scientific reports

OPEN

Identification of diagnostic biomarks and immune cell infiltration in ulcerative colitis

Qin Chen [1,4]✉, Shaosheng Bei[2,4], Zhiyun Zhang[1], Xiaofeng Wang[3] & Yunying Zhu[1]

We aimed to explore diagnostic biomarks and immune cell infiltration characteristics in ulcerative colitis (UC). We used the dataset GSE38713 as the training set and dataset GSE94648 as the test set. A total of 402 differentially expressed genes (DEGs) were obtained from GSE38713. Annotating, visualizing, and integrating discovery of these differential genes was performed using Gene Ontology (GO), Kyoto Gene and Genome Encyclopedia Pathway (KEGG), and Gene Set Enrichment Analysis (GSEA). Protein–protein interaction networks were constructed from the STRING database, and protein functional modules were identified using the CytoHubba plugin of Cytoscape. Random forest and LASSO regression were used to screen for UC-related diagnostic markers, and ROC curves were generated to validate their diagnostic value. The composition of 22 immune cells was analyzed, and the immune cell infiltration in UC was analyzed using CIBERSORT. Results: Seven diagnostic markers associated with UC were identified: TLCD3A, KLF9, EFNA1, NAAA,WDR4, CKAP4, and CHRNA1. Immune cell infiltration assessment revealed that macrophages M1, activated dendritic cells, and neutrophil cells infiltrated relatively more compared to normal control samples. Our results suggest a new functional feature of UC and suggest potential biomarkers for UC through comprehensive analysis of integrated gene expression data.

Ulcerative colitis (UC) is a chronic nonspecific inflammatory disease of the intestine, that is characterized by persistent or recurrent abdominal pain, diarrhea, and mucopurulent stools. The incidence and prevalence of UC are increasing worldwide[1], thus increasing the medical and economic burden on society. Therefore, the exploration of diagnostic biomarkers and therapeutic targets has become a focal issue for improving the prognosis of UC.

The etiology and pathogenesis of UC are still unclear, and current research suggests that it is mainly caused by the interaction of genetic susceptibility, epithelial barrier defects, immune system dysfunction, and environmental factors[2,3]. Among all the factors, an impaired immune response plays an important role in the development and progression of UC[4]. Both innate and adaptive immunity have been shown to play important roles in intestinal inflammation[5]. When the tolerance mechanisms of the intestinal barrier fail, local immune cells are stimulated, resulting in production of chemokines and subsequent infiltration of immune cells. Thus, the inflammatory process is further exacerbated[4]. Studies have shown that the cytokines interleukin (IL)-13, TNF, IL-23, IL-9, and IL-36 promote inflammatory immune cell infiltration and are important in the pathogenesis of UC[6]. Different types of immune cells, whether in an activated or inactivated state, can modulate the immune response by inhibiting, maintaining, or promoting the development of UC[7].

With the completion of the Human Genome Project, histological technologies, mainly high-throughput microarray analysis and bioinformatics analysis, have provided reliable technical support for studying the pathological mechanisms of complex diseases[8]. Several relevant studies have used microarray analysis to show the involvement of differentially expressed genes (DEGs) in biological functions and pathways contributing to the development of UC, as well as potential biomarkers that are immunologically relevant to patients with UC[9,10]. However, the biomarkers that have been identified are still less accurate in the diagnosis and prognosis of UC, mainly due to the complexity of UC pathogenesis. Different microarray platforms and small sample sizes may have led to inconsistent results in these studies. Further comprehensive analyses are necessary to identify new, more reliable diagnostic biomarkers and therapeutic targets to overcome these inconsistencies.

[1]Department of Anorectal, Kunming Municipal Hospital of Traditional Chinese Medicine, The Third Affiliated Hospital of Yunnan University of Chinese Medicine, No. 25 Dongfeng East Road, Panlong District, Kunming 650011, Yunnan Province, China. [2]Department of Anorectal, Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing, China. [3]Department of Colorectal Surgery, Guang'an Men Hospital, China Academy of Chinese Medical Sciences, Beijing, China. [4]These authors contributed equally: Qin Chen and Shaosheng Bei. ✉email: hichenqin2011@163.com

1

Therefore, in the current study, we screened DEGs in UG samples using microarray sequencing of UC from the Gene Expression Omnibus (GEO) database, which included 30 UC patients and 13 normal controls and performed functional enrichment analysis of Gene Ontology (GO), Kyoto Gene and Genome Encyclopedia Pathway (KEGG), and Gene Set Enrichment Analysis (GSEA), and constructed a protein–protein interaction (PPI) network to explore important protein action modules, while random forest and LASSO regression were used to screen for the diagnostic markers of UC. We also used another dataset for ROC validation, and used CIBERSORT[11] to calculate its immune cell composition and analyze its correlation with UC. We assessed the immune cell infiltration in UC, which provides new ideas for further research on the molecular mechanism underlying UC pathogenesis.

Materials and methods

Data download and pre-processing. UC expression profiles with reliable sample sources were downloaded from the GEO (https://www.ncbi.nlm.nih.gov/geo/) database using the GEOquery package[12] of the R software (version 3.6.5, http://r-project.org/). The dataset GSE38713[13] and GSE94648[14] with samples from *Homo sapiens* and platforms based on GPL570 and GPL19109 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array were used. The GSE38713 dataset included 30 UC patient samples and 13 normal samples and GSE94648 dataset included 25 UC patient samples and 22 normal samples, both of which were included in this study (Table 1). The raw data of the GSE38713 and GSE94648 datasets were read using the affy package[15]; RMA background correction, and data normalization were performed to obtain the gene expression matrices of the two datasets. HUGO Gene Nomenclature Committee (HGNC)[16] is responsible for providing a unique, standardized and widely disseminated symbol for all genes on the human genome including protein-coding genes, non-coding RNA genes, methyl genes and other genes; for each human gene, mRNA expression profiles were obtained using the HGNC mRNA gene annotation file.

Identification of DEGs. We used GSE38713 as the training set and GSE94648 as the test set. The GSE38713 dataset was screened for differentially expressed genes (DEGs) using the limma package[17],and the volcano plot of DEGs was plotted using the ggplot2[18] package. Criterion for selection was adj. p value < 0.05, and | log2FC|> 1. Around 402 genes were found to be differentially expressed.

PPI Network Analysis and Identification of Key Genes. The STRING[19] database searches were used to identify interactions between known proteins and predicted proteins. We used the DEGs obtained from differential expression analysis and put them into the STRING database to obtain their protein interaction networks, and then put the networks into Cytoscape[20] software to identify the genes that interact more strongly with other genes and visualize them. Using the MCODE[21] plugin to identify its sub-networks and based on the score, the three highest-rated sub-networks were obtained, which we believed may serve a specific function.

Functional enrichment analysis. GO[22] is a database established by the Gene Ontology Consortium to create a semantic vocabulary standard for qualifying and describing gene and protein functions for a wide range of species that can be updated as research progresses. GO annotations are divided into three broad categories: molecular function (MF), biological process (BP), and cellular components (CC). KEGG[23–25] is a comprehensive database that integrates genomic, chemical, and systemic functional information. KEGG database specifically stores information about gene pathways in different species. Metascape[26] is a web tool that provides a variety of functions such as gene enrichment analysis and protein interaction network analysis. The website integrates more than 40 gene function annotation databases and provides diverse visualizations. We used Metascape to perform GO/KEGG functional enrichment analysis of differentially expressed genes, selecting functions with p < 0.01, minimum count of 3, and enrichment factor > 1.5. We also used the R package Pathview[27] to visualize the more important pathways in KEGG and R package ggplot2 to visualize the more important functions in GO.

GSEA functional enrichment analysis. GSEA[28] is based on the idea of using predefined gene sets (usually from functional annotations or results of previous experiments) to rank genes according to their differential expression in two types of samples, and then testing whether the predefined set of genes is enriched at the top or bottom of the ranking table. We used the clusterProfiler package[29] to analyze the gene expression profile of GSE38713 using the GSEA method, selecting "c2.cp.kegg.v7.4.symbols.gmt" and "c5.go.bp.v7.4.symbols.gmt" as the reference gene set[30] , and $p < 0.05$ was considered significantly enriched.

Random Forest identification for signature genes. For the 402 differentially expressed genes obtained, the RandomForest package was used to filter the feature genes. RandomForest[31] package in R was used to construct a random forest for the 402 differentially expressed genes. The larger the Gini coefficient, the better the classification, and the larger the decrease in the Gini value when selecting a certain point, the better the clas-

Dataset	Patient	Control
GSE38713	30	13
GSE94648	25	22

Table1. Related information of dataset Platform (Affymetrix Human Genome U133 Plus 2.0 Array).

sification. The parameter MeanDecreaseGini[30] is the average decreasing GINI value, the more it decreases, the better the classification effect of this node. This node is chosen as the classification node, that is, the node with the largest GINI value as the classification node. The first 2/3 nodes with the best classification effect (keeping the largest MeanDecreaseGini and removing the first 1/3 nodes with small ones) were selected, and the nodes with poor classification effect were removed. Finally, a total of six rounds were screened to obtain 54 feature genes that contributed more to the classification.

