of communications. (**B**) Scatter plot of cell signal intensity. (**C**) Histogram of cellular communication family strength. (**D**) Violin plots of CD44 and CD55 expression in cells of various cell types. (**E**-**J**) Cellular communication networks mediated by specific ligand-receptor pairs.

networks. For further identification of disease biomarkers based on disease-associated genes, we employed machine learning and PPI analysis. PPI focuses on the biological connections and interactions among proteins, unveiling key participants in disease mechanisms. Machine learning algorithms analyze complex data patterns and predict disease markers based on feature importance and classification accuracy. These complementary

approaches address the limitations imposed by protein databases on PPI analyses and the constraints of machine learning algorithms, thereby enhancing the stability and accuracy of the biomarkers.

CD44, CD55, REG4, and PDKK1IP1 were upregulated in the disease group, while B4GALNT2, FAM195A, MTMR11, and FLJ35024 were downregulated. Single-gene GSEA analysis revealed their enrichment patterns, with most genes associated with the complement system and REG4 enriched in adherens junctions. CD44 contributes to cell adhesion, migration, and activation of inflammatory cells, while CD55 protects intestinal epithelial cells and maintains intestinal integrity. REG4 promotes intestinal epithelial cell proliferation and migration but can also worsen inflammation and injury. PDZK1IP1 regulates immune cell activation and apoptosis in intestinal epithelial cells. B4GALNT2 affects glycosylation reactions and downregulation disrupts the intestinal mucus layer, leading to cell destruction and dysbiosis. MTMR11 is involved in intracellular signaling and autophagy, and downregulation may increase apoptosis and inflammation. Limited research is available on the roles of FAM195A and FLJ35024 in UC, but FAM195A exhibits significant immune correlations with CD4 T cells and activated NK cells, providing new insights for further investigation. Furthermore, we utilized these potential biomarkers to develop a highly sensitive diagnostic model using a neural network. The model was trained with the training set and validated with an external dataset, demonstrating strong predictive capabilities and generalizability for UC diagnosis. Accurate UC diagnosis is crucial for early detection, improved disease management, reduced complications, and precise treatment guidance. A reliable diagnostic model minimizes medical resource waste, enhances medical care efficiency, and alleviates patient burden. To aid physicians in identifying high-risk patients accurately and efficiently, we created diagnostic line graphs that can assist in developing effective treatment plans.

