



OPEN Autophagy related biomarkers in ulcerative colitis revealed by bioinformatics analysis and immune correlation

Xinyi Xue Chen^{1,2}, Shichen Min^{1,2}, Zhaofeng Shen^{1,2}, Mengyuan Zhang^{1,2}, Lejuan Huang^{1,2}, Mengran Su^{1,2}, Hong Shen^{1,2}✉ & Lei Zhu^{1,2}✉

Ulcerative colitis (UC) is a chronic inflammatory bowel disease of unknown etiology characterized by abnormal mucosal immune responses and persistent intestinal inflammation. UC significantly reduces quality of life and imposes a substantial economic burden. Early diagnosis and personalized treatment are crucial for improving outcomes; however, significant challenges remain. This study aimed to identify novel molecular targets for UC treatment. Gene expression data related to UC were obtained from the Gene Expression Omnibus, and differentially expressed genes (DEGs) were identified based on the criteria $P < 0.05$ and $|\log FC| > 0.5$. The DEGs were intersected with autophagy-related genes (ARGs) to explore the relationship between UC pathogenesis and autophagy. LASSO regression, random forest regression, and support vector machine-recursive feature elimination were employed to screen for differentially expressed ARGs. The robustness and generalizability of the intersected genes were validated in external datasets and mouse models, and immune infiltration was analyzed to assess the interaction between UC and the immune system. *PEA15* was identified as a potential UC diagnostic marker, and the diagnostic value of *SERPINA1*, *CASP4*, and *CASP1* was reconfirmed. Our findings provide new insights into the pathogenesis of UC, facilitating early diagnosis and the development of personalized therapeutic strategies.

Keywords Ulcerative colitis (UC), Autophagy, Machine learning, Immune infiltration, R studio

Ulcerative colitis (UC) is a chronic inflammatory bowel disease primarily characterized by abdominal pain, diarrhea, and colorectal bleeding¹. It often requires lifelong treatment and is associated with various complications during disease progression, significantly affecting patients' quality of life and imposing a substantial economic burden². In recent years, the global incidence of UC has steadily increased^{3,4}. To date, the exact etiology of UC remains unclear, but it is widely believed to be closely related to genetic factors, gut microbiota, and immune responses⁵. Early diagnosis is crucial for timely symptom management and prevention of complications. However, current diagnostic methods are inadequate for meeting early detection needs⁶. Gastrointestinal endoscopy and histopathological biopsy are the primary diagnostic tools for UC, but they often fail to provide a definitive diagnosis in cases with atypical endoscopic or pathological features, leading to delayed treatment⁷. Therefore, investigating potential biomarkers for UC holds significant value for early diagnosis and personalized treatment.

Autophagy is generally considered a self-protective cell mechanism that enables the clearance of damaged proteins, organelles, and invading pathogens, which is critical for maintaining intracellular homeostasis⁸. The mechanisms of autophagy are primarily classified into three types: chaperone-mediated autophagy, microautophagy, and macroautophagy⁹. Autophagy is closely associated with the development and progression of neurodegenerative diseases¹⁰, metabolic disorders¹¹, and cancers¹². Autophagy can regulate the apoptosis of intestinal epithelial cells and maintain the function of the intestinal epithelial barrier¹³. Dysregulation of autophagy may lead to intestinal homeostasis imbalance, gut microbiota dysbiosis, and exacerbation of intestinal inflammation¹⁴. Given that autophagy plays a vital role in regulating intestinal homeostasis, we hypothesized that autophagy may participate in the pathogenesis and progression of UC. We employed bioinformatics techniques to investigate key genes associated with autophagy and their relevance to UC. Our objective was to elucidate the

¹Affiliated Hospital of Nanjing University of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, Nanjing, Jiangsu Province, China. ²Nanjing University of Chinese Medicine, Jiangsu Province, China. ✉email: shenhong999@njucm.edu.cn; zhulei5100@njucm.edu.cn

relationship between autophagy-related genes (ARGs) and the development of UC and provide novel targets and insights to facilitate early diagnosis and personalized treatment.

Materials and methods

Data acquisition

Data were obtained from the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>), and GSE87466¹⁵ and GSE92415¹⁶ were selected as the training set. GSE87466 contained 87 UC samples and 21 normal samples. GSE92415 contained 162 UC samples and 21 normal samples. GSE38713¹⁷ and GSE75214¹⁸ were selected as the validation set, with the former containing 30 UC samples and 13 normal samples, and the latter containing 74 UC samples and 11 normal samples (data from Crohn's disease samples and inactive patient samples were excluded; Table 1). The ARG set was obtained from the HADb database (<http://www.autophagy.lu/index.html>).