Identification and validation of diagnostic markers. LASSO is a shrinkage estimation method that allows variable selection by constructing a penalty function that can compress the coefficients of variables and make regression coefficients zero. We used the LASSO regression algorithm for feature selection to screen for diagnostic markers of UC based on feature genes obtained from random forest. The GSE94648 dataset was used as a test set to validate the diagnostic efficacy of the obtained diagnostic markers, and use GEPIA2[32] to analyze the prognosis of the obtained diagnostic markers and UC-related Colorectal Cancer (CRC) in the TCGA database.

Immune cell infiltration analysis. CIBERSORT[11] is based on the principle of linear support vector regression to deconvolute the transcriptome expression matrix and to estimate the composition and abundance of immune cells in a mixture of cells[33]. We downloaded the original code and the corresponding immune cell files from the CIBERSORT official website and derived the immune cell infiltration matrix in R based on the gene expression profile of GSE38713 and the immune cell files. We used the corrplot package[34] to plot correlation heat maps and visualize the correlation of the 22 immune cell infiltrates. The ggplot2 package was used to plot box line plots for visualizing the differences between the infiltration of 22 immune cells; igraph package[35] was used to plot correlation network plots of immune cell infiltrates for visualizing the interactions of the 22 immune cell infiltrates, and $p < 0.05$, |correlation coefficient > 0.4) were used as the criteria for interactions. We correlated the obtained diagnostic markers with immune cell infiltrates and then visualized the results using the pheatmap[36] package.

Results

Data download and pre-processing. The data analysis process is illustrated in Fig. 1. First, the gene expression matrices from the GEO official website GSE38713 and GSE94648 datasets (Table 1) were normalized and processed based on the RMA method using the affy package. The two datasets were found to be more suitable for analysis since they had more positive data (Fig. 2). The protein gene annotation files were downloaded from HGNC, and 16,930 mRNAs were obtained after matching.

DEG Analysis. After data preprocessing, we performed differential expression analysis on the GSE38713 expression matrix using the R package limma, with $|logFC| > 1$ and adj. p value < 0.05 as the threshold screening, and a total of 402 DEGs, 242 upregulated genes, and 160 downregulated genes were extracted from the gene expression matrix. The distribution of DEGs is shown in the volcano plot (Fig. 3A). We then performed a hierarchical clustering analysis of the 402 DEGs in GSE38713 and GSE94648, and found that the majority of
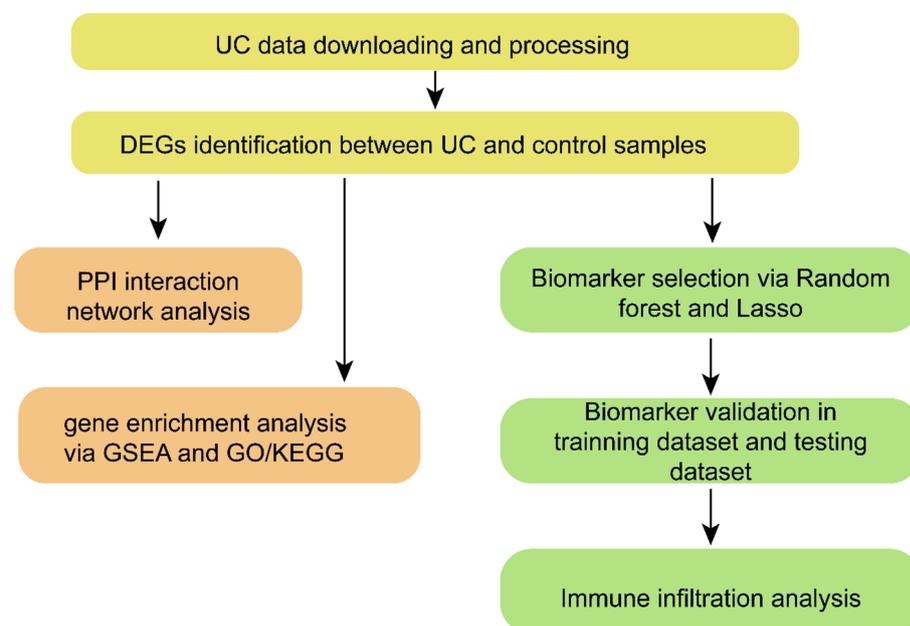

Figure 1. Flow chart of data analysis.
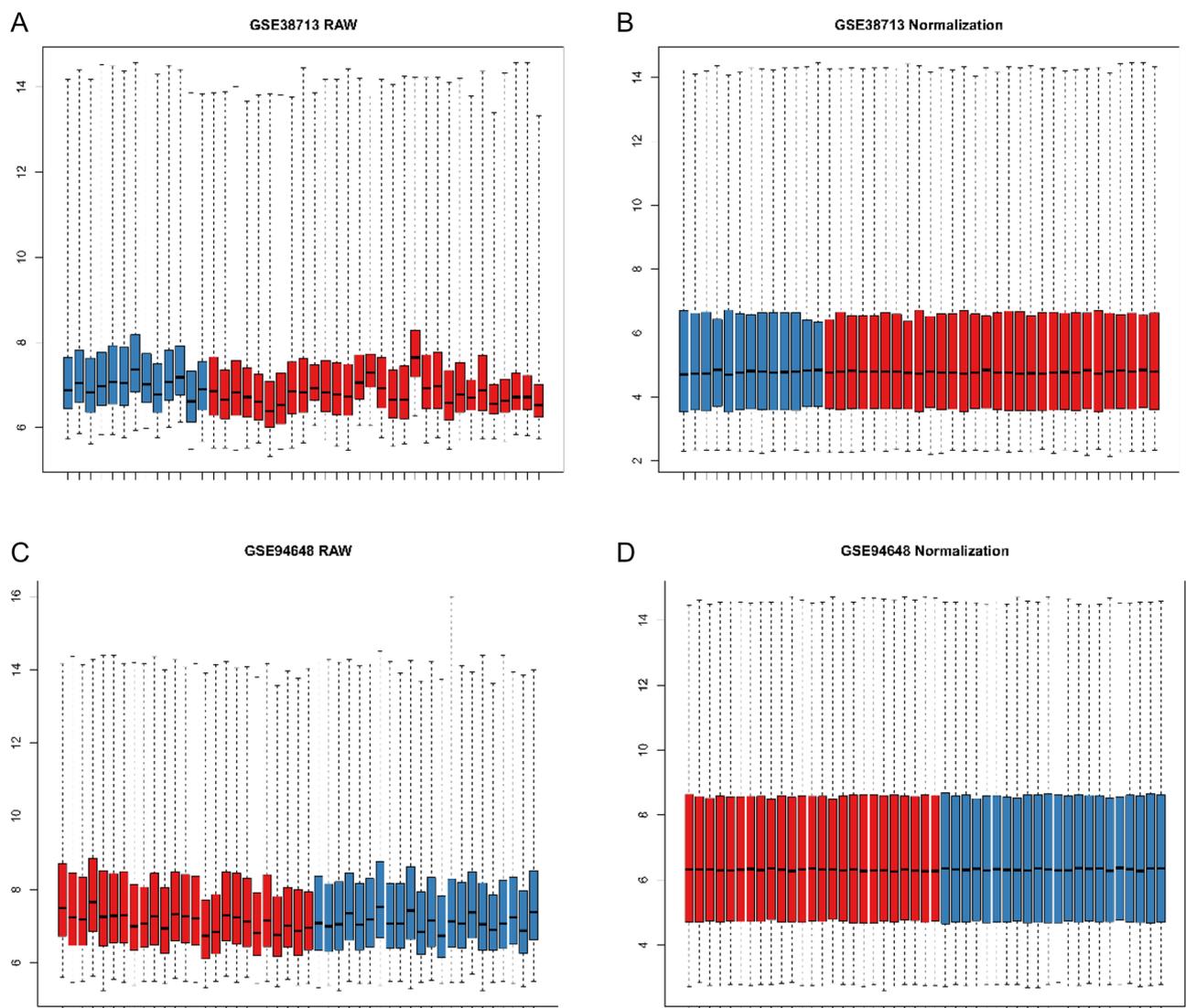
Figure 2. Data box plots. (**A**) GSE38713 uncorrected box plot; (**B**) GSE38713 corrected box plot; (**C**) GSE94648 uncorrected box plot; (**D**) GSE94648 corrected box plot. Blue represents normal samples, red represents disease samples.

disease samples were clustered into one category and normal samples were clustered into a different category (Fig. 3B,C).

PPI Network Analysis and Identification of Key Genes. We placed the 402 DEGs into the STRING database to obtain their PPI networks (Fig. 4A), and the PPI networks were placed into Cytoscape to identify and visualize important genes with strong interactions with other genes (Fig. 4B). The MCODE plug-in was used to identify the three sub-networks with the highest scores (Fig. 4C–E) (tableS2-Supplement 1).