When exploring the immune infiltrative features of UC, we observed elevated expression levels of multiple immune cell types in the disease group, including plasma cells, follicular helper T cells(Tfh cells), activated NK cells, macrophages, and neutrophils. Plasma cells, as mature B lymphocytes that produce antibodies, secrete mainly IgA in UC. This immunoglobulin is essential for neutralizing bacterial toxins and other pathogenic factors in the intestine, which helps to reduce the degree of intestinal inflammation. Additionally, the secretion of various cytokines and chemokines by plasma cells can also modulate the inflammatory response and influence immune regulation[27–30]. T follicular helper (Tfh) cells are a specialized subset of CD4 + T cells that play a key role in regulating the immune response of B cells. They provide critical help to B cells in the germinal centers of lymphoid tissues and promote B cell development, proliferation and antibody production. Inappropriate activation or abnormal function of Tfh cells can lead to an imbalance in the immune response and cause autoimmune diseases. Studies have shown that in patients with UC, Tfh cell numbers and activity are significantly elevated and positively correlate with disease severity. These cells can migrate into the intestinal mucosa and interact with B cells, leading to over-activation and release of large amounts of inflammatory cytokines and antibodies, thereby contributing to the development and progression of UC[31–33]. In patients with UC, dysbiosis and infections of the intestinal flora may worsen the intestinal inflammatory response. NK cells can kill pathogenic microorganisms, and their activity may help to remove these pathogens and reduce the intestinal inflammatory response. NK cells secrete a variety of cytokines that suppress the intestinal inflammatory response and promote epithelial cell repair and proliferation. In addition to this, NK cells make cellular contact with T cells and DC cells to regulate the immune response, thereby reducing the intestinal inflammatory response[34,35]. Macrophages assume a pivotal role in orchestrating the intricate dynamics of the inflammatory response within the intestinal milieu, alongside their indispensable contribution to the reparative processes of intestinal tissues. Initially, undifferentiated monocytes are recruited to the inflamed tissue and differentiate into macrophages (M0 macrophages). These M0 macrophages can further differentiate into two distinct phenotypes: M1 macrophages, which mainly promote the onset and development of the inflammatory response, and M2 macrophages, which play a role in suppressing the extent of the inflammatory response and promoting tissue repair[36]. Neutrophils, as the vanguard of the immune system, promptly migrate to the site of infection or tissue injury, assuming a pivotal role in their capacity to phagocytose and eradicate invading microorganisms. In UC, damage to the intestinal mucosal epithelium leads to a dysbiosis of the intestinal flora, which can cause an increase in the number and activity of neutrophils. However, this increase in neutrophil activity may also contribute to the exacerbation of the inflammatory response. In addition, neutrophils also release pro-inflammatory cytokines and reactive oxygen species, which can further damage the intestinal mucosa and promote the development of UC[37]. Additionally, the significant downregulation of Treg cells in the disease group caught our attention. Regulatory T cells (Tregs) are a type of immune cell that plays a role in immune regulation and autoimmune tolerance, including cellular contact and secretion of immunosuppressive factors. Treg cells can exert immunosuppressive effects by making cellular contacts with other immune cells, which include suppressing the function of immune cells such as CD4 + T cells, CD8 + T cells, NK cells and DC cells. In addition, Treg cells can also secrete immunosuppressive factors to suppress the activity of immune cells and reduce the inflammatory response in the gut[38,39]. On the other hand, Treg cells can also promote the restoration and repair of the intestinal epithelial barrier, thereby reducing the intestinal inflammatory response[40]. In summary, the complex interplay among Tfh cells, NK cells, Treg cells, and other immune cells highlights their potential importance in the pathogenesis of UC and may offer novel insights for the development of immunotherapeutic approaches for the disease.

We used gene expression profiling data from UC patients to identify disease subtypes and analyzed the resulting two disease subtypes for immune infiltration. The PCA analysis of the disease typing results demonstrated a clear clustering of UC patients into two distinct categories, providing further evidence for the validity of our disease classification into two subtypes. When comparing the level of immune infiltration between the subtypes, the two showed significant differences in a variety of immune cells, including Tfh cells, NK cells, macrophages, and neutrophils. For the disease, there may be significant differences in gene expression, metabolic profiles and levels of immune infiltration between subtypes, and these differences may lead to different susceptibilities to drug therapy in different subtypes. In UC, some patients may respond well to immunosuppressive or biological

agents, while others may require surgical treatment. By classifying and exploring UC subtypes, it is possible to identify subtypes that are associated with specific treatment responses. This can help doctors in selecting appropriate treatments and optimizing personalized treatment plans for patients.

CD44 and CD55 play a very important role in the cellular communication network when analyzed using single cell data. Especially CD44, which forms different ligand-receptor pairs with multiple ligands, is very active in cellular regulation. At the same time, it plays a complex role, either promoting immune infiltration and inflammatory response or suppressing the response of immune cells to control inflammation. By analyzing the cellular communication function of CD44 can provide new insights into the pathomechanism of UC, and also suggests to us that CD44 may be a potential therapeutic target and strategy. Therapeutic strategies targeting CD44 (e.g., anti-CD44 monoclonal antibodies) may be a new therapeutic tool for UC, an autoimmune disease.

