Construction of molecular subgroups of ulcerative colitis

J.-L. MA[1], H.-J. ZHANG[2], C.-F. ZHANG[3], Y.-Y. ZHANG[1], G.-M. WANG[1,4]

[1]School of Clinical Medicine, Dali University, Dali, China
[2]Department of Obstetrics, The Affiliated Yantai Yuhuangding Hospital of Qingdao University, Yantai, China
[3]Department of Neurology, The Affiliated Yantai Yuhuangding Hospital of Qingdao University, Yantai, China
[4]Genetic Testing Center, The First Affiliated Hospital of Dali University, Dali, China

Abstract. – OBJECTIVE: Ulcerative colitis (UC), a chronic inflammatory disease of the colon with unknown etiology, is characterized by remission and recurrence. At present, a considerable number of UC cases are misdiagnosed or delayed in diagnosis and treatment. We aimed to identify UC-related genes to aid the development of drugs for this condition.

PATIENTS AND METHODS: Transcriptome data of 362 patients with UC and 126 control subjects were obtained from the Gene Expression Omnibus. The 362 patients with UC were subgrouped using unsupervised machine learning. R software was used to analyze the clinical characteristics of the subgroups, screen subgroup-specific genes, assess the relationships between gene modules and clinical characteristics using weighted gene co-expression network analysis, and perform Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses of the subgroups.

RESULTS: Patients with UC were classified into two subgroups. Genes specific to subgroup I included *IL21R, ATP8B2,* and *PLEKHO1*. Severe disease tended to be associated with immune cell infiltration; anti-tumor necrosis factor (TNF)-α antibodies and ustekinumab may have been effective in this subgroup. Subgroup II-specific genes included *SLC4A4, EPB41L4B,* and *PLCE1*. Patients in this subgroup had mild clinical conditions; however, their disease was more likely to progress to colorectal cancer. Thus, 5-aminosalicylic acid-based drugs may be effective for the treatment of UC in these patients.

CONCLUSIONS: We divided UC into two molecular subgroups based on transcriptome data, providing molecular evidence for the development of diagnostic methods and individualized treatment strategies for UC.

Key Words:
Ulcerative colitis, Bioinformatics, Subgroup, Gene.

Introduction

Ulcerative colitis (UC) is a chronic inflammatory disease of the colon with unknown etiology. The disease is characterized by chronic inflammation of the intestinal mucosa that starts from the rectum and can affect the entire colon, with alternating remission and recurrence[1]. Immune, genetic, environmental, dietary, and other factors have been associated with disease occurrence[2]. UC is a global disease; in North America and northern Europe, the estimated prevalence rate is 90-505 cases per 100,000 people[3]. From 1990 to 2003, the prevalence of UC in China was approximately 11.6 per 100,000 people, and the incidence rate showed an upward trend[4]. UC can occur at any age, particularly in late adolescence or early adulthood[5]. Although medical treatments have made significant progress, a considerable number of cases are misdiagnosed or delayed in diagnosis and treatment. Reportedly, 22.9% of patients with UC are misdiagnosed as having other diseases before admission[4]. The risk of colon cancer is high in patients with UC. The overall prevalence of colorectal cancer in patients with UC is 3-7%[6].

Correct scientific classification and targeted treatment are key to effectively managing UC. Clinically, disease activity is classified as mild (purulent and bloody stool <4 times per day), moderate (purulent and bloody stool 4-6 times per day), and severe (purulent and bloody stool >6 times per day as well as poisoning symptoms, such as fever and tachycardia). In recent years, with the development of molecular medicine, the management of moderate to severe UC has transitioned from the hospital setting to the patient's home. The treatment methods include subcutaneous administration of biological agents and oral drugs[7,8]. Among the UC drugs that have

Corresponding Authors: Guangming Wang, Ph.D; e-mail: wgm1991@dali.edu.cn;
Yuanyuan Zhang, Ph.D; e-mail: zhyy@dali.edu.cn

been recently approved or are awaiting approval, the anti-interleukin (IL)-23 antibody seems to be the safest, with the fewest adverse events[9]. Although considerable progress has been made in drug development, there remains an urgent need for effective biomarkers to guide rational treatment choices. To obtain more effective treatment outcomes for UC, further progress must be made in the management beyond drug development.

In this study, we aimed to identify UC pathogenesis-related genes to aid in drug development. We used unsupervised machine learning and statistical analysis to identify molecular subgroups of healthy and patient cohorts and validate their feasibility through genetic analysis to develop more effective treatments for UC.

Patients and Methods

Data Sources and Processing

Original UC-related array expression profiling datasets (GSE11223, GSE13367, GSE53306, GSE87466, and GSE212849) and corresponding platform files (GPL1708, GPL570, GPL14951, GPL13158, and GPL570) were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). The chip datasets contain information on patients with UC and healthy control subjects. We used the 'limma' (v.3.54.2) and 'sva' (v.3.46.0) packages in R software (v.4.2.2; R Foundation, Vienna, Austria) to remove batch effects and obtain a single, unified dataset via principal component analysis (PCA)[10,11].

Establishment of Molecular Subgroups

The 'ConsensusClusterPlus' (v.1.62.0) package in R (v.4.2.2) was used for class discovery. The gene expression matrix of 362 patients with UC was subjected to cluster analysis using unsupervised machine learning[12]. The unsupervised class-learning method can identify groups of similar data points, defined as clusters, from a dataset and visually represent them as heatmaps. The gene expression matrix cluster numbers were set to 2-10. Cases with similar genetic characteristics were clustered. Finally, we selected the 'best cluster' with a high cluster consistency score and higher clustering density within the group as the grouping.

Analysis of Subgroup Clinical Characteristics

We collated the clinical characteristics of the normalized dataset, including age, smoking status, lesion range (extensive or localized), and the UC scoring system (UCSS) score. Using age and the UCSS score as continuous variables, the values were transferred to SPSS software (v.26.0; IBM Corp., Armonk, NY, USA), and the mean and standard deviation were obtained through Chi-square testing. Then, the 'rstatix', 'ggplot2', and 'ggpubr' packages in R (v.4.2.2) were used to perform box graph analysis on the age and UCSS score. The smoking and lesion ranges were then used as discrete variables for histogram analysis. Finally, a comparison between groups was carried out through a one-way analysis of variance to obtain the clinical characteristics of the subgroups. Patients who had previously smoked were classified in the smoking group.

Subgroup-Specific Gene Screening

The 'limma' R (v.4.2.2) package was used to compare the subgroups with each other and with the control group to identify significantly differentially expressed genes (DEGs) based on a mean filter >0.2 and corrected p<0.05. Only DEGs upregulated in each subgroup were obtained and defined as genes specific to the subgroup.

Gene set enrichment analysis of the subgroups

To identify DEG sets that were enriched for annotated biological functions, we used gene set enrichment analysis (GSEA). Perl software (v.5.32.1; https://www.perl.org/) was used to convert the comparison results between each subgroup and the control group before GSEA and to obtain gene list files and gene set files for each subgroup. The files were then transferred into GSEA software (v.4.1.0; https://www.gsea-msigdb.org/gsea/msigdb/) by subgroup to analyze the different genes in each subgroup. As the collated dataset was large, the maximum running value was set to 5,000.