Functional enrichment analysis. We first performed a functional enrichment analysis of DEGs using Metascape to screen for function at $p < 0.01$, a minimum count of 3, and an enrichment factor > 1.5. The DEGs were mainly associated with extracellular matrix organization, inflammatory response, humoral immune response, apical part of cell, external encapsulating structure, carboxylic acid transmembrane transporter activity, protein digestion and absorption, ECM-receptor interaction, complement and coagulation cascades, and PI3K-Akt signaling pathway (Fig. 5) (tableS3-Supplement 2). The detailed enrichment results are shown in Supplement 3.

GSEA functional enrichment analysis. We first downloaded the gene sets "c2.cp.kegg.v7.4.symbols. gmt" and "c5.go.bp.v7.4.symbols.gmt." The GSEA function in the clusterProfiler package was used to enrich the GSE38713 expression profile with "c2.cp.kegg.v7.4.symbols.gmt" and "c5.go.bp.v7.4.symbols.gmt" as reference gene sets[30] (these two are more commonly used in functional enrichment). We used p value < 0.05 as the
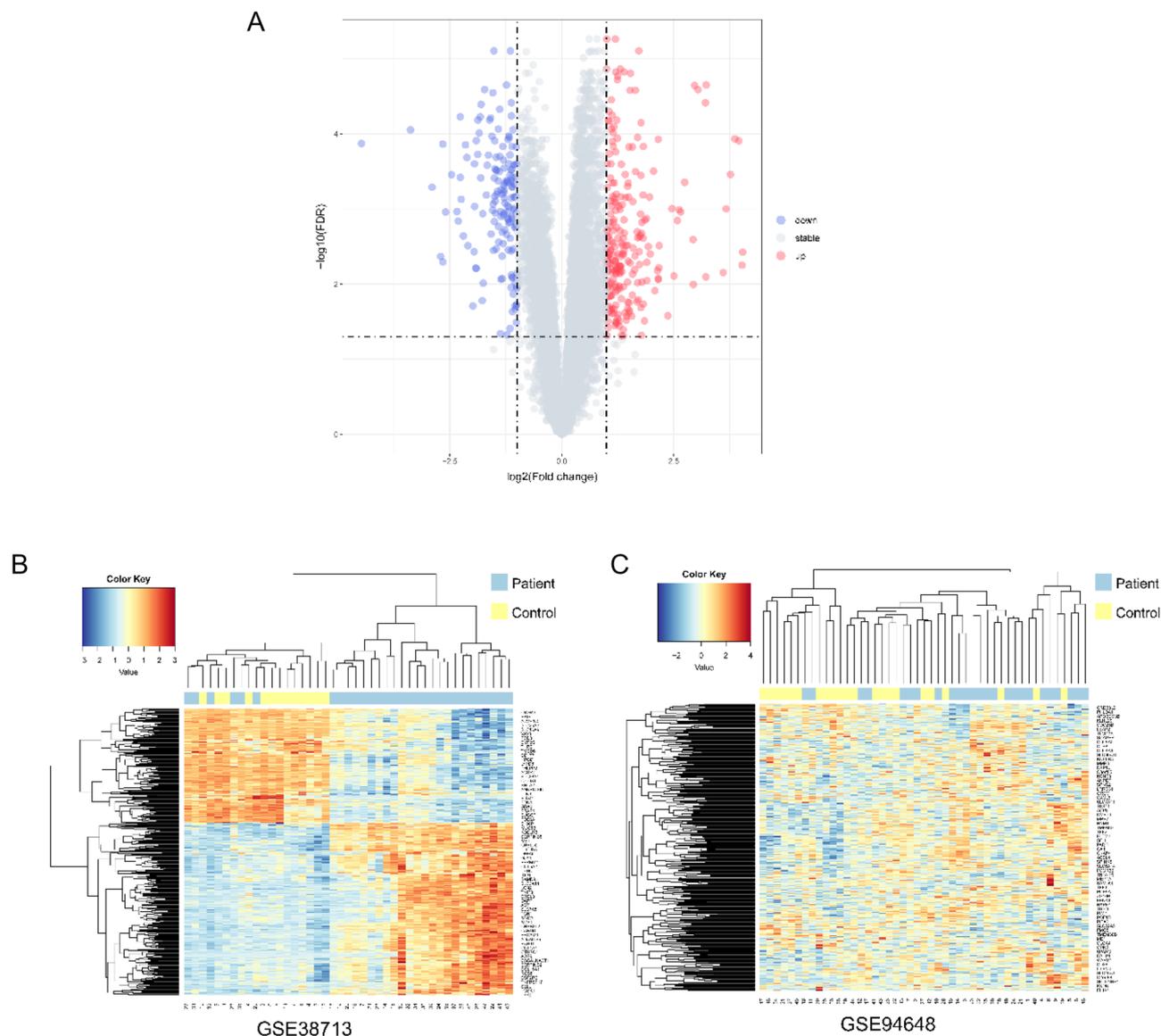
Figure 3. Volcano and Heat map. (**A**) Volcano map of GSE38713 DEGs, red represents up-regulated differential genes, blue represents down-regulated differential genes, and gray represents no differential genes; (**B**) heat map of GSE38713 clusters; (**C**) heat map of GSE94648 clusters. Yellow represents the control group; blue represents the UC group.

threshold to screen for differential functions. The results of KEGG and GO enrichments are shown in Fig. 6 (Table 4-Supplement 4) The main enrichments in GO BP were: Divalent inorganic cation homeostasis, regulation of body fluid levels, positive regulation of MAPK cascade. The main enrichments in KEGG were: cytokine receptor interaction, focal adhesion, chemokine signaling pathway, among others. The detailed enrichment results are shown in Supplement 5.

Random Forest and LASSO Identification for Diagnostic Markers. A total of 402 DEGs were obtained from the gene expression matrix, and these 402 DEGs were used to construct a random forest. The top 2/3 nodes with the best classification effect were selected (keeping the largest MeanDecreaseGini and removing the top 1/3 nodes with small ones), and the poorly classified nodes were removed. We screened a total of six rounds and obtained 54 important genes. We then used the LASSO regression algorithm to identify seven diagnostic markers associated with UC, namely TLCD3A, KLF9, EFNA1, NAAA,WDR4, CKAP4, and CHRNA1, from the important genes obtained (Fig. 7A). The single gene ROC analysis using the expression values of the seven diagnostic markers in GSE38713 and GSE94648 revealed that the AUC values of all diagnostic markers in GSE38713 were greater than 0.9 (Fig. 7B,C). Good diagnostic values were also demonstrated in GSE94648 (Fig. 7D,E), with most genes having AUC values above 0.65. We then performed a hierarchical clustering analysis using these seven genes in GSE38713 and GSE94648 (Fig. 7H,I); the samples in both datasets were clustered into two categories, with one category clustering most of the normal samples and one category clustering most
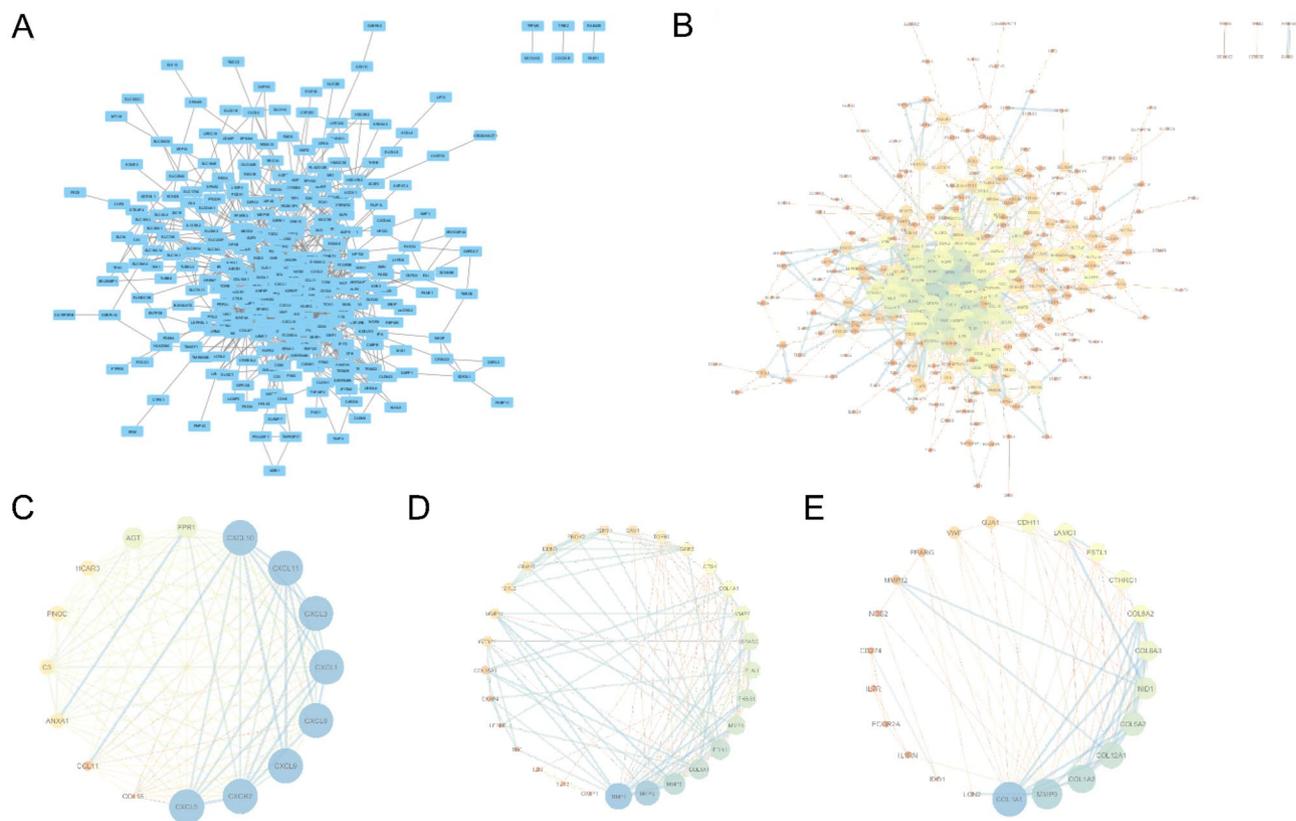
Figure 4. Protein–protein interaction (PPI) network analysis. (**A**) PPI network obtained from STRING database; (**B**) network analysis and visualization using Network Analyzer in Cytoscape; (**C–E**) the three highest rated word networks identified with the MCODE plugin, which were considered as the three more important functional modules.