In developing our diagnostic model, we faced several key data challenges. First, the number of disease samples used to train the model was larger than the normal control samples. Considering to avoid the batch effect caused by different datasets, we did not adopt the up-sampling method; at the same time, in order to maintain the stability of the total number of samples, we also did not adopt the down-sampling method. Therefore, in order to validate the performance of the diagnostic model and optimize the model, more normal samples equivalent to the number of disease samples need to be collected in the clinic in the future. In addition, there are some limitations of our study. Ulcerative colitis (UC) is a complex disease, and its risk is influenced by a variety of factors, including genetics, and the prevalence varies significantly among different races. For example, the prevalence of UC is lower in Asian populations compared to European and American populations, whereas it is relatively high among Caucasians in Europe and the United States, but lower among blacks. Patients of different races also showed differences in disease symptoms, clinical manifestations, and pathologic features[41]. Our study did not cover the differences in gene expression among different racial groups. Therefore, further testing of the performance of the diagnostic model will need to consider its accuracy in different populations as well as its ability to generalize. Given the limitations of data availability, external validation of the model should use supplemental datasets to ensure the accuracy and robustness of the diagnostic model. Validation through clinical trials using UC patient samples will be a critical part of this process. Although we have successfully identified diagnostic genes associated with UC, more experimental studies are needed to reveal the complex mechanisms controlled by these key genes.

Conclusion

In conclusion, this study utilized various methods, including WGCNA and machine learning algorithms, to identify potential biomarkers for UC, develop a diagnostic model, and explore the underlying immune mechanisms of the disease, which may provide new insights for immunotherapy. These results are significant in enhancing the accuracy of diagnosis and improving the treatment outcomes of UC.

Data availability

The datasets analyzed in the current study are available in the GEO repository (https://www.ncbi.nlm.nih.gov/geo/). All raw data, codes are available at https://www.jianguoyun.com/p/DXpSfhUQ85jYCxj-rIkFIAA.

Received: 18 November 2023; Accepted: 8 October 2024
Published online: 16 October 2024

References

1. Adams, S. M. & Bornemann, P. H. Ulcerative colitis. *Am. Fam Physician*. **87**, 699–705 (2013).
2. Du, L. & Ha, C. Epidemiology and Pathogenesis of Ulcerative Colitis. *Gastroenterol. Clin. North. Am.* **49**, 643–654. https://doi.org/10.1016/j.gtc.2020.07.005 (2020).
3. Guan, Q. A. & Comprehensive review and update on the pathogenesis of inflammatory bowel disease. *J. Immunol. Res.* 7247238. https://doi.org/10.1155/2019/7247238 (2019).
4. Kappelman, M. D. & Bousvaros, A. Nutritional concerns in pediatric inflammatory bowel disease patients. *Mol. Nutr. Food Res.* **52**, 867–874. https://doi.org/10.1002/mnfr.200700156 (2008).
5. Kaenkumchorn, T. & Wahbeh, G. Ulcerative Colitis: making the diagnosis. *Gastroenterol. Clin. North. Am.* **49**, 655–669. https://doi.org/10.1016/j.gtc.2020.07.001 (2020).
6. Matsumoto, T., Nakamura, S., Okawa, K. & Kitano, A. Differential diagnosis of ulcerative colitis. *Nihon Rinsho* **57**, 2461–2465 (1999).
7. Viennot, S. et al. Colon cancer in inflammatory bowel disease: recent trends, questions and answers. *Gastroenterol. Clin. Biol.* **33**(Suppl 3), 190–201. https://doi.org/10.1016/s0399-8320(09)73154-9 (2009).
8. Le Berre, C., Honap, S. & Peyrin-Biroulet, L. Ulcerative colitis. *Lancet*. **402**, 571–584. https://doi.org/10.1016/s0140-6736(23)00966-2 (2023).
9. Gros, B. & Kaplan, G. G. Ulcerative colitis in adults: a review. *Jama*. **330**, 951–965. https://doi.org/10.1001/jama.2023.15389 (2023).
10. Wehkamp, J. & Stange, E. F. Recent advances and emerging therapies in the non-surgical management of ulcerative colitis. *F1000Res* **7** (2018). https://doi.org/10.12688/f1000research.15159.1
11. Zhang, J. et al. Investigation of Potential Genetic Biomarkers and Molecular Mechanism of Ulcerative Colitis Utilizing Bioinformatics Analysis. *Biomed. Res. Int.* **2020**, 4921387. https://doi.org/10.1155/2020/4921387 (2020).
12. Tan, R. et al. Identification of early diagnostic and prognostic biomarkers via WGCNA in stomach adenocarcinoma. *Front. Oncol.* **11**, 636461. https://doi.org/10.3389/fonc.2021.636461 (2021).
13. Zhang, S. et al. Uncovering the immune microenvironment and molecular subtypes of hepatitis B-related liver cirrhosis and developing stable a diagnostic differential model by machine learning and artificial neural networks. *Front. Mol. Biosci.* **10**, 1275897. https://doi.org/10.3389/fmolb.2023.1275897 (2023).
14. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell*. **184**, 3573–3587e3529. https://doi.org/10.1016/j.cell.2021.04.048 (2021).
15. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47. https://doi.org/10.1093/nar/gkv007 (2015).

16. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559. https://doi.org/10.1186/1471-2105-9-559 (2008).
17. Langfelder, P., Horvath, S. & Fast, R. Functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* **46**, 1–17. https://doi.org/10.18637/jss.v046.i11 (2012).
18. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics.* **16**, 284–287. https://doi.org/10.1089/omi.2011.0118 (2012).
19. Zhang, S. et al. Construction of a diagnostic model for hepatitis B-related hepatocellular carcinoma using machine learning and artificial neural networks and revealing the correlation by immunoassay. *Tumour Virus Res.* **16**, 200271. https://doi.org/10.1016/j.tvr.2023.200271 (2023).
20. Szklarczyk, D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–d646. https://doi.org/10.1093/nar/gkac1000 (2023).
21. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized Linear models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22. https://doi.org/10.18637/jss.v033.i01 (2010).
22. Chi, H. et al. Proposing new early detection indicators for pancreatic cancer: combining machine learning and neural networks for serum miRNA-based diagnostic model. *Front. Oncol.* **13**, 1244578. https://doi.org/10.3389/fonc.2023.1244578 (2023).
23. Beck, M. W. & NeuralNetTools Visualization and analysis tools for neural networks. *J. Stat. Softw.* **85**, 1–20. https://doi.org/10.18637/jss.v085.i11 (2018).
24. Robin, X. et al. pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77. https://doi.org/10.1186/1471-2105-12-77 (2011).
25. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. **26**, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170 (2010).
26. Tontini, G. E., Vecchi, M., Pastorelli, L., Neurath, M. F. & Neumann, H. Differential diagnosis in inflammatory bowel disease colitis: state of the art and future perspectives. *World J. Gastroenterol.* **21**, 21–46. https://doi.org/10.3748/wjg.v21.i1.21 (2015).
27. Arató, A. & Savilahti, E. [Distribution of cells containing different IgG and IgA subclasses in the colonic mucosa in childhood ulcerative colitis and Crohn disease]. *Orv Hetil.* **132**, 2027–2031 (1991).
28. Baklien, K. & Brandtzaeg, P. Comparative mapping of the local distribution of immunoglobulin-containing cells in ulcerative colitis and Crohn's disease of the colon. *Clin. Exp. Immunol.* **22**, 197–209 (1975).
29. Spencer, J. & Bemark, M. Human intestinal B cells in inflammatory diseases. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 254–265. https://doi.org/10.1038/s41575-023-00755-6 (2023).
30. Uzzan, M. et al. Ulcerative colitis is characterized by a plasmablast-skewed humoral response associated with disease activity. *Nat. Med.* **28**, 766–779. https://doi.org/10.1038/s41591-022-01680-y (2022).
31. Long, Y. et al. Increased circulating PD-1(hi)CXCR5- peripheral helper T cells are associated with disease severity of active ulcerative colitis patients. *Immunol. Lett.* **233**, 2–10. https://doi.org/10.1016/j.imlet.2021.03.001 (2021).