Weighted Gene Co-Expression Network Analysis

We used the 'WGCNA' (v.1.71) R package (v.4.2.2) for weighted gene co-expression network (WGCN) construction and analysis (WGCNA) using the datasets. WGCNA clusters genes based on high expression similarity and the co-expression modules can then be correlated with clinical characteristics[13]. First, the genes were clustered to assess whether there were significant outliers. Then, a co-expression network was constructed based on the standard scale-free network analysis function, and the power function 'pickSoftThreshold' was used to calculate the soft threshold power β, which enhances strong correlation and reduces weak correlation at the expression level. The co-expression gene modules were identified using the dynamic tree-cutting method, and a gene tree map was generated according to different degrees of topological overlap measure[13]. Genes with similar expression levels were grouped into differently colored modules. Finally, Pearson's correlation was used to analyze the relationships between the modules and age, smoking status, UCSS score, and lesion range.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Enrichment Analyses

We used the 'clusterProfiler' (v.4.6.2), 'org.Hs.eg.db' (v.3.16.0), and 'enrichplot' (v.1.18.3) packages in R (v.4.2.2) for Gene Ontology (GO) term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses in the obtained colored gene modules[14]. The GO categories were biological process (BP), molecular function (MF), and cellular component (CC). The most significantly enriched pathway in each gene module was selected from the KEGG results, and the relationships between the pathway and the subgroups were analyzed.

Statistical Analysis

Continuous variables were analyzed using SPSS and are reported as mean ± standard deviation. Then, a single-factor analysis of variance was used for comparison between the two groups. Next, R was used to statistically analyze the data from the GEO database. Results at $p<0.05$ were considered significant in a two-tailed significance test.

Results

Processing of Transcriptome Data and Establishment of Subgroups

To obtain a normalized, unified dataset, batch effects were removed from the individual datasets GSE11223, GSE13367, GSE53306, GSE87466, and GSE212849. The main components of the processed dataset were clustered and visualized. In the PCA plot of the datasets before pre-processing for batch effects, the datasets were clearly distinct (Figure 1A). After pre-processing, all samples were scattered and could be regarded as a single dataset (Figure 1B). In unsupervised learning, the maximum cluster number was set to 10. After the analysis, nine clusters were obtained, and the consistency of each subgroup was evaluated. When the data were classified into two subgroups, the consistency evaluation score of each subgroup was close to 1.0, indicating that the clustering within each group was good (Figure 1C). To verify the accuracy of the clustering results, we generated a consensus matrix plot. This plot confirmed that when the data were classified into two subgroups, the correlation within each subgroup was high (close to 1.0), but that between groups was low (Figure 1D).

Subgroup Clinical Characteristics

A total of 362 UC and 126 control samples were included in the grouping analysis (**Supplementary Table I** for detailed clinical information). Analysis of clinical characteristics, namely, smoking status, lesion range, age, and the UCSS score, in each subgroup revealed a slightly higher number of smoking patients in subgroup II than in subgroup I, although the difference was not significant (Figure 2A). The proportion of patients with extensive lesions was higher in subgroup I than in subgroup II (Figure 2B), whereas the proportion of patients with localized lesions was lower in subgroup I than in subgroup II (Figure 2C). The mean age was 65.79±14.941 years in subgroup I and 69.75±15.731 years in subgroup II, revealing no significant difference (Figure 2D). The mean UCSS scores in subgroups I and II were 4.33±3.270 and 2.04±2.182, respectively, which differed significantly ($p<0.001$; Figure 2E).

Subgroup-Specific Gene Identification

To identify specific genes in each subgroup, we screened DEGs that were specifically upregulated in each subgroup by comparing the subgroups with each other and the control group. There were 1,798 specific genes in subgroup I and 1,660 in subgroup II. Table I lists the top 10 specific genes in each subgroup. A Venn diagram plotted to determine the accuracy of the screening results revealed no genes in the intersection (Figure 3), indicating that these genes met the screening criteria.

Subgroup GSEA

To analyze whether the specific genes in the two subgroups and the DEGs between the subgroups and control samples were consistent, we conducted GSEA. The results showed that the compared DEGs were concentrated on the left side of the image with a p-value and error detection rate <0.01 (Figure 4 A-B). This indicated that the uniqueness of the DEGs between the subgroups and of those between the subgroups and control samples was consistent.

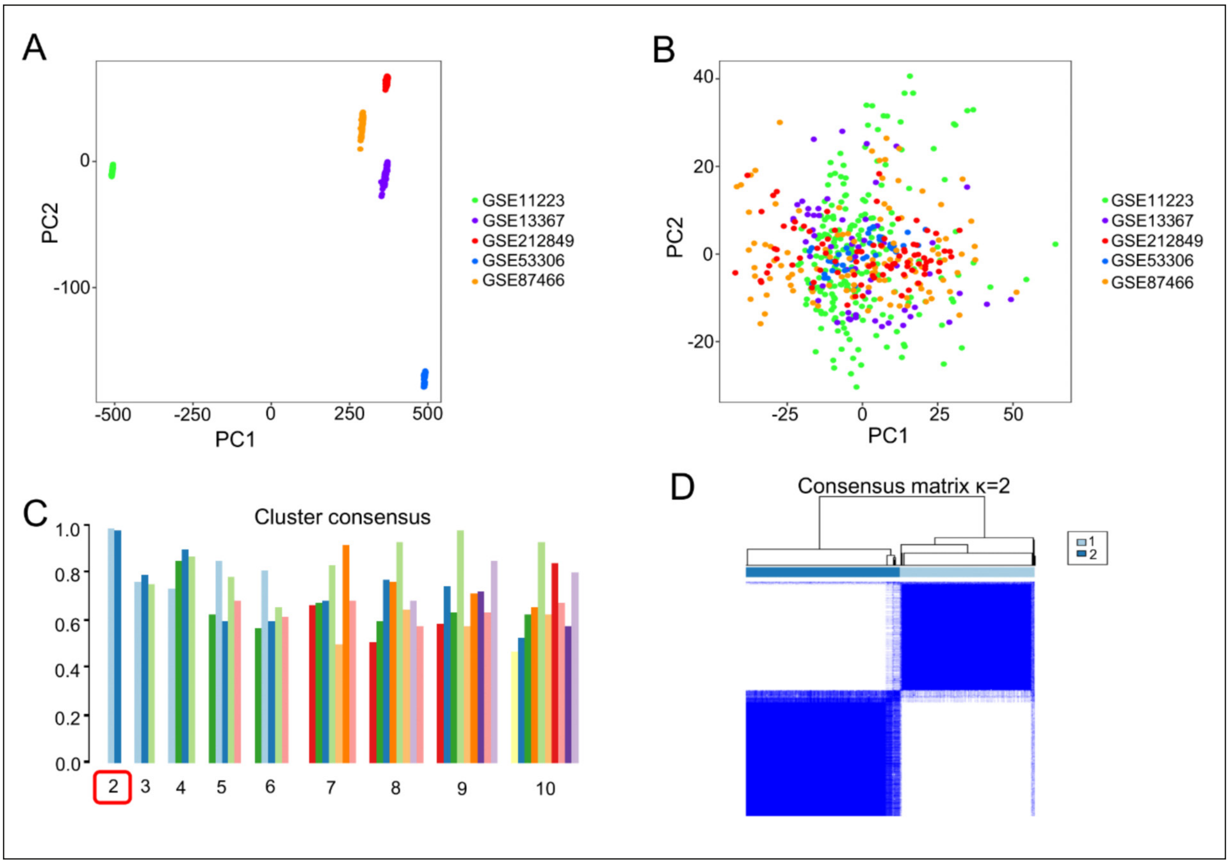

Figure 1. Normalization and cluster analysis of the datasets. **A**, Principal component analysis (PCA) plot of the datasets before pre-processing for batch effects, showing no intersection among the datasets. **B**, PCA plot of the datasets after pre-processing for batch effects, showing a more distributed dataset. **C**, Consistency evaluation scores of the clusters. The values below the chart represent the number of clusters, and each color represents a subgroup. When the number of groups is 2, the consistency score is highest. **D**, Consensus matrix plot showing that when the samples were classified into two subgroups, the darker the blue rectangle within the group, the higher the genetic correlation. PC, principal component.