of the disease samples. We then inserted these seven genes into GEPIA2 and found that NAAA and CHRNA1 had a significant effect on the survival prognosis of UC-associated CRC (Fig. 7F,G).

Immune Cell Infiltration Analysis and its Correlation with Diagnostic Markers. The results of interactions of 22 immune cells (Fig. 8A) showed that follicular helper T cells had the strongest interactions with other immune cells, while resting mast cells, monocytes, macrophages, M0, and T cells had weaker interactions with other immune cells. The results of the correlation heat map of 22 immune cells showed (Fig. 8B) that T cells CD4 memory resting, dendritic cells activated, neutrophil, M1 macrophages, T cells gamma delta, and mast cells resting showed a significant negative correlation with follicular helper T cells, activated mast cell, dendritic cells resting; monocytes showed a significant negative correlation with monocytes and showed a significant positive correlation with macrophages M0; activated dendritic cells showed a significant positive correlation with macrophages M0, neutrophils, and T cells follicular helper cells, whereas showed a significant negative correlation with T cells CD4 memory resting. The box line plot of immune cell infiltration differences (Fig. 8C) showed that macrophages M1, activated dendritic cells, and neutrophil cells infiltrated relatively more, while NK cells activated cells infiltrated relatively less, compared with normal control samples. The results of the correlation analysis (Fig. 8D) showed that immune cells were clustered into two categories: macrophages M2, B cells naïve, NK cells resting, T cells regulatory, NK cells activated, T cells CD4 memory resting, eosinophils, T cells CD8, and resting mast cells showed a significant positive correlation with CKAP4, TLCD3A, WDR4 and a negative correlation with KLF9, EFNA1, NAA, and CHRNA1, while the rest of the immune cells showed the opposite trend.

Discussion

Ulcerative colitis (UC) is a refractory disease characterized by a long duration, recurrence, and difficulty in healing[37]. The exact pathogenesis of this disease remains unknown. However, understanding the pathology of UC and underlying molecular mechanisms is essential for its clinical diagnosis and treatment. The use of efficient genome-wide gene expression microarray data and bioinformatics analysis can help us understand the molecular mechanisms of disease onset and progression, and is necessary for the identification of potential diagnostic biomarkers. To date, relevant reports have been published in terms of immune infiltration. Xiu et al.[9] by raw letter analysis has predicted central genes, namely CDC42, POLR2A, RAC1, PIK3R1, MAPK1, and SRC, that have important roles in the pathological differences between children and adults with UC as well as immune cells, namely B cells, T cells, monocytes, macrophages, and mast cells, which may be potential biomarkers for the diagnosis and treatment of UC. Xue et al.[10] showed that DPP10, S100P, AMPD1, and ASS1 may serve as
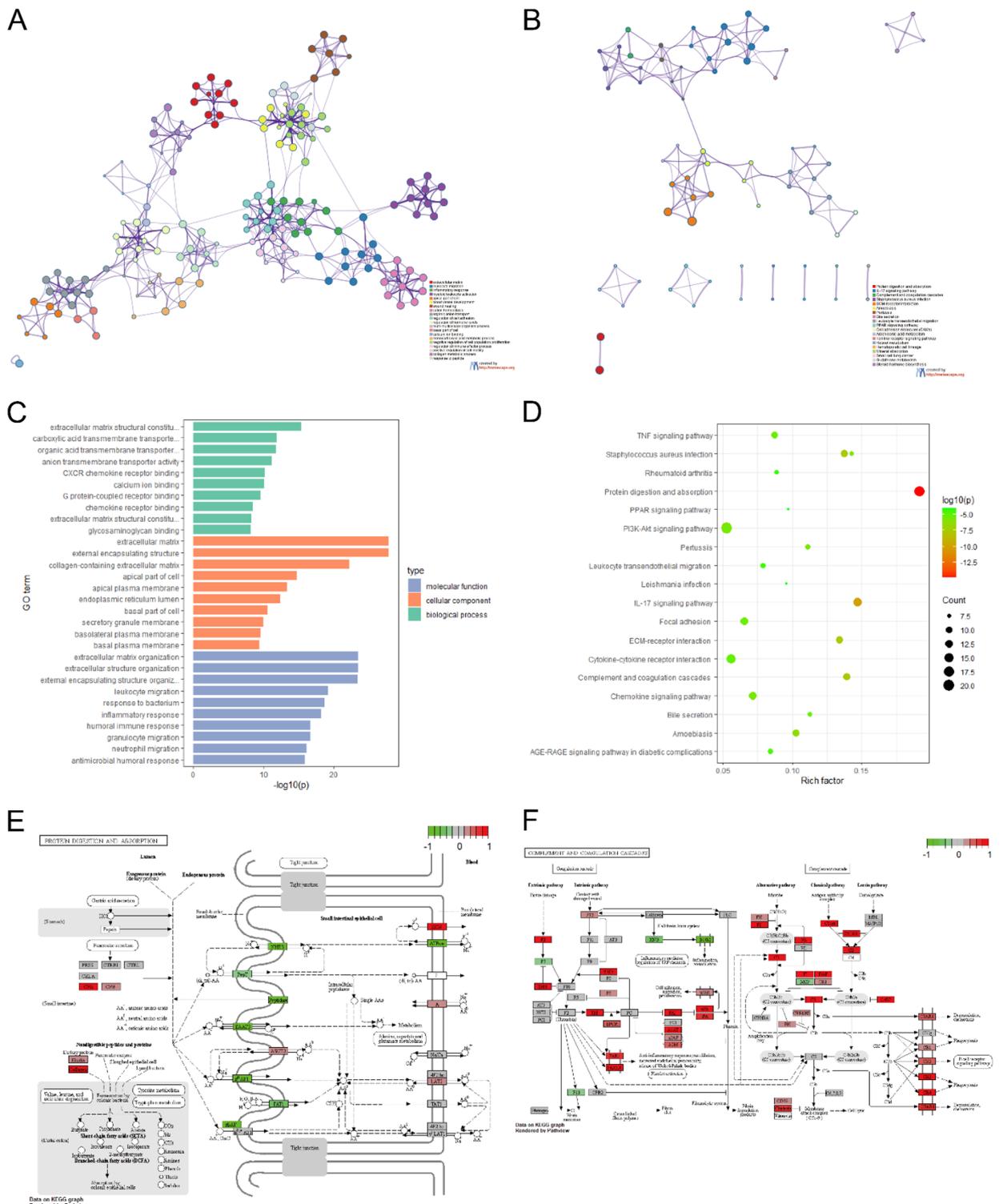
Figure 5. Functional enrichment analysis of differentially expressed genes. (**A**) Network diagram of top 20 GO enrichment functions, with cluster IDs to indicate color, each node is an enriched term; (**B**) network diagram of top 20 KEGG enrichment functions, with cluster IDs to indicate color; (**C**) barplot of GO enrichment functions, the length of the function bar is shown by *p* value; (**D**) dotplot of top 20 KEGG enrichment results; (**E**) pathway map of Protein digestion and absorption; (**F**) pathway map of Complement and coagulation cascades.