Data processing

The probe IDs of each gene in the dataset downloaded from the GEO database were mapped to the gene symbols. When multiple probe IDs corresponded to the same gene, the average expression value of the corresponding probe IDs was taken as the expression value of the gene. If there were missing data in the dataset, the entry was deleted. In order to reduce interference caused by technical reasons, the limma package was used to normalize the data. After merging the GSE87466 and GSE92415 datasets, the removeBatchEffect function was used to remove the batch effect of both datasets and construct a unified gene dataset. The same data processing method was used to process the GSE38713 and GSE75214 datasets for subsequent validation of the results.

Screening and functional enrichment of UC and mitochondria-related differentially expressed genes

The limma package was used to identify differentially expressed genes (DEGs) between UC samples and normal samples in the combined dataset, with $P < 0.05$ and $|\log FC| > 0.5$ used as the criteria for DEG selection. We used the ggplot2 and pheatmap packages to visualize the DEGs, and draw volcano.

maps and clustered heat maps. The intersect function was used to identify the overlap between the selected DEGs and 222 ARGs, yielding the genes associated with both UC and autophagy. The results were visualized using a Venn diagram. To explore the biological functions of the DEGs and their associated pathways, gene ontology (GO)¹⁹ and Kyoto Encyclopedia of Genes and Genomes (KEGG)²⁰ enrichment analyses were performed using the clusterProfiler package. Enriched pathways were selected based on a threshold of $P < 0.05$, and the results were visualized. Gene sets were analyzed using the GSVA package²¹ to investigate the enrichment of sample genes in hallmark gene sets, which were obtained from the MSigDB database. The normalized enrichment score was used to assess the magnitude of enrichment, with a value greater than 0 indicating a positive association with the pathway, and a value less than 0 indicating a negative association.

Machine learning model construction and identification of target genes

To identify potential biomarkers for UC, three machine learning models were employed to further screen the differentially expressed ARGs (DE-ARGs): least absolute shrinkage and selection operator (LASSO) regression, random forest (RF) regression, and support vector machine-recursive feature elimination (SVM-RFE). The LASSO algorithm was used to find the best model by introducing λ (the shrinkage operator), and the target genes were screened by 10-fold cross-validation. The Lasso algorithm was operated via the glmnet package in R. The RF algorithm was constructed using the randomForest package, which is designed to produce smaller fluctuations in DEG prediction. The RF algorithm was constructed with 500 decision trees, employing leave-group-out cross-validation with classification accuracy as the evaluation metric. The SVM-RFE model was constructed using the e1071 and sigFeature packages. The SVM-RFE model scores and ranks each gene, removes the gene with the.

smallest score, and then uses the remaining genes to model again. Finally, we obtained an optimal model. The SVM-RFE model was implemented using a radial basis function kernel, with 5-fold cross-validation used to identify the optimal DEGs set. The genes identified by each algorithm were intersected to obtain the characteristic genes associated with UC and autophagy. The final target genes were screened based on the importance score of each gene in the model.

Evaluation and validation of targeted gene diagnostic capabilities

The pROC package was used to plot ROC curves for individual target genes, and the area under the curve (AUC) was calculated to evaluate the ability of the target genes to distinguish between UC samples and normal samples.

Series	Platforms	Samples	Tissue	Type
GSE87466	GPL13158	87 UC and 21 Normal	Mucosal biopsy samples	mRNA
GSE92415	GPL13158	162 UC and 21 Normal	Mucosal biopsy samples	mRNA
GSE38713	GPL570	30 UC and 13 Normal	Mucosal biopsy samples	mRNA
GSE75214	GPL6244	74 UC and 11 Normal (removal of Crohn's disease samples and inactive UC samples)	Mucosal biopsy samples	mRNA

Table 1. Details of datasets.

Similarly, the discriminative performance of the target genes was validated using the GSE38713 and GSE75214 datasets.