32. Long, Y. et al. The Imbalance of Circulating Follicular Helper T Cells and Follicular Regulatory T Cells is Associated with Disease activity in patients with Ulcerative Colitis. *Front. Immunol.* **11**, 104. https://doi.org/10.3389/fimmu.2020.00104 (2020).
33. Wang, X. et al. The shifted balance between circulating follicular regulatory T cells and follicular helper T cells in patients with ulcerative colitis. *Clin. Sci. (Lond).* **131**, 2933–2945. https://doi.org/10.1042/cs20171258 (2017).
34. Fuss, I. J. & Strober, W. The role of IL-13 and NK T cells in experimental and human ulcerative colitis. *Mucosal Immunol.* **1** (Suppl 1), 31–33. https://doi.org/10.1038/mi.2008.40 (2008).
35. Geremia, A., Biancheri, P., Allan, P., Corazza, G. R. & Di Sabatino, A. Innate and adaptive immunity in inflammatory bowel disease. *Autoimmun. Rev.* **13**, 3–10. https://doi.org/10.1016/j.autrev.2013.06.004 (2014).
36. Dharmasiri, S. et al. Human intestinal macrophages are involved in the Pathology of both Ulcerative Colitis and Crohn Disease. *Inflamm. Bowel Dis.* **27**, 1641–1652. https://doi.org/10.1093/ibd/izab029 (2021).
37. Dinallo, V. et al. Neutrophil Extracellular traps sustain inflammatory signals in Ulcerative Colitis. *J. Crohns Colitis*. **13**, 772–784. https://doi.org/10.1093/ecco-jcc/jjy215 (2019).
38. Barbi, J., Pardoll, D. & Pan, F. Treg functional stability and its responsiveness to the microenvironment. *Immunol. Rev.* **259**, 115–139. https://doi.org/10.1111/imr.12172 (2014).
39. Göschl, L., Scheinecker, C. & Bonelli, M. Treg cells in autoimmunity: from identification to Treg-based therapies. *Semin Immunopathol.* **41**, 301–314. https://doi.org/10.1007/s00281-019-00741-8 (2019).
40. Larabi, A., Barnich, N. & Nguyen, H. T. T. New insights into the interplay between autophagy, gut microbiota and inflammatory responses in IBD. *Autophagy*. **16**, 38–51. https://doi.org/10.1080/15548627.2019.1635384 (2020).
41. Barnes, E. L., Loftus, E. V., Kappelman, M. D. & Jr. & Effects of Race and ethnicity on diagnosis and management of Inflammatory Bowel diseases. *Gastroenterology*. **160**, 677–689. https://doi.org/10.1053/j.gastro.2020.08.064 (2021).

Author contributions

LJ , HC, GY and QW conceived the study. LJ, SZ, CJ, HQC, JY and JH drafted the manuscript. LJ, SZ and CJ performed the literature search and collected the data. LJ, HQC and JH analyzed and visualized the data. HC, GY and QW helped with the final revision of this manuscript. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the Science and Technology Development Fund, Macau SAR (No.: 0098/2021/A2 and 0048/2023/AFJ), the National Natural Science Foundation of China (No.: 82361168663), Chinese Medicine Guangdong Laboratory (HQCML-C-2024007), and Macau University of Science and Technology's Faculty Research Grant (No: FRG-23-003-FC).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study was exclusively based on public sequencing data from the GEO public database, without involvement of any human or animal subjects.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-75797-0.

Correspondence and requests for materials should be addressed to H.C., Q.W. or G.Y.

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.