diagnostic biomarkers for UC and that differentially infiltrating immune cells may help indicate the progression of UC. Zhu et al.[38] revealed that immunity and infection are the two most important factors in the pathogenesis
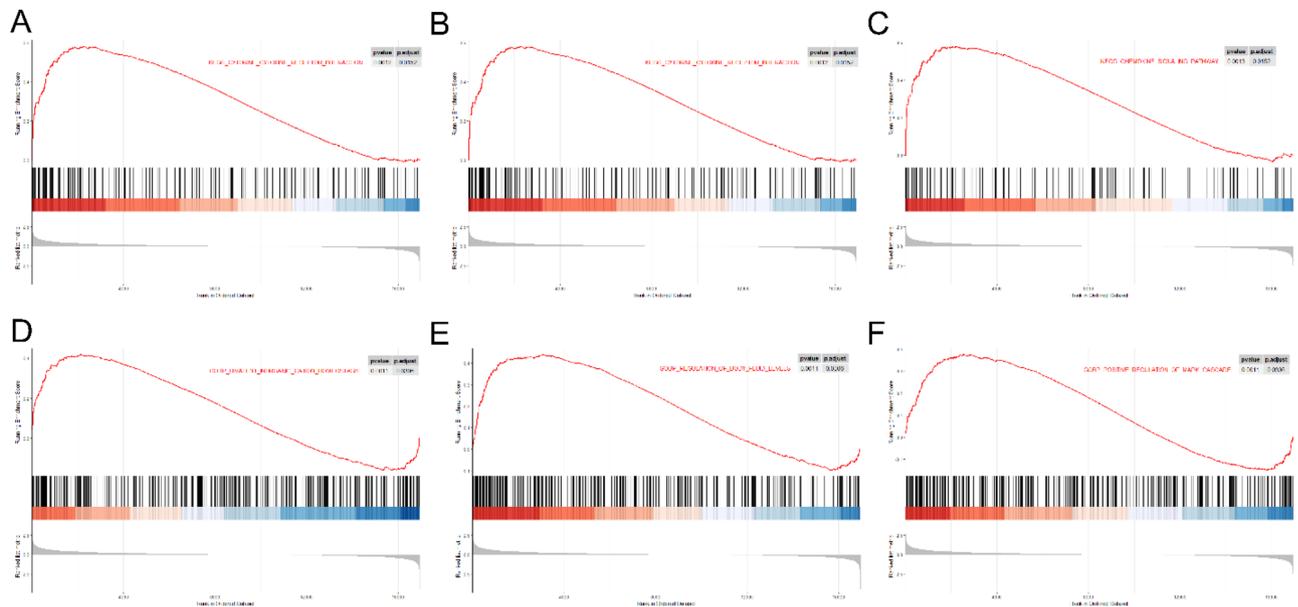
Figure 6. GSEA function enrichment analysis, (**A–C**) the top 3 GSEA enrich plots of enriched KEGG.D-F. The top 3 GSEA enrich plots of enriched GO. (**A**) KEGG_cytokine cytokine receptor interaction; (**B**) KEGG_focal adhesion; (**C**) KEGG_ chemokine signaling pathway. (**D**) GOBP_divalent inorganic cation homeostasis; (**E**) GOBP_regulation of body fluid levels; F. GOBP_ positive regulation of MAPK cascade.

of UC; in this study, we used microarray-based bioinformatics analysis to explore the gene expression profile and pathogenesis of UC. To avoid a high false positive rate and one-sided results, we selected two gene microarray datasets (GSE38713 and GSE94648) for comprehensive analysis. We screened differential genes by performing GO, KEGG, GESA, and PPI analysis; used random forest and LASSO regression to screen for diagnostic markers; and used CIBERSORT to screen for immune cells associated with UC.

We used the dataset GSE38713 as the training set and GSE94648 as the validation set, and identified 402 DEGs, 242 upregulated genes and 160 downregulated genes in the dataset GSE38713. PPI network analysis of the obtained differential genes yielded three important sub-networks, and we suggest that these three modules may play a special role in the pathogenesis of UC. The largest of these subnetworks is dominated by the chemokine family, including CXCL1, CXCL3, CXCL5, CXCL6, CXCL9, CXCL10, CXCL11, and the chemokine receptor CXCR2, a large class of peptides that play a key role in the regulation of inflammation[39]. Chemokines are classified into the four families C, CC, CXC, and CX3C, according to the number and arrangement of their N-terminal cysteine residues; the CC subfamily mainly recruits lymphocytes and dendritic cells, and the CXC subfamily mainly recruits neutrophils and monocytes. Chemokines can chemotacticize leukocytes to participate in immune and inflammatory responses[40,41]. Blocking chemokines or their receptors significantly reduced intestinal inflammation and mucosal damage in animals with UC, suggesting that chemokines play a key role in the pathogenesis of UC[42]. CXCL1 upregulates and recruits circulating white blood cells, allowing the inflammation cycle to continue. Relevant studies have shown that CXCL1 is significantly upregulated in the colon tissues of UC patients and rats, and may be a potential biomarker of UC tissue biopsy[39,43,44]. CXCL9 is a small cytokine called MIG, which serves as a chemoattractant for T cells induced by IFN-γ. Serum CXCL9 level is related to UC disease activity, and its expression increases in patients with UC and UC mouse models. Thus, it may become a marker of patient response to treatment[45,46]. Animal studies have reported that CXCL10 inhibits the proliferation of intestinal epithelial cells and regulates the proliferation of crypt cells during acute colitis in mice, making it a new therapeutic target for inflammatory bowel disease[47]. CXCR2 plays a key role in the pathogenesis of UC by regulating the immune response of neutrophils. Blocking CXCR2 can improve DSS-induced intestinal mucosal inflammation in mice, and CXCR2 can be used as a new target for UC drug therapy[48,49].