WGCNA of Patients With UC

We performed WGCNA on 12,587 standardized genes after removing batch effects and found that the gene connectivity conformed to a scale-free network with a soft threshold power β of 10 (Figure 5A). A WGCN was constructed according to the determined genes. Three colored gene modules (Figure 5B) were obtained using the dynamic tree-cutting method: blue (1,594 genes), brown (1,216 genes), and grey (648 genes). The relationships between the modules and clinical characteristics were determined. Age was significantly negatively correlated with the brown module (*p*=0.04) and significantly positively correlated with the grey module (*p*=0.005). The UCSS score was negatively correlated with the blue module (*p*=2e-13) and positively correlated with the brown (*p*=7e-15) and grey (*p*=5e-8) modules. Smoking was positively correlated with the blue module (*p*=0.008) and negatively correlated with the grey module (*p*=0.006). Extensive lesions were significantly negatively correlated with the blue module (*p*=0.01) and significantly positively correlated with the brown (*p*=0.02) and grey (*p*=0.001) modules (Figure 5C), similar to the

Table I. List of the top 10 specifically upregulated genes in both subgroups.

Subgroup I	Subgroup II
IL21R	*SLC4A4*
ATP8B2	*EPB41L4B*
PLEKHO1	*PLCE1*
CMTM7	*SAMD13*
ADA	*XK*
CHI3L2	*PANK3*
HCLS1	*ENTPD5*
GMFG	*OCLN*
SH3BP5	*GPD1L*
FXYD5	*PXMP2*

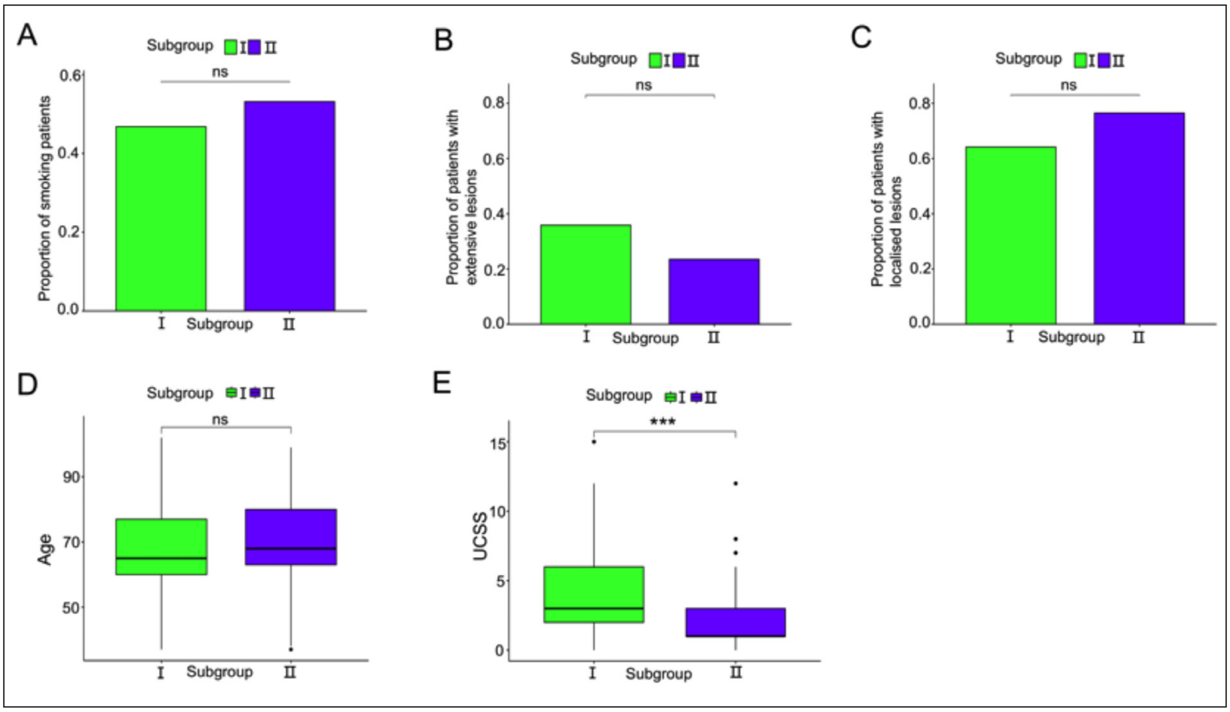

Figure 2. Analysis of the clinical characteristics in the subgroups. Comparison of smoking status (**A**), presence of extensive lesions (**B**), presence of localized lesions (**C**), age (**D**), and the UCSS score between the two subgroups (**E**). Single factor analysis of variance shows: *p<0.05, **p<0.01, ***p<0.001, ns=not significant. UCSS, ulcerative colitis scoring system.

UCSS score. Through correlation analysis of the gene modules and subgroups, we found that the expression of the blue module was lower, whereas that of the brown module was higher in subgroup I than in the control group. The results of subgroup II contrasted with those of subgroup I (Figure 6).

GO and KEGG Enrichment Analyses

To analyze the biological characteristics of the gene modules, we conducted enrichment analyses. The GO analysis results showed that the BPs ‘fatty acid catalytic process’ and ‘cellular lipid catalytic process’ were enriched in the blue module, whereas ‘regulation of leukocyte cell-cell adhesion’, ‘regulation of T cell activation’, and ‘regulation of protein secret’ were enriched in the brown module (Figure 7A). The CCs ‘mitochondrial matrix’ and ‘plasma membrane’ were enriched in the blue module (Figure 7B). The MFs ‘electric transfer activity’, ‘oxidoreductase activity’, and ‘phosphoric ester hydrolase activity’ were enriched in the blue module, and ‘extracellular matrix structural constraint’ was enriched in the brown module (Figure 7C). The KEGG pathway enrichment analysis showed that the ‘peroxisome’ and ‘peroxisome proliferator-activated receptor (PPAR) signalling’ pathways were enriched in the blue module, and ‘Th17 cell differentiation’, ‘B cell receiver signalling pathway’, and ‘nuclear factor (NF)-kappa B signalling pathway’ were enriched in the brown module (Figure 7D). To further understand the correlations between KEGG pathway enrichment and the subgroups, we selected the pathway that was the most significantly enriched in each gene module and analyzed the expression of each pathway in each subgroup. The ‘peroxisome’ pathway showed low expression in subgroup I but high expression in

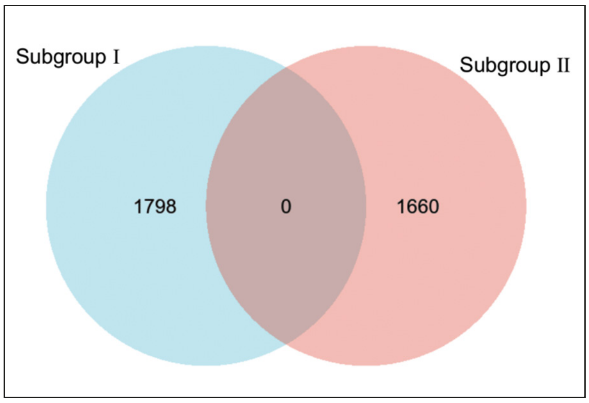

Figure 3. Venn diagram showing the numbers of subgroup-specific genes.