Animal model establishment and treatment procedures

C57BL/6J mice (male, 6–8 weeks) were supplied by Zhejiang Vital River Laboratory Animal Technology Co. Ltd (Zhejiang, China) and kept in a specific-pathogen-free room with a 12/12 h light/dark cycle (animal usage license number: SYXK(Su)2022-0070). All animal experiments were conducted in accordance with the ARRIVE guidelines (<https://arriveguidelines.org>) and were approved by the Institutional Animal Care and Use Committee of Nanjing University of Chinese Medicine. (application number: 2023DW-082-01). After 7 days of acclimatization, 36 C57BL/6J mice were randomly divided into six groups: the normal (Ctrl) group ($n=6$), and the DSS treatment (DSS) group ($n=6$). To establish the UC model, 3% DSS was administered in the drinking water for 7 days. On the morning of day 15, mice were anesthetized using an isoflurane vaporizer and maintained in a surgical plane of anesthesia, as confirmed by a negative pedal withdrawal reflex test. After confirming adequate anesthesia, mice were humanely euthanized via cervical dislocation. Colonic tissues were immediately collected, rinsed with cold phosphate-buffered saline (PBS), and prepared for subsequent histological and molecular analyses.

Hematoxylin and Eosin (H&E) staining

The colon was fixed with paraformaldehyde and embedded in paraffin, after which 4 μ m thick specimens were sliced off and deparaffinized with xylene and ethanol. After the samples were rinsed with tap water, hematoxylin and eosin (H&E) staining was performed using a standard protocol; specimens were soaked in hematoxylin for 5 min and then in eosin dye for another 5 min. The specimens were dehydrated in ethanol and finally washed with xylene.

Real-time PCR

Total RNA was extracted and used to synthesize cDNA with a cDNA Synthesis Kit. The cycle threshold values were determined using the LightCycler[®] 96 system (Roche) and a qRT-PCR SYBR Green Kit (Vazyme). β -actin served as the reference gene. All primers were designed and synthesized by Genaray Biotech (Shanghai), and their sequences are listed in Table 2.

Protein-protein interaction network construction

Protein-protein interaction (PPI) networks were constructed to analyze interactions between proteins encoded by DE-ARGs, using the STRING online database (<https://cn.string-db.org/>). A medium confidence level of 0.4 was set.

Immune infiltration analysis

Immune infiltration analysis of the dataset was performed by using the GSVA package to evaluate the enrichment levels of immune cells in different groups, and boxplots were generated. To assess the correlation between the target genes and immune infiltration levels, a correlation analysis was conducted, and a correlation heatmap was created using the pheatmap package. The immunity-related gene dataset was obtained from <http://cis.hku.hk/TISIDB/data/download/CellReports.txt>.

Statistical analysis

All data processing, statistical analysis, and plotting were performed using R software (version: 4.3.2). Differences between the two groups of samples were analyzed using the Wilcoxon rank sum test. Pearson correlation analysis was used to determine the correlation between immune cells and marker genes. The level of statistical significance was set at $P < 0.05$.

Results

Data processing

The GSE87466 and GSE92415 datasets were downloaded from the GEO database as the training set, and GSE38713 and GSE75214, as the validation set. All datasets were judged by the function without log2 logarithmic transformation. To eliminate variation in expression due to experimental techniques and other factors, we used the limma package to normalize the data across all datasets and boxplots to observe baseline levels (Fig. 1a). The data from GSE87466 and GSE92415 were merged and de-batched using the limma package (Fig. 1b, Supplementary Fig. 1).

Gene	Forward (5' to 3')	Reverse (5' to 3')
<i>Serpine1</i>	AGAATCCACACAGCCCATCAG	CACCGAGGACACGCCATAGG
<i>Pea15a</i>	TCCTCACTATGGTGGTTGACTACAG	TGGCACTGGGAATACGGGTTAG
<i>Hspa5</i>	GCCGAGGAGGAGGACAAGAAG	GAACACACCGACGCAGGAATAG
<i>Casp4</i>	GCTACGATGTGGTGGTGAAAGAG	GTGCTGTCTGATGTCTGGTGTTC
<i>Casp1</i>	ACTCGTACACGTCTTGCCCTCA	CTGGGCAGGCAGCAAATTC

Table 2. Primer sequences.