The second subnetwork is dominated by the matrix metalloproteinase family, including MMP1, MMP2, MMP3, MMP7, MMP10, and MMP9(in the third subnetwork), and its associated matrix metalloprotein inhibitor TIMP1. UC lesions are strongly associated with excessive ECM degradation. Matrix metalloproteinases (MMPs) can degrade proteins in the (ECM) and have an important role in ulceration and tissue remodeling[50,51]. Increased levels of MMP1 and MMP2 have been shown to play a major role in degradation of the intestinal matrix[52]. Both protein and mRNA levels of MMP-2 and MMP9 were significantly increased in inflammatory bowel disease tissues, with the highest expression levels in severely inflamed tissues[53]. MMP9 activates myosin light chain kinase (MLCK) to impair colonic epithelial permeability and plays an important role in enhancing the degree of inflammation[54]. Upregulation of intestinal mucosal MMP9 expression in patients with UC correlates with severity, and its increase suggests severe mucosal damage in active UC[55]. MMP3 is produced by mesenchymal cells and immune cells in the lamina propria. Some studies have reported serum MMP3 as a potential biomarker for endoscopic and histological activity of UC[56].The primary cellular source of MMP7 in patients with active UC is most likely leukocytes[57]. In UC, MMP7 was found to be expressed in epithelial cells at
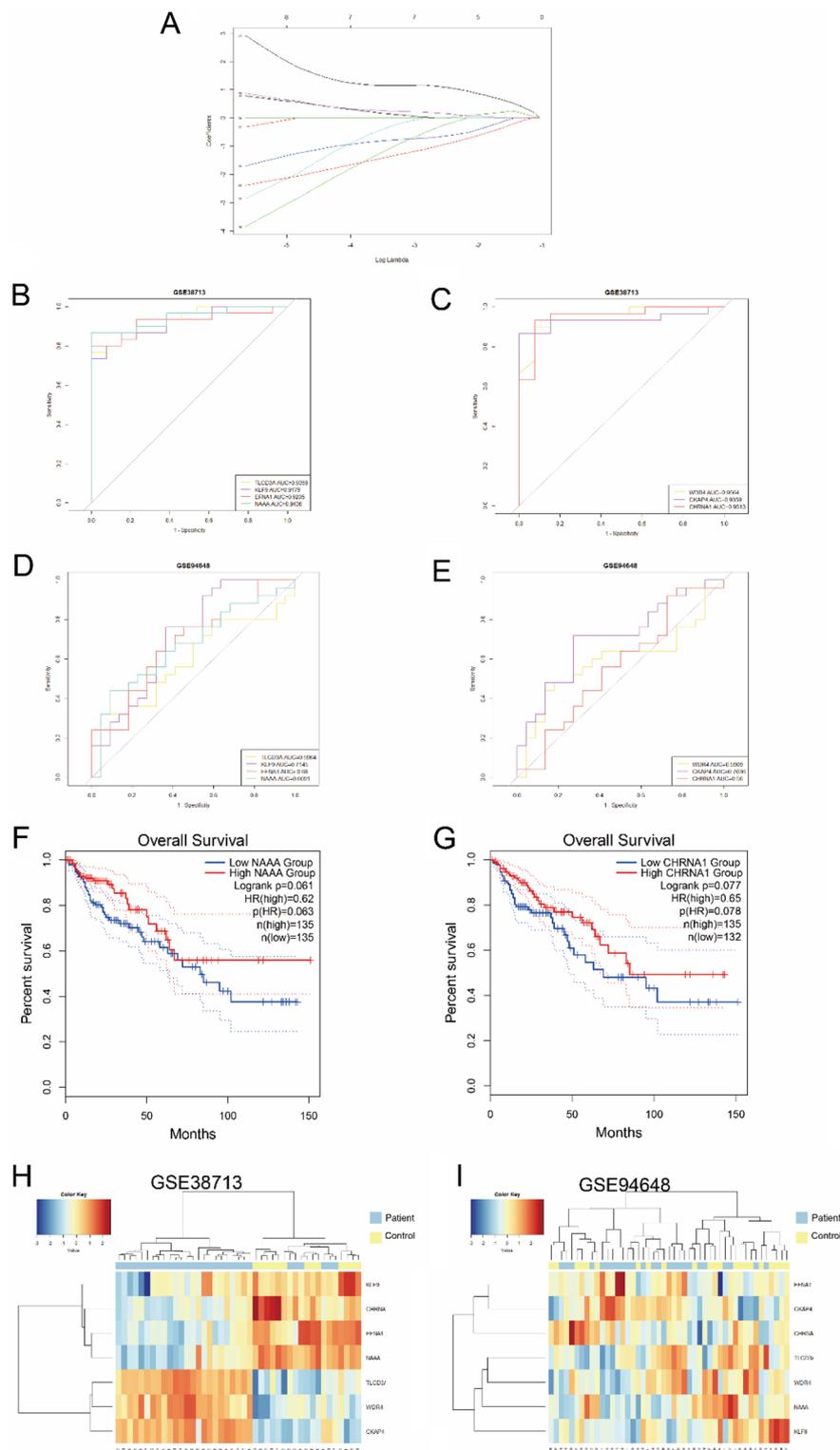
Figure 7. Identification, validation and prognostic analysis of diagnostic markers. (**A**) LASSO logistic regression algorithm for screening diagnostic markers; (**B–E**) ROC curves of diagnostic markers in GSE38713 and GSE94648; (**F,G**) GEPIA2 database showing the impact of NAAA and CHRNA1 on survival prognosis of UC-associated colorectal cancer; (**H,I**) Hierarchical clustering of diagnostic markers for GSE38713 and GSE94648. Normal samples are in yellow and disease samples are in blue.
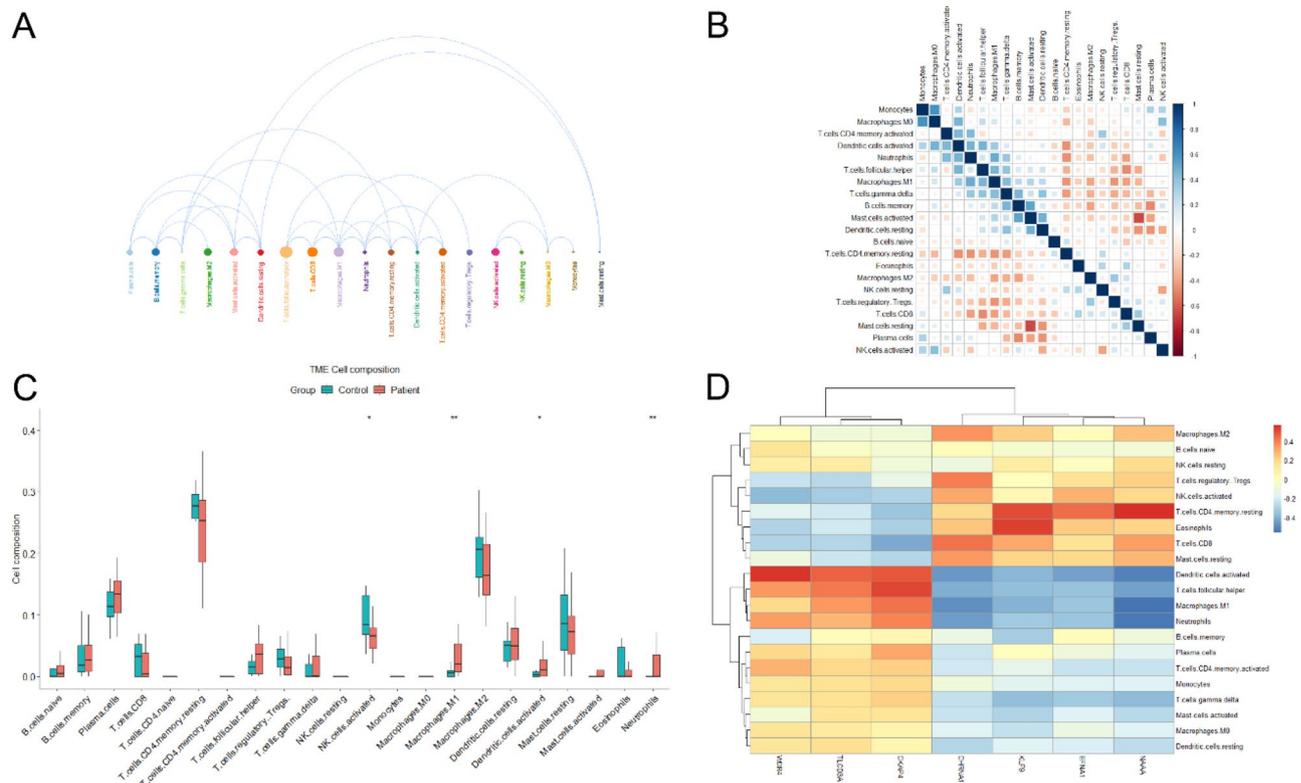
Figure 8. Visualization of immune cell infiltration and its correlation analysis with diagnostic markers. (**A**) Interaction plot of 22 immune cell infiltrations; circle size represents the strength of interactions with other immune cells, the larger the circle, the stronger the interactions with other immune cells. (**B**) Correlation heat map of 22 immune cell infiltrations; blue indicates positive correlation, red indicates negative correlation, the darker the color, the stronger the correlation. (**C**) Box line plot of 22 immune cell infiltrations, red represents UC group, blue represents control group. Box line plot of the proportion of 22 immune cell infiltrates; red represents UC group, blue represents Control group. (**D**) Correlation analysis of 22 immune cell infiltrates and diagnostic markers; red represents positive correlation, blue represents negative correlation. $*p < 0.05$; $**p < 0.01$.

the ulcer margin, developmental abnormalities, and transformed cells, and its expression was correlated with the degree of endoscopic inflammation[58]. Tissue inhibitors of metalloproteinases (TIMPs) are natural inhibitors of MMPs, which in turn are a group of secreted glycoproteins that are widely found in tissues and body[59]. TIMP-1 is one of the four isoforms of TIMPs and mainly inhibits MMP1, MMP3, and MMP9 activities[60]. MMP-1 mRNA, TIMP-1 mRNA, and MMP-1 mRNA/TIMP-1 mRNA ratio in the diseased colonic mucosa of patients with UC can be used as biomarkers to determine the severity of the patients' clinical symptoms[61]. As shown by GO and KEGG analysis of DEGs, in the BP annotation of GO, extracellular matrix organization, inflammatory response, humoral immune response; in the CC annotation of GO, extracellular matrix; and in the MF annotation of GO, extracellular matrix structural components and CXCR chemokine receptor binding were significantly associated with the occurrence and development of UC.

Inflammation is an important pathological response in UC pathogenesis. Pathologically, inflammation occurs in the lining of the colon and rectum, and is manifested by infiltration of neutrophils, macrophages, lymphocytes, and mast cells. Intestinal inflammation further destroys the mucosa and submucosa, eventually leading to intestinal ulceration[62]. ECM constitutes the framework structure for cell survival and affects the basic life activities of cells, and its components are in dynamic balance; imbalance will cause various pathological changes, such as ulcer formation[63]. Degradation of ECM is involved in the pathological development of UC, and quantitative changes in its component composition and structure play an important role in the pathogenesis of inflammatory bowel disease[64,65]. The pathogenesis of UC is related to immunological abnormalities, and various factors involved in the immune system may be directly or indirectly associated with UC[66]. Generally, immune responses are divided into cellular and humoral immune responses according to different effectors[67]. Chemokine receptors are a class of G protein-coupled receptors that play an important role in inflammatory cells of injured or infected organs. Chemokine receptor expression is upregulated during the active phase of UC[68]. In KEGG, among the important pathways of enrichment, complement and coagulation cascades, chemokine signaling pathway, IL-17 signaling pathway, and ECM-receptor interaction were significantly enriched, which is consistent with previous studies[44,69,70]. In our study, the PI3K-Akt signaling pathway was enriched with the highest number of genes. The PI3K/AKT signaling pathway is closely related to the regulation of cytokines and plays an important role in the process of intestinal inflammation, which can lead to dysregulation of the inflammatory response[71]. In UC, UC-associated colon carcinogenesis can be induced by upregulating the PI3K/Akt signaling pathway. When this pathway is blocked, the activation of nuclear factor kappa B (NF-κB) is inhibited, and cytokine release is reduced[72,73].

To investigate the biological functions of the DEGs associated with UC, GSEA was performed. The top3 enriched GO-BPs were divalent inorganic cation homeostasis, regulation of body fluid levels and positive regulation of MAPK cascade, respectively. Cytokine cytokine receptor interaction, focal adhesion, and chemokine signaling pathways were the first three significantly enriched KEGG pathways. The enrichment of cytokine-cytokine receptor interaction and chemokine signaling pathway were consistent with the results of a previous study[44].