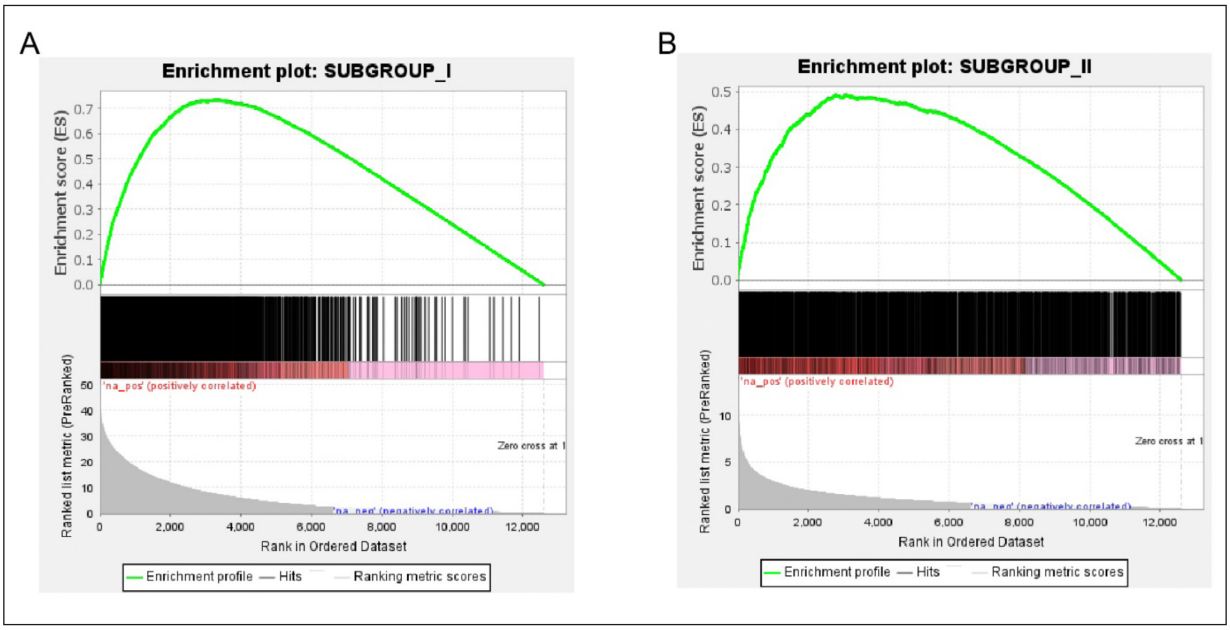

Figure 4. GSEA of the two subgroups. **A**, GSEA of significant DEGs in subgroup I. **B**, GSEA of significant DEGs in subgroup II. The green line represents the gene enrichment scores, each black line represents a gene, and the grey area represents the signal-to-noise ratio between each subgroup and the control group. GSEA, gene set enrichment analysis; DEGs, differentially expressed genes.

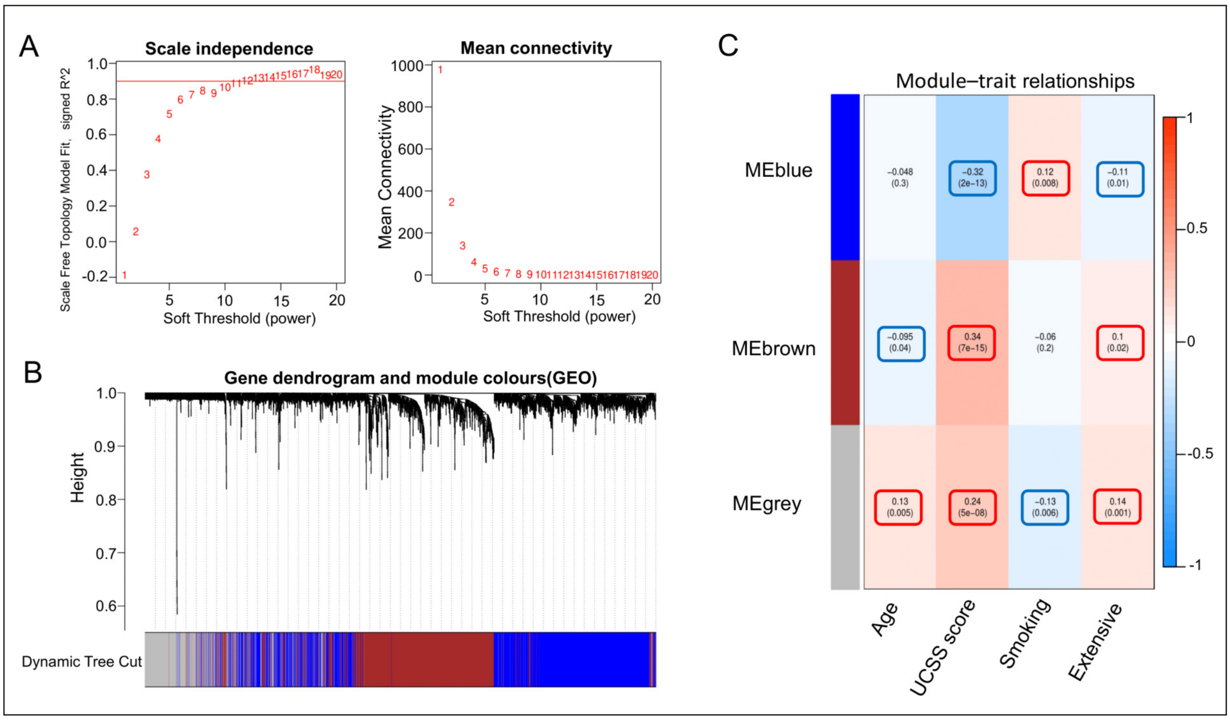

Figure 5. Weighted gene co-expression network analysis of the datasets. **A**, Left panel: the relationship between the power value and scale-free fitting index; right panel: the relationship between the power value and average connectivity. The red line represents the optimal power value. **B**, Colored gene module tree. Each black line represents a gene and corresponds to different colored modules. **C**, Relationships between the modules and clinical characteristics (age, UCSS score, smoking, and presence of extensive lesions) through Pearson correlation analysis.