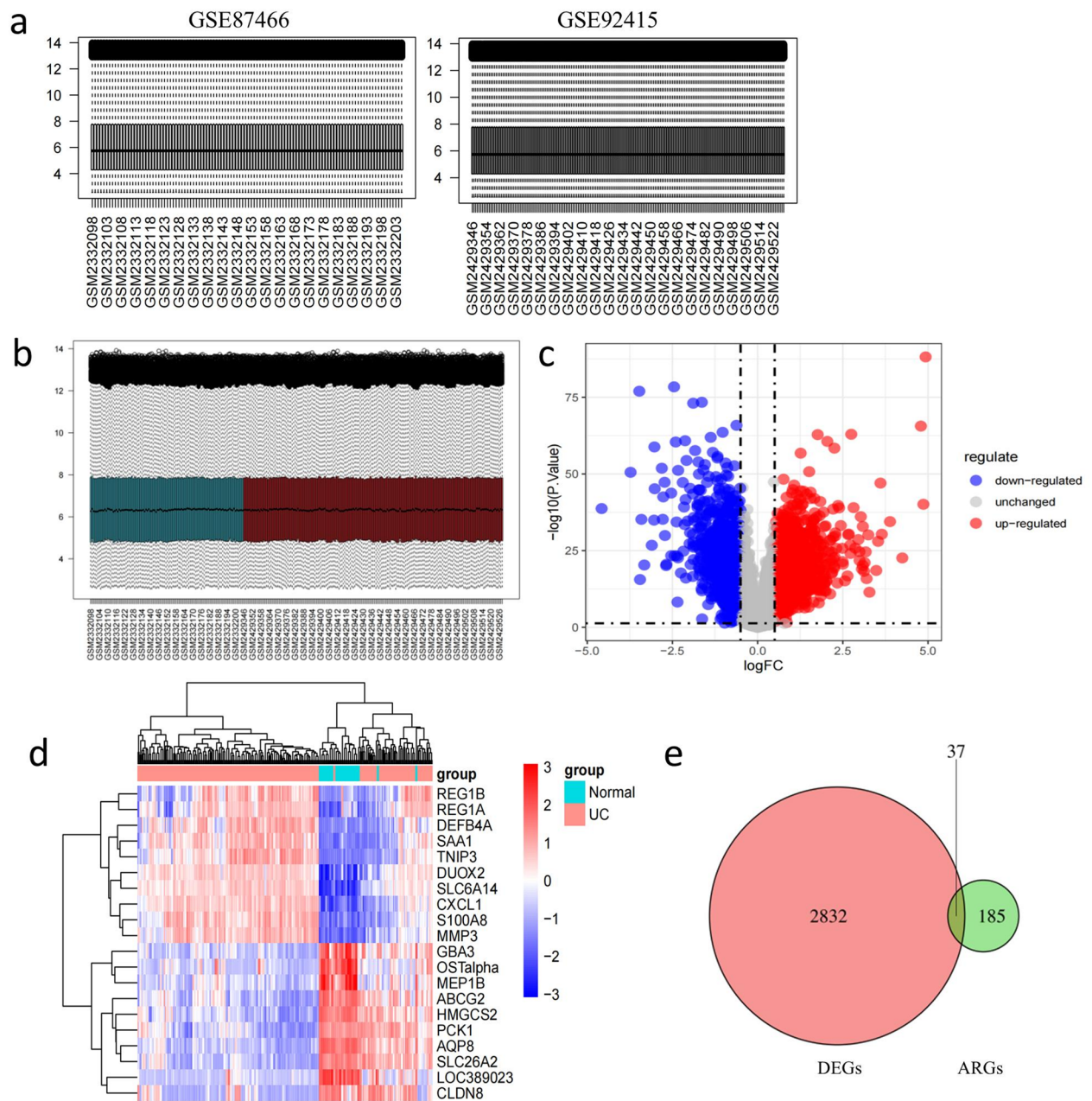


Fig. 1. Data preprocessing and differentially expressed gene visualization. (a) Boxplot after data preprocessing. (b) Boxplot after data merging and batch effect removal. (c) Volcano plot of differentially expressed genes between UC patients and normal population, with red representing upregulated genes and blue representing downregulated genes. (d) Heat map illustrating clustering of the top 20 differentially expressed genes. The color gradient from red to blue represents the change from high to low expression. (e) Venn diagram of the intersection between UC differentially expressed genes and autophagy-related genes. Represented are differentially expressed genes (DEGs, red), autophagy-related genes (ARGs, green), and their intersection (overlap).