The seven diagnostic markers screened using random forest screen and LASSO regression were TLCD3A, KLF9, EFNA1, NAAA, WDR4, CKAP4, and CHRNA1. Among them, only NAAA has a small number of reports in the literature on its association with UC. A decreased number of NAAA-positive immune cells detected in active UC has been reported[74]. NAAA-targeted drugs have potential value in the treatment of human inflammatory diseases[75].

From the analysis of immune cell infiltration assessment and its correlation with diagnostic markers, it is clear that among 22 immune cells, follicular helper T cells have the strongest interactions with other immune cells. The immune balance between follicular helper T (TFH) cells and follicular regulatory T (TFR) cells is important for regulating B-cell responses, and changes in the ratio between the two, shifts the balance from immune tolerance to an immune response state, leading to B-cell immune dysregulation and the pathogenesis of UC[76]. It has been shown that increased inducible co-stimulation positive (ICOS) + programmed cell death 1 positive (PD-1) + TFH cells are associated with B-cell activation in UC pathogenesis and may act as potential biomarkers for UC disease monitoring[77]. In the differential analysis of immune cell infiltration, macrophages M1 ($p < 0.01$), activated dendritic cells ($p < 0.05$), and neutrophils ($p < 0.01$) were highly expressed in UC tissues compared to normal tissues. Neutrophils are predominantly present in areas of colonic mucosal injury in patients with UC, forming their characteristic crypt abscesses, producing reactive oxygen species and releasing serine proteases, matrix metalloproteinases, and myeloperoxidase (MPO)[78,79]. Macrophages are the main effector cells of the innate immune system and play various roles, such as phagocytosis of pathogens, secretion of cytokines and chemokines, and antigen presentation. They are divided into M1 macrophages (classically called activated macrophages) and M2 macrophages (alternatively called activated macrophages). M1 macrophages are more frequently present in the lamina propria of the colonic mucosa of UC and produce large amounts of pro-inflammatory cytokines. Their abnormal activation is an important part of UC development[80–83]. Activated and mature dendritic cells may play a role in inducing an immune response that is exacerbated in UC, and their increased function may be related to the inflammatory mucosal environment found in patients with UC[84].

Our study has certain limitations. First, the sample size was increased to further clarify the diagnostic accuracy of the core genes associated with UC. Second, the results of the two microarrays as training and validation sets, respectively, may be more one-sided, and external validation is needed to avoid false-positive rates; third, further ex vivo experiments are needed to validate the potential mechanisms by which the obtained important gene modules act on UC.

Conclusions

In conclusion, the aim of this study was to explore the molecular mechanisms underlying UC pathogenesis through bioinformatics analysis. We aimed to identify the relevant biological functions and signaling pathways involved in the development of UC. We identified three functional modules that play an important role in the development of UC occurrence through PPI network analysis. Seven genes were identified by LASSO regression as potential diagnostic markers for UC, and the area under the curve for most genes was greater than 0.65 was estimated by ROC curve analysis and further by GEPIA2 analysis. NAAA and CHRNA1 were predicted to also serve as prognostic markers for survival in UC-associated CRC. The relationship between immune cell infiltration and seven diagnostic markers was also analyzed by CIBERSORT, and positive relationships were obtained between Macrophages M2, NK cells resting, T cell regulatory and CKAP4, TLCD3A, WDR4. In contrast, these immune cells were inversely correlated with KLF9, EFNA1, NAA, and CHRNA1; however, further experiments are required to validate the current findings.

Data availability

Our data can be found in the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/, GSE38713 and GSE94648) database.

References

1. Eaden, J. A., Abrams, K. R. & Mayberry, J. F. The risk of colorectal cancer in ulcerative colitis: A meta-analysis. *Gut* **48**, 526–535 (2001).
2. Lai, L. J., Shen, J. & Ran, Z. H. Natural killer T cells and ulcerative colitis. *Cell Immunol.* **335**, 1–5 (2019).
3. Hibi, T. & Ogata, H. Novel pathophysiological concepts of inflammatory bowel disease. *J. Gastroenterol.* **41**, 10–16 (2006).
4. Neurath, M. F. Targeting immune cell circuits and trafficking in inflammatory bowel disease. *Nat. Immunol.* **20**, 970–979 (2019).
5. de Souza, H. S. & Fiocchi, C. Immunopathogenesis of IBD: Current state of the art. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 13–27 (2016).
6. Kobayashi, T. *et al.* Ulcerative colitis. *Nat. Rev. Dis. Primers* **6**, 74 (2020).
7. Tatiya-Aphiradee, N., Chatuphonprasert, W. & Jarukamjorn, K. Immune response and inflammatory pathway of ulcerative colitis. *J. Basic Clin. Physiol. Pharmacol.* **30**, 1–10 (2018).
8. Lan, K. *et al.* A survey of data mining and deep learning in bioinformatics. *J. Med. Syst.* **42**, 139 (2018).
9. Xiu, M. X., Liu, Y. M., Chen, G. Y., Hu, C. & Kuang, B. H. Identifying hub genes, key pathways and immune cell infiltration characteristics in pediatric and adult ulcerative colitis by integrated bioinformatic analysis. *Dig. Dis. Sci.* **66**, 3002–3014 (2021).

10. Xue, G., Hua, L., Zhou, N. & Li, J. Characteristics of immune cell infiltration and associated diagnostic biomarkers in ulcerative colitis: Results from bioinformatics analysis. *Bioengineered* **12**, 252–265 (2021).

11. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol.* **1711**, 243–259 (2018).

12. Davis, S. & Meltzer, P. S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).

13. Planell, N. *et al.* Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. *Gut* **62**, 967–976 (2013).

14. Planell, N. *et al.* Usefulness of transcriptional blood biomarkers as a non-invasive surrogate marker of mucosal healing and endoscopic response in ulcerative colitis. *J. Crohns Colitis* **11**, 1335–1346 (2017).

15. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. Affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).

16. Braschi, B. *et al.* Genenames.org: The HGNC and VGNC resources in 2019. *Nucleic Acids Res.* **47**, D786–D792 (2019).

17. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

18. Ito, K. & Murphy, D. Application of ggplot2 to pharmacometric graphics. *CPT Pharmacomet. Syst. Pharmacol.* **2**, e79 (2013).

19. Szklarczyk, D. *et al.* The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res* **45**, D362–D368 (2017).

20. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

21. Chen, S. *et al.* Identification of crucial genes in abdominal aortic aneurysm by WGCNA. *Peer J* **7**, e7873 (2019).

22. Gene Ontology Consortium. Gene Ontology Consortium: Going forward. *Nucleic Acids Res.* **43**(Database issue), D1049–D1056 (2015).

23. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

24. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).

25. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).

26. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

27. Luo, W. & Cory, B. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**(14), 1830–1831 (2013).

28. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: A desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251–3253 (2007).

29. 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 (2012).

30. Zhou, M. *et al.* Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* **12**, 518–529 (2018).

31. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **23**, 18–22 (2002).

32. Tang, Z., Kang, B., Li, C., Chen, T. & Zhang, Z. GEPIA2: An enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* **47**, W556–W560 (2019).

33. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

34. Pesenti, C. *et al.* The genetic landscape of human glioblastoma and matched primary cancer stem cells reveals intratumour similarity and intertumour heterogeneity. *Stem Cells Int.* **2019**, 2617030 (2019).

35. Sasaki, K. *et al.* Clarifying the structure of serious head and spine injury in youth Rugby Union players. *PLoS ONE* **15**, e0235035 (2020).

36. Cheng, Q. & Wang, L. LncRNA XIST serves as a ceRNA to regulate the expression of ASF1A, BRWD1M, and PFKFB2 in kidney transplant acute kidney injury via sponging hsa-miR-212-3p and hsa-miR-122-5p. *Cell Cycle* **19**, 290–299 (2020).

37. Daperno, M. *et al.* Unmet medical needs in the management of ulcerative colitis: Results of an Italian Delphi consensus. *Gastroenterol. Res. Pract.* **2019**, 3108025 (2019).

38. Zhu, J., Wang, Z., Chen, F. & Liu, C. Identification of genes and functional coexpression modules closely related to ulcerative colitis by gene datasets analysis. *Peer J* **7**, 8061 (2019).

39. Egesten, A. *et al.* The proinflammatory CXC-chemokines GRO-α/CXCL1 and MIG/CXCL9 are concomitantly expressed in ulcerative colitis and decrease during treatment with topical corticosteroids. *Int. J. Colorectal Dis.* **22**, 1421–1427 (2007).

40. Zlotnik, A. & Yoshie, O. Chemokines: A new classification system and their role in immunity. *Immunity* **12**, 121–127 (2000).

41. Rossi, D. & Zlotnik, A. The biology of chemokines and their receptors. *Annu. Rev. Immunol.* **18**, 217–242 (2000).

42. Xia, X. M. *et al.* CXCR4 antagonist AMD3100 attenuates colonic damage in mice with experimental colitis. *World J. Gastroenterol.* **16**, 2873–2880 (2010).

43. Boshagh, M. A., Foroutan, P., Moloudi, M. R., Fakhari, S. & Jalili, A. ELR positive CXCL chemokines are highly expressed in an animal model of ulcerative colitis. *J. Inflamm. Res.* **12**, 167–174 (2019).