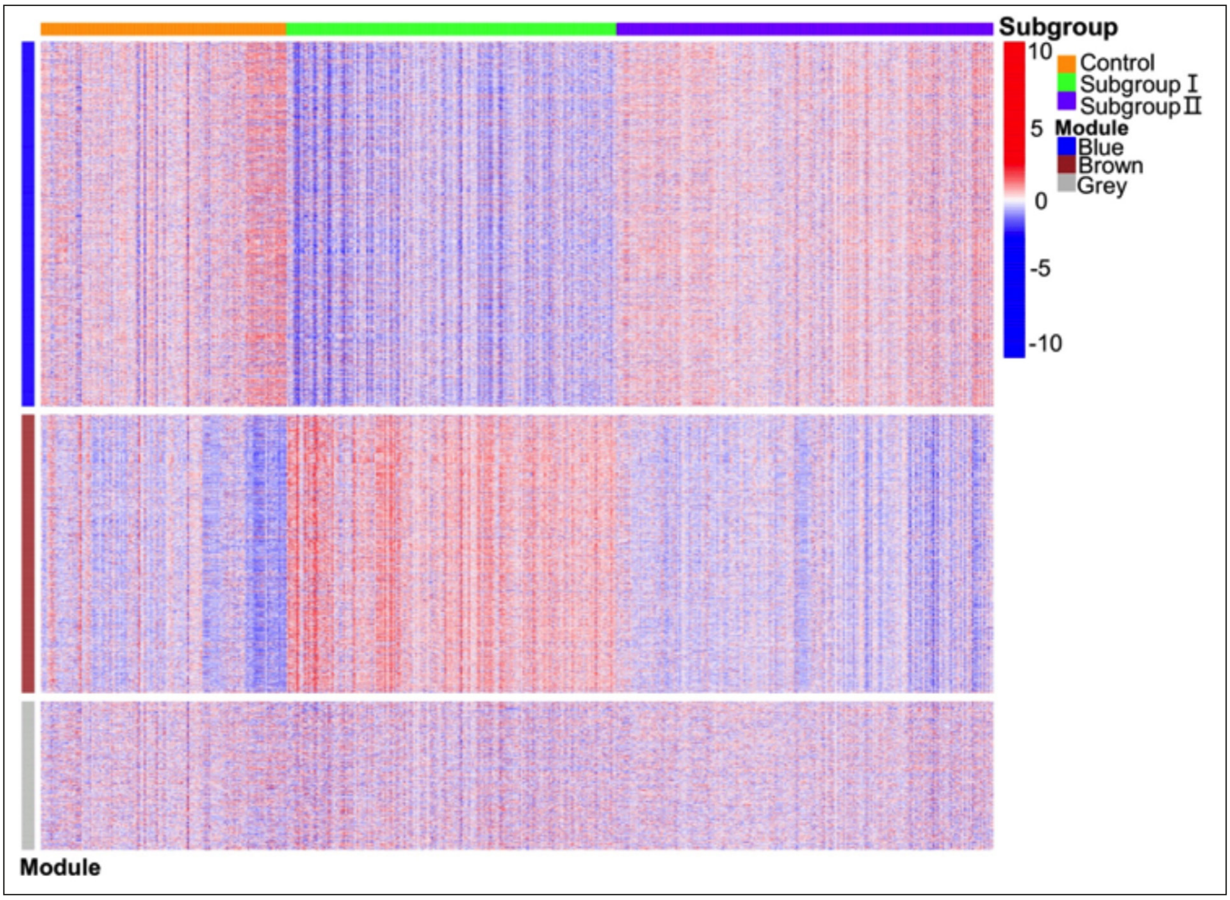

Figure 6. Heatmap of the relationships between the colored gene modules and subgroups. Red represents a positive correlation, and blue represents a negative correlation.

the control group and subgroup II. 'Hematopoietic cell lineage' was highly expressed in subgroup I but had low expression in the control group and subgroup II. There was no significant difference in the expression of cytokine-cytokine receptor interaction among the groups (Figure 8).

Summary of Subgroup Characteristics

To better understand the clinical, molecular, and biological characteristics of each subgroup, we summarized the characteristics of each subgroup based on the above results in Table II. We found that the characteristics of each subgroup differed and could be used to classify UC.

Discussion

With the development of molecular medicine, more disease-related genes have received increasing attention. The information from these genes can provide new methods for developing treatments. To the best of our knowledge, this is the first study to classify the transcriptome data of 362 patients with UC using bioinformatics. With recent progress in our understanding of the genetics of UC, it has become increasingly important to construct suitable genetic UC subtypes[15,16]. In our study, UC was classified into two genetic subgroups. Through analysis of the clinical characteristics of the subgroups, we found that disease lesions were more extensive and localized disease was less prevalent in subgroup I than in subgroup II, indicating that the patients in subgroup I had more severe disease. The UCSS score is based on a combination of stool frequency, rectal bleeding, endoscopic activity, and an overall rating scale assessed by the physician (0-3 for each item, representing inactive to serious disease). The total UCSS score for severe disease is 12 points[17]. The UCSS score in subgroup I (4.33±3.270) was significantly higher than that in subgroup II (2.04±2.182), further indicating that the symptoms were more severe in subgroup I. Furthermore,

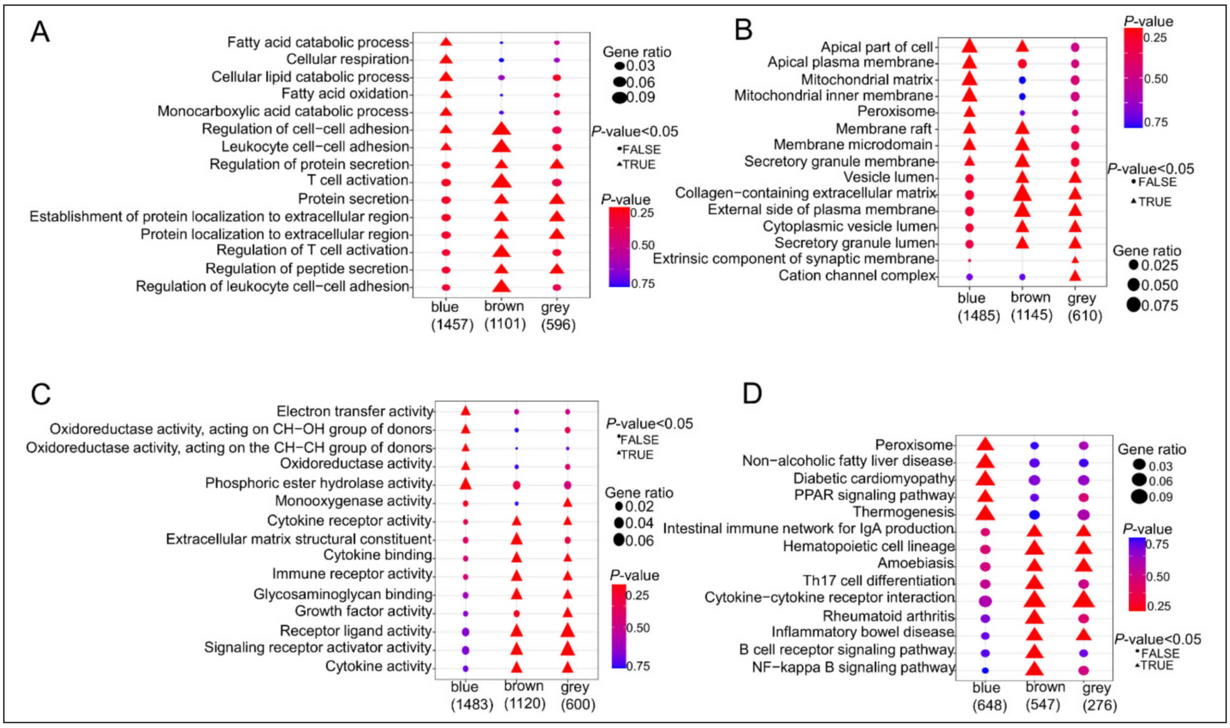

Figure 7. GO and KEGG pathway enrichment analyses of the gene modules. **A**, BP analysis results. **B**, CC analysis results. **C**, MF analysis results. **D**, KEGG pathway enrichment results. The 15 most significantly enriched terms are shown for each analysis. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological process; MF, molecular function; CC, cellular component.

Table II. Summary of both subgroups.

Category	Characteristics
Subgroup I	
Clinical features	UCSS score (High), Disease extent (extensive)
Specific gene	*IL21R, ATP8B2, PLEKHO1, CMTM7, ADA, CHI3L2, HCLS1, GMFG, SH3BP5, FXYD5*
BP	Regulation of protein secretion, T cell activation, Protein secretion, Protein localisation to extracellular region, Regulation of leukocyte cell–cell adhesion
CC	Vesicle lumen, Collagen-containing extracellular matrix, External side of plasma membrane, Cytoplasmic vesicle lumen, Extrinsic component of synaptic membrane
MF	Cytokine receptor activity, Extracellular matrix structural constituent, Cytokine binding, Immune receptor activity, Glycosaminoglycan binding, Receptor ligand activity, Signalling receptor activator activity, Cytokine activity
KEGG	Intestinal immune network for IgA production, Haematopoietic cell lineage, Amoebiasis, Th17 cell differentiation, Cytokine–cytokine receptor interaction, Rheumatoid arthritis, Inflammatory bowel disease, B cell receptor signalling pathway, NF-kappa B signalling pathway
Effective drugs	Anti-TNF-α antibodies, Ustekinumab
Subgroup II	
Clinical features	UCSS score (Low), Disease extent (limited), Colorectal cancer predisposition
Specific gene	*SLC4A4, EPB41L4B, PLCE1, SAMD13, XK, PANK3, ENTPD5, OCLN, GPD1L, PXMP2*
BP	Fatty acid catabolic process, Cellular respiration, Monocarboxylic acid catabolic process
CC	Apical part of cell, Apical plasma membrane, Mitochondrial matrix, Mitochondrial inner membrane, Peroxisome
MF	Electron transfer activity, Oxidoreductase activity, Phosphoric ester hydrolase activity
KEGG	Peroxisome, Non-alcoholic fatty liver disease, Diabetic cardiomyopathy, PPAR signalling pathway, Thermogenesis
Effective drugs	5-ASA

BP: biological process; CC: cellular component; KEGG: Kyoto Encyclopedia of Genes and Genomes; MF: molecular function; PPAR: proliferator-activated receptor; TNF: tumor necrosis factor; UCSS: ulcerative colitis scoring system; 5-ASA: 5-aminosalicylic acid; Ig, immunoglobulin.

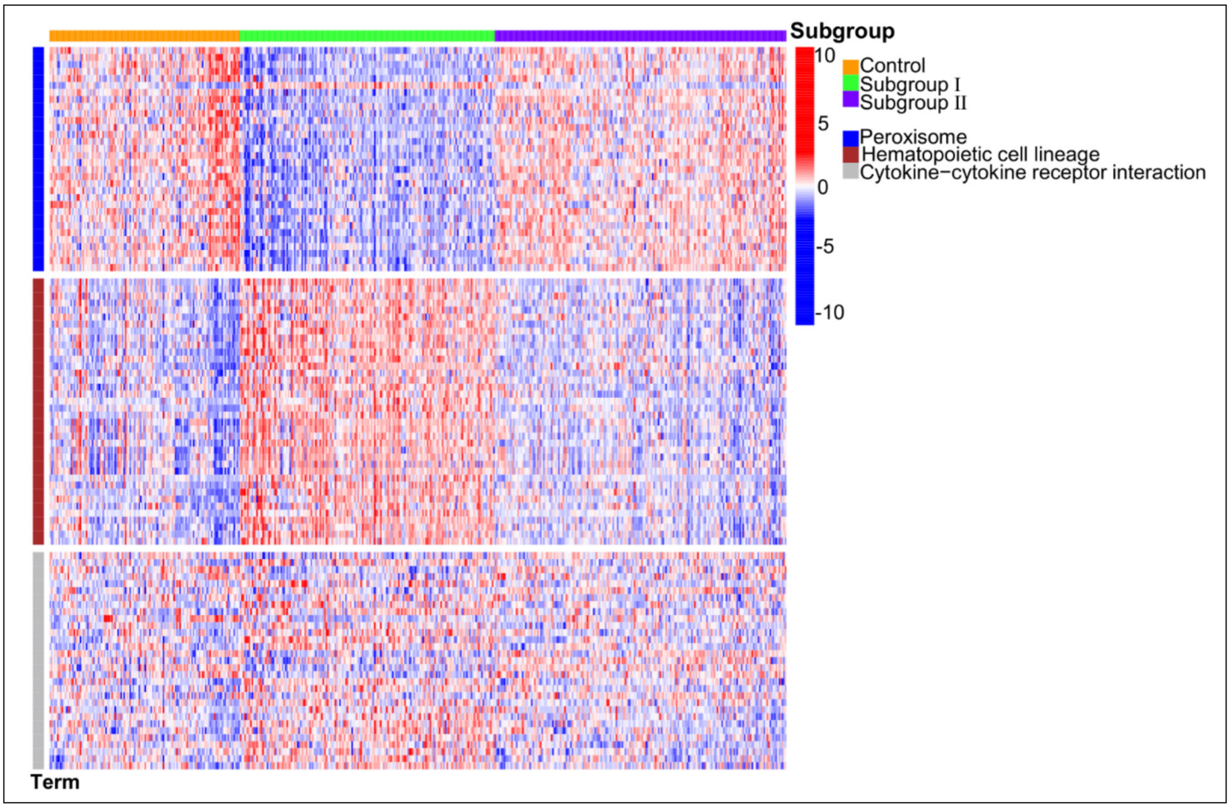

Figure 8. Heatmap showing the relationships between KEGG pathways and subgroups. KEGG, Kyoto Encyclopedia of Genes and Genomes.

there were more smokers in subgroup II than in subgroup I. Smoking is a risk factor for UC[18]. However, UC attack, drug dose, and risk of hospitalization and colon resection are similar in smokers and non-smokers with UC, and smoking cessation does not affect disease severity[19].

The *IL-21* gene (*IL-21*) comprises nine exons, spanning approximately 20 kb of genomic DNA on human chromosome 16p11[20]. IL-21 is a cytokine produced by follicular helper T cells that regulates innate and adaptive immune responses[21]. This protein exerts its effects by binding to a heterodimeric receptor consisting of the class I cytokine family receptor IL-21R and the common γ chain receptor[22]. IL-21 regulates T, B, and natural killer lymphocyte growth and functional activity[21]. Furthermore, IL-21 controls the activity of intestinal epithelial cells and fibroblasts and is regarded as an important mediator of communication between immune cells and non-immune cells[23]. It plays a key role in various inflammatory bowel disease (IBD) conditions[24,25]. Sarra et al[25] found that *IL-21* expression is enhanced in IBD and that IL-21 is mainly produced by CD4$^+$ T intestinal lamina propria lymphocytes co-expressing Interferon-gamma (IFNγ). In UC, *IL-21* expression is also increased[26]. In wild-type mice treated with sodium dextran sulphate to induce UC, the use of anti-*IL-21* reduced colonic invasion of follicular helper T cells and mucosal damage[27]. The significant difference in *IL-21* expression in subgroup I suggests that subgroup I is related to immune infiltration.

The GO analysis of subgroup I resulted in terms related to T cell activation and cytokine activity. In UC, large numbers of T lymphocytes that are activated and secrete various cytokines infiltrate and aggregate in inflammatory tissues and play a major role in disease pathogenesis[28]. *In vivo* and *in vitro* studies[29] have shown that the T cell response in UC is dominated by type 2 T helper cells that produce IL-4, IL-5, and IL-13. Tissue damage results from the interaction with cytokines and cells, including antigen-presenting cells, T helper cells, regulatory T cells, and natural killer T cells. These cells regulate and inhibit inflammation and play an extremely important role in UC[30].

In the pathological process of UC, several inflammatory cell types produce inflammatory cytokines. Ample experimental and clinical evidence[31] shows that the onset and progression of UC involve a multifactorial process mainly manifested in immune response disorders and epithelial barrier defects. When the microfold cells in the intestinal epithelium recognize exogenous antigens, they activate immune cells (e.g., B, T, and dendritic cells) in the Peyer's patches. This results in antigen-antibody reactions and eventually leads to the progressive injury of the intestinal endothelium. Based on research on molecularly targeted drugs, immune cells and soluble ligands (such as cytokines) are considered important targets for the treatment of UC[32]. For example, in UC, tumor necrosis factor (TNF)-α induces proinflammatory effects, including necrosis, activation of macrophages and T cells, and injury of epithelial cells[33]. Therefore, anti-TNF-α antibodies play an important role in treating inflamed mucosa. IL-23 is a heterodimeric cytokine produced mainly by inflammatory myeloid cells[34]. IL-12 shares the p40 subunit with IL-23 and can stimulate the differentiation of naïve CD4$^+$ T cells into Th1 cells. IL-23 mainly induces naïve CD4$^+$ T cells to differentiate into Th17 cells[35]. The pivotal role of Th-17 cells and their pathways in the development of UC has been confirmed[36]. Therefore, blocking IL-12 and IL-23 can effectively treat chronic intestinal inflammation in UC. Ustekinumab is a human IgG1κ monoclonal antibody that specifically binds to the shared p40 protein subunits of IL-12 and IL-23. Blocking the binding of the IL-12 receptor (IL-12R) on the surface of immune cells to p40 inhibits the biological activity of human IL-12 and IL-23, thereby inhibiting the production of Th1- and Th17-related cytokines. This has been reported[37] as an effective treatment for patients with moderate to severe UC. Therefore, it is inferred that anti-TNF-α antibodies and ustekinumab could be effective for patients in subgroup I.