44. Zhang, J. *et al.* Investigation of potential genetic biomarkers and molecular mechanism of ulcerative colitis utilizing bioinformatics analysis. *Biomed. Res. Int.* **2020**, 4921387 (2020).

45. Elia, G. & Guglielmi, G. CXCL9 chemokine in ulcerative colitis. *Clin. Ter.* **169**, e235–e241 (2018).

46. Lacher, M. *et al.* Association of a CXCL9 polymorphism with pediatric Crohn's disease. *Biochem. Biophys. Res. Commun.* **363**, 701–707 (2007).

47. Sasaki, S. *et al.* Blockade of CXCL10 protects mice from acute colitis and enhances crypt cell survival. *Eur. J. Immunol.* **32**, 3197–3205 (2002).

48. Zhu, F. *et al.* Blockade of CXCR2 suppresses proinflammatory activities of neutrophils in ulcerative colitis. *Am. J. Transl. Res.* **12**, 5237–5251 (2020).

49. Buanne, P. *et al.* Crucial pathophysiological role of CXCR2 in experimental ulcerative colitis in mice. *J. Leukoc. Biol.* **82**, 1239–1246 (2007).

50. Yang, K. *et al.* Visualization of protease activity in vivo using an activatable photo-acoustic imaging probe based on CuS nanoparticles. *Theranostics* **4**, 134–141 (2014).

51. Garg, P. *et al.* Notch1 regulates the effects of matrix metalloproteinase-9 on colitis-associated cancer in mice. *Gastroenterology* **141**, 1381–1392 (2011).

52. Stallmach, A. *et al.* Comparression of matrix metalloproteinases 1 and 2 in pouchitis and ulcerative colitis. *Gut* **47**, 415–422 (2000).

53. Gao, Q. *et al.* Expression of matrix metalloproteinases-2 and -9 in intestinal tissue of patients with inflammatory bowel diseases. *Dig. Liver Dis.* **37**, 584–592 (2005).

54. Neurath, M. Current and emerging therapeutic targets for IBD. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 688 (2017).

55. Lakatos, G. *et al.* The behavior of matrix metalloproteinase-9 in lymphocytic colitis, collagenous colitis and ulcerative colitis. *Pathol. Oncol. Res.* **18**, 85–91 (2012).

56. Kourkoulis, P. *et al.* Leucine-rich alpha-2 glycoprotein 1, high mobility group box 1, matrix metalloproteinase 3 and annexin A1 as biomarkers of ulcerative colitis endoscopic and histological activity. *Eur. J. Gastroenterol. Hepatol.* **32**, 1106–1115 (2020).
57. Rath, T. *et al.* Cellular sources of MMP-7, MMP-13 and MMP-28 in ulcerative colitis. *Scand. J. Gastroenterol.* **45**, 1186–1196 (2010).
58. Matsuno, K. *et al.* The expression of matrix metalloproteinase matrilysin indicates the degree of inflammation in ulcerative colitis. *J. Gastroenterol.* **38**, 348–354 (2003).
59. von Lampe, B., Barthel, B., Coupland, S. E., Riecken, E. O. & Rosewicz, S. Differential expression of matrix metalloproteinases and their tissue inhibitors in colon mucosa of patients with inflammatory bowel disease. *Gut* **47**, 63–73 (2000).
60. Wang, Y. D., Tan, X. Y. & Zhang, K. Correlation of plasma MMP-1 and TIMP-1 levels and the colonic mucosa expressions in patients with ulcerative colitis. *Mediat. Inflamm.* **2009**, 275072 (2009).
61. Wang, Y. D. & Yan, P. Y. Expression of matrix metalloproteinase-1 and tissue inhibitor of metalloproteinase-1 in ulcerative colitis. *World J. Gastroenterol.* **12**, 6050–6053 (2006).
62. Ge, H. *et al.* Rhein attenuates inflammation through inhibition of NF-κB and NALP3 inflammasome in vivo and in vitro. *Drug Des. Dev. Ther.* **11**, 1663–1671 (2017).
63. Shi, L., Ramsay, S., Ermis, R. & Carson, D. In vitro and in vivo studies on matrix metalloproteinases interacting with small intestine submucosa wound matrix. *Int. Wound J.* **9**, 44–53 (2012).
64. Kirov, S. *et al.* Degradation of the extracellular matrix is part of the pathology of ulcerative colitis. *Mol. Omics* **15**, 67–76 (2019).
65. Derkacz, A., Olczyk, P., Jura-Półtorak, A., Olczyk, K. & Komosinska-Vassev, K. The diagnostic usefulness of circulating profile of extracellular matrix components: Sulfated Glycosaminoglycans (sGAG), Hyaluronan (HA) and extracellular part of Syndecan-1 (sCD138) in patients with Crohn's disease and ulcerative colitis. *J. Clin. Med.* **10**, 1722 (2021).
66. Han, Y. *et al.* Role of moxibustion in inflammatory responses during treatment of rat ulcerative colitis. *World J. Gastroenterol.* **20**, 11297–11304 (2014).
67. Zhang, L. X., Zhao, L. F., Zhang, A. S., Chen, X. G. & Xu, C. S. Expression patterns and action analysis of genes associated with physiological responses during rat liver regeneration: Cellular immune response. *World J. Gastroenterol.* **12**, 7514–7521 (2006).
68. Lin, X. *et al.* Functional characterization of CXCR4 in mediating the expression of protein C system in experimental ulcerative colitis. *Am. J. Transl. Res.* **9**, 4821–4835 (2017).
69. Cao, F., Cheng, Y. S., Yu, L., Xu, Y. Y. & Wang, Y. Bioinformatics analysis of differentially expressed genes and protein-protein interaction networks associated with functional pathways in ulcerative colitis. *Med. Sci. Monit.* **27**, e927917 (2021).
70. Song, R. *et al.* Identification and analysis of key genes associated with ulcerative colitis based on DNA microarray data. *Medicine (Baltimore)* **97**, e10658 (2018).
71. Wei, J. & Feng, J. Signaling pathways associated with inflammatory bowel disease. *Recent Pat. Inflamm. Allergy Drug Discov.* **4**, 105–117 (2010).
72. Setia, S., Nehru, B. & Sanyal, S. N. Upregulation of MAPK/Erk and PI3K/Akt pathways in ulcerative colitis-associated colon cancer. *Biomed. Pharmacother.* **68**, 1023–1029 (2014).
73. Huang, X. L. *et al.* PI3K/Akt signaling pathway is involved in the pathogenesis of ulcerative colitis. *Inflamm. Res.* **60**, 727–734 (2011).
74. Suárez, J. *et al.* Ulcerative colitis impairs the acylethanolamide-based anti-inflammatory system reversal by 5-aminosalicylic acid and glucocorticoids. *PLoS ONE* **7**, e37729 (2012).
75. Piomelli, D. *et al.* N-Acylethanolamine Acid Amidase (NAAA): Structure, function, and inhibition. *J. Med. Chem.* **63**, 7475–7490 (2020).
76. 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 (2017).
77. Long, Y., Zhao, X., Liu, C., Xia, C. & Liu, C. Activated inducible co-stimulator-positive programmed cell death 1-positive follicular helper T cells indicate disease activity and severity in ulcerative colitis patients. *Clin. Exp. Immunol.* **202**, 106–118 (2020).
78. Dinallo, V. *et al.* Neutrophil extracellular traps sustain inflammatory signals in ulcerative colitis. *J. Crohns Colitis* **13**, 772–784 (2019).
79. Muthas, D. *et al.* Neutrophils in ulcerative colitis: A review of selected biomarkers and their potential therapeutic implications. *Scand. J. Gastroenterol.* **52**, 125–135 (2017).
80. Nie, M. F. *et al.* Serum and ectopic endometrium from women with endometriosis modulate macrophage M1/M2 polarization via the Smad2/Smad3 pathway. *J. Immunol. Res.* **2018**, 6285813 (2018).
81. Ham, M. *et al.* Macrophage glucose-6-phosphate dehydrogenase stimulates proinflammatory responses with oxidative stress. *Mol. Cell. Biol.* **33**, 2425–2435 (2013).
82. Magnusson, M. K. *et al.* Macrophage and dendritic cell subsets in IBD: ALDH+ cells are reduced in colon tissue of patients with ulcerative colitis regardless of inflammation. *Mucosal Immunol.* **9**, 171–182 (2016).
83. Wong, W. Y. *et al.* Proteomic profiling of dextran sulfate sodium induced acute ulcerative colitis mice serum exosomes and their immunomodulatory impact on macrophages. *Proteomics* **16**, 1131–1145 (2016).
84. Ikeda, Y., Akbar, F., Matsui, H. & Onji, M. Characterization of antigen-presenting dendritic cells in the peripheral blood and colonic mucosa of patients with ulcerative colitis. *Eur. J. Gastroenterol. Hepatol.* **13**, 841–850 (2001).

Author contributions

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-33388-5.

Correspondence and requests for materials should be addressed to Q.C.