The solid carrier family 4 member 4 (*SLC4A4*) gene encodes a Na$^+$/HCO$_3$-cotransporter and participates in the secretion and absorption of sodium bicarbonate to maintain the acid-base balance in the body[38]. *SLC4A4* plays a role in tumor onset and development. In colorectal cancer, *SLC4A4* expression is significantly downregulated, which is associated with poor prognosis, lymph node invasion, and distant metastasis[39]. Through the overexpression of *SLC4A4*, it has been found[40] that high levels of *SLC4A4* significantly inhibit colorectal cancer cell proliferation and migration and promote cancer cell apoptosis. Therefore, *SLC4A4* is considered a potential biomarker of colorectal cancer. As *SLC4A4* is a highly abundant bicarbonate transporter in cancer cells, the accumulation of bicarbonate in the extracellular space and reduction in glycolysis leads to a reduction in lactic acid production, which can alleviate acidosis[38]. The overall risk of colorectal cancer in patients with UC is approximately 7% 20 years after UC onset and 30% after 35 years[41]. In our study, *SLC4A4* was the most significant DEG among subgroup II-specific genes, indicating that patients with UC in subgroup II may have a higher risk of colorectal cancer than those in subgroup I.

5-aminosalicylic acid (5-ASA) is a peroxisome proliferator-activated receptor γ (PPAR-γ) agonist. In 1962, it was found[42] that sulfasalazine, composed of sulfapyridine and 5-ASA linked via an azo bond, could be used to treat mild to moderate UC. To date, it is widely used as the first-line drug for UC treatment. In mice with dextran sulphate-treated UC, 5-ASA activates PPAR-γ signalling in the intestinal epithelial cells to improve colitis[43]. In addition to improving inflammation, 5-ASA reduces the number of *Escherichia coli* cells in the fecal flora of patients with UC and inhibits the growth of this bacterium[44]. According to the GO and KEGG analyses, peroxisome, PPAR-γ signalling, and oxidoreductase activity were closely related to subgroup II, indicating that 5-ASA may be effective in patients in subgroup II.

Limitations

This study has some limitations. First, the results of this study need to be verified through *in vivo* and *in vitro* experiments. Second, the collected datasets do not contain clinical information such as patient sex and ethnicity. Future research should aim to provide a more comprehensive analysis of the datasets. Third, this study only focused on transcriptomics, which needs to be combined with genomics methods, such as metabolomics and proteomics, to gain a more comprehensive understanding of the molecular mechanism underlying UC. These shortcomings should be addressed in future studies.

Conclusions

We used transcriptome data to classify patients with UC into two subgroups. The specific genes in subgroup I were *IL21R, ATP8B2,* and *PLEKHO1.* Severe disease tended to be associated with immune cell infiltration. The genes specific to subgroup II were *SLC4A4*, *EPB41L4B*, and *PLCE1.* The clinical conditions in this group were mild, but patients in this group may be more likely to develop colorectal cancer, and 5-ASA may be effective for the treatment of UC in this group. Our results provide molecular evidence for the development of diagnostic methods and individualized treatment of UC.

Acknowledgments

We thank the contributors of all the datasets uploaded to the GEO platform.

Funding

This work was supported by the Yunnan Provincial Natural Science Foundation (grant number 202001BA070001-133), Medical Discipline Leader of the Yunnan Provincial Commission of Health and Family Planning (grant number D-2017057), and Key Construction Discipline of The First Affiliated Hospital of Dali University. These funders had no role in the design, conduct, or reporting of this work.

Ethics Approval

The data used in this study were sourced from the GEO public database. The contributors to the database have obtained ethical approval. Our research has no ethical issues.

Informed Consent

There was no requirement for informed consent.

Authors' Contributions

YZ and GW designed the research; JM performed the research; JM, HZ, and CZ performed the data collection and analysis; JM wrote the paper; YZ and GW revised the manuscript. All authors read and approved the final manuscript.

ORCID ID

Jianglei Ma: 0000-0002-2770-7597
Huijie Zhang: 0000-0002-1976-5543
Chenfang Zhang: 0009-0007-8981-6143
Yuanyuan Zhang: 0000-0002-5263-4895
Guangming Wang: 0000-0002-0220-1493

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability

The datasets generated during and/or analyzed during the current study are publicly available.

References

1) Turner D, Ricciuto A, Lewis A, D'amico F, Dhaliwal J, Griffiths AM, Bettenworth D, Sandborn WJ, Sands BE, Reinisch W, Schölmerich J. STRIDE-II: an update on the selecting therapeutic targets in inflammatory bowel disease (STRIDE) initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. Gastroenterology 2021; 160: 1570-1583.
2) Ordas I, Eckmann L, Talamini M, Baumgart DC, Sandborn WJ. Ulcerative colitis. Lancet 2012; 380: 1606-1619.
3) Conrad K, Roggenbuck D, Laass MW. Diagnosis and classification of ulcerative colitis. Autoimmun Rev 2014; 13: 463-466.
4) Wang Y, Ouyang Q, APDW 2004 Chinese IBD working group. Ulcerative colitis in China: retrospective analysis of 3100 hospitalized patients. J Gastroenterol Hepatol 2007; 22: 1450-1455.
5) da Silva BC, Lyra AC, Rocha R, Santana GO. Epidemiology, demographic characteristics and prognostic predictors of ulcerative colitis. World J Gastroenterol 2014; 20: 9458-9467.
6) Chambers WM, Warren BF, Jewell DP, McC. Mortensen NJ. Cancer surveillance in ulcerative colitis. Br J Surg 2005; 92: 928-936.
7) Sandborn WJ, Su C, Sands BE, D'Haens GR, Vermeire S, Schreiber S, Danese S, Feagan BG, Reinisch W, Niezychowski W, Friedman G. Tofacitinib as induction and maintenance therapy for ulcerative colitis. N Engl J Med 2017; 376: 1723-1736.
8) Schreiber S, Ben-Horin S, Leszczyszyn J, Dudkowiak R, Lahat A, Gawdis-Wojnarska B, Pukitis A, Horynski M, Farkas K, Kierkus J, Kowalski M. Randomized controlled trial: subcutaneous vs intravenous infliximab CT-P13 maintenance in inflammatory bowel disease. Gastroenterology 2021; 160: 2340-2353.
9) Hanzel J, Hulshoff MS, Grootjans J, D'Haens G. Emerging therapies for ulcerative colitis. Expert Rev Clin Immunol 2022; 18: 513-524.
10) Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012; 28: 882-883.
11) Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.
12) Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010; 26: 1572-1573.
13) Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.
14) Zhang C, Zheng Y, Li X, Hu X, Qi F, Luo J. Genome-wide mutation profiling and related risk signature for prognosis of papillary renal cell carcinoma. Ann Transl Med 2019; 7: 427.
15) Li Y, Tang M, Zhang FJ, Huang Y, Zhang J, Li J, Wang Y, Yang J, Zhu S. Screening of ulcerative

colitis biomarkers and potential pathways based on weighted gene co-expression network, machine learning and ceRNA hypothesis. Hereditas 2022; 159: 42.

16) Nanki K, Fujii M, Shimokawa M, Matano M, Nishikori S, Date S, Takano A, Toshimitsu K, Ohta Y, Takahashi S, Sugimoto S. Somatic inflammatory gene mutations in human ulcerative colitis epithelium. Nature 2020; 577: 254-259.

17) Nikolaus S, Rutgeerts P, Fedorak R, Steinhart AH, Wild GE, Theuer D, Möhrle J, Schreiber S. Interferon beta-1a in ulcerative colitis: a placebo controlled, randomised, dose escalating study. Gut 2003; 52: 1286-1290.

18) Georgiou AN, Ntritsos G, Papadimitriou N, Dimou N, Evangelou E. Cigarette smoking, coffee consumption, alcohol intake, and risk of Crohn's disease and ulcerative colitis: a Mendelian randomization study. Inflamm Bowel Dis 2021; 27: 162-168.

19) Blackwell J, Saxena S, Alexakis C, Bottle A, Cecil E, Majeed A, Pollok RC. The impact of smoking and smoking cessation on disease outcomes in ulcerative colitis: a nationwide population-based study. Aliment Pharmacol Ther 2019; 50: 556-567.

20) Parrish-Novak J, Dillon SR, Nelson A, Hammond A, Sprecher C, Gross JA, Johnston J, Madden K, Xu W, West J, Schrader S. Interleukin 21 and its receptor are involved in NK cell expansion and regulation of lymphocyte function. Nature 2000; 408: 57-63.

21) Habib T, Nelson A, Kaushansky K. IL-21: a novel IL-2-family lymphokine that modulates B, T, and natural killer cell responses. J Allergy Clin Immunol 2003; 112: 1033-1045.

22) Collins M, Whitters MJ, Young DA. IL-21 and IL-21 receptor: a new cytokine pathway modulates innate and adaptive immunity. Immunol Res 2003; 28: 131-140.

23) Fina D, Caruso R, Pallone F, Monteleone G. Interleukin-21 (IL-21) controls inflammatory pathways in the gut. Endocr Metab Immune Disord Drug Targets 2007; 7: 288-291.

24) Monteleone G, Caruso R, Fina D, Peluso I, Gioia V, Stolfi C, Fantini MC, Caprioli F, Tersigni R, Alessandroni L, MacDonald TT. Control of matrix metalloproteinase production in human intestinal fibroblasts by interleukin 21. Gut 2006; 55: 1774-1780.

25) Sarra M, Monteleone I, Stolfi C, Fantini MC, Sileri P, Sica G, Tersigni R, Macdonald TT, Pallone F, Monteleone G. Interferon-gamma-expressing cells are a major source of interleukin-21 in inflammatory bowel diseases. Inflamm Bowel Dis 2010; 16: 1332-1339.

26) Monteleone G, Monteleone I, Fina D, Vavassori P, Blanco GD, Caruso R, Tersigni R, Alessandroni L, Biancone L, Naccari GC, MacDonald TT. Interleukin-21 enhances T-helper cell type I signaling and interferon-gamma production in Crohn's disease. Gastroenterology 2005; 128: 687-694.

27) Yu J, He S, Liu P, Hu Y, Wang L, Wang X, Han Y, Zhu X. Interleukin-21 promotes the development of ulcerative colitis and regulates the proliferation and secretion of follicular T helper cells in the colitides microenvironment. Mol Med Rep 2015; 11: 1049-1056.

28) Bhan AK, Mizoguchi E, Smith RN, Mizoguchi A. Lessons for human inflammatory bowel disease from experimental models. Curr Opin Gastroenterol 1999; 15: 285-290.

29) Targan SR, Karp LC. Defects in mucosal immunity leading to ulcerative colitis. Immunol Rev 2005; 206: 296-305.

30) Tatiya-Aphiradee N, Chatuphonprasert W, Jarukamjorn K. Immune response and inflammatory pathway of ulcerative colitis. J Basic Clin Physiol Pharmacol 2018; 30: 1-10.

31) Ungaro R, Mehandru S, Allen PB, Peyrin-Biroulet L, Colombel JF. Ulcerative colitis. Lancet 2017; 389: 1756-1770.

32) Tong X, Zheng Y, Li Y, Xiong Y, Chen D. Soluble ligands as drug targets for treatment of inflammatory bowel disease. Pharmacol Ther 2021; 226: 107859.

33) Nakase H, Sato N, Mizuno N, Ikawa Y. The influence of cytokines on the complex pathology of ulcerative colitis. Autoimmun Rev 2022; 21: 103017.

34) Teng MW, Bowman EP, Mcelwee JJ, Smyth MJ, Casanova JL, Cooper AM, Cua DJ. IL-12 and IL-23 cytokines: from discovery to targeted therapies for immune-mediated inflammatory diseases. Nat Med 2015; 21: 719-729.

35) Jauregui-Amezaga A, Somers M, De Schepper H, Macken E. Next generation of biologics for the treatment of Crohn's disease: an evidence-based review on ustekinumab. Clin Exp Gastroenterol 2017; 10: 293-301.

36) Allocca M, Furfaro F, Fiorino G, Gilardi D, D'Alessio S, Danese S. Can IL-23 be a good target for ulcerative colitis? Best Pract Res Clin Gastroentrol 2018; 32: 95-102.

37) Fiorino G, Allocca M, Correale C, Roda G, Furfaro F, Loy L, Zilli A, Peyrin-Biroulet L, Danese S. Positioning ustekinumab in moderate-to-severe ulcerative colitis: new kid on the block. Expert Opin Biol Ther 2020; 20: 421-427.

38) Cappellesso F, Orban MP, Shirgaonkar N, Berardi E, Serneels J, Neveu MA, Di Molfetta D, Piccapane F, Caroppo R, Debellis L, Ostyn T. Targeting the bicarbonate transporter SLC4A4 overcomes immunosuppression and immunotherapy resistance in pancreatic cancer. Nat Cancer 2022; 3: 1464-1483.

39) Chen X, Chen J, Feng Y, Guan W. Prognostic value of SLC4A4 and its correlation with immune infiltration in colon adenocarcinoma. Med Sci Monit 2020; 26: e925016-1.

40) Zheng Y, Zeng J, Lin D, Xia H, Wang X, Chen L, Chen H, Huang L, Zeng C. Extracellular vesicles derived from cancer-associated fibroblast carries miR-224-5p targeting SLC4A4 to promote the proliferation, invasion and migration of colorectal cancer cells. Carcinogenesis 2021; 42: 1143-1153.

41) Rogler G. Chronic ulcerative colitis and colorectal cancer. Cancer Lett 2014; 345: 235-241.

42) Baron JH, Connell AM, Lennard-Jones JE, Jones FA. Sulphasalazine and salicylazosulphadimidine in ulcerative colitis. Lancet 1962; 1: 1094-1096.

43) Cevallos SA, Lee JY, Velazquez EM, Foegeding NJ, Shelton CD, Tiffany CR, Parry BH, Stull-Lane AR, Olsan EE, Savage HP, Nguyen H. 5-Aminosalicylic acid ameliorates colitis and checks dysbiotic Escherichia coli expansion by activating PPAR-gamma signaling in the intestinal epithelium. mBio 2021; 12: e03227-20.

44) Xu J, Chen N, Wu Z, Song Y, Zhang Y, Wu N, Zhang F, Ren X, Liu Y. 5-Aminosalicylic acid alters the gut bacterial microbiota in patients with ulcerative colitis. Front Microbiol 2018; 9: 1274.