Screening for differentially expressed genes

The limma package was used to construct the comparison matrix, and the eBayes function within the package was applied to fit a linear model for identifying DEGs. The criteria for DEG selection were $P < 0.05$ and $|\log_{2}(FC)| > 0.5$. In UC patients, 1,607 and 1,262 genes were upregulated and downregulated, respectively, compared with their expression levels in the normal population. The ggplot2 package was employed to create a volcano plot of the DEGs (Fig. 1c). The top 10 most significant DEGs among the upregulated and downregulated genes were selected, and the pheatmap package was used to generate a heatmap illustrating the clustering of these genes (Fig. 1d). As shown, the expression levels of the top 20 genes differed distinctly between UC samples and normal samples. To explore the relationship between UC and autophagy, the identified DEGs were intersected with

ARGs, resulting in a total of 37 DE-ARGs. Among these, 28 genes were upregulated and 9 were downregulated in UC patient samples (Fig. 1e).

Functional enrichment analysis

To explore the biological properties of the DEGs, GO enrichment analysis was performed on upregulated and downregulated genes. The graph shows the top six pathways with the most significant differences in biological process (BP), cellular component (CC) and molecular function (MF) terms, at a significance level of $P < 0.05$ (Figs. 2a, b). Based on the enrichment analysis graph, upregulated DEGs were mainly involved in the cellular response to organic substance, response to organic substance, apoptotic process, programmed cell death, cell death and response to chemicals, and they were mainly localized in CCs such as the aggresome, endoplasmic reticulum lumen, inclusion body, and canonical inflammasome complex. Enriched MFs included cytokine activity, molecular function regular activity, and misfolded protein binding. The downregulated genes were mainly associated with protein autophosphorylation, vacuole organization, autophagy, process utilizing autophagy mechanism, macroautophagy, and autophagosome assembly and were mainly localized in the autophagosome and vacuole. They were also associated with protein kinase activity, phosphotransferase activity alcohol group as acceptors, kinase activity, and transferase activity transferring.

Functional features and related pathways of the DEGs were analyzed using KEGG pathway enrichment analysis, and the top 10 entries were visualized. The upregulated DEGs were associated with PD-L1 expression and the PD-1 checkpoint pathway in cancer, the NOD-like receptor signaling pathway, and viral protein interaction with cytokines and cytokine receptors. The downregulated DEGs were enriched in the autophagy-animal pathway (Figs. 2c, d).

Gene set variation analysis (GSVA) differs from the GO and KEGG pathway enrichment analyses. Whereas GO and KEGG analyze individual DEGs, GSVA analyzes the entire set of genes and assesses the enrichment differences between two sets of samples across hallmark gene sets. GSVA preserves all the genes, maps the gene expression data to a predefined set of genes, calculates enrichment scores for each sample in each gene set, and sorts the genes based on the enrichment score (Fig. 2e). UC-related genes were enriched in TNFA signaling via NFkB, epithelial mesenchymal transition, and inflammatory response pathways and were negatively correlated with bile acid metabolism, xenobiotic metabolism, and oxidative phosphorylation pathways.

Machine learning to identify target genes

To identify genes with the most significant predictive power, we used LASSO regression analysis, RF

regression, and SVM-RFE. In the LASSO regression analysis, after 10-fold cross-validation, we identified 0.0018 as the optimal λ value, and 14 characterized genes were screened from 37 DE-ARGs (Fig. 3a). In the RF algorithm, 19 characterized genes were screened, and importance ranking histograms were plotted by constructing 1000 decision trees (Fig. 3b). When using SVM-RFE to screen DE-ARGs, we performed a 5-fold cross-validation and repeated the model construction 5 times to produce a characterized genes selection scheme that incorporates approximately 1–37 genes simultaneously. Error was minimized when the model incorporated all the genes. A ranked histogram (Fig. 3c) was plotted based on the importance of each gene in the model. The intersection of the feature genes obtained by using the three algorithms was found, and 11 feature genes (*PEA15*, *HSPA5*, *SERPINA1*, *CASP4*, *CASP1*, *CCL2*, *CX3CL1*, *BAG3*, *ULK3*, *TP53INP2*, and *DAPK2*) were ultimately obtained (Fig. 3d). By combining the three algorithms to rank the importance of the DEGs, we found that *SERPINA1*, *PEA15*, *CASP4*, and *CASP1* had the greatest importance in all three algorithms. Thus, we used these four genes as target genes for subsequent analysis and exploration.

Target gene expression and diagnostic performance evaluation

Using boxplots, we evaluated the differential expression of the four target genes between UC patients and a healthy population (Fig. 4a). The boxplots revealed significant differential expression for all four genes between the two groups. To assess the discriminative ability of the target genes, we constructed ROC curves (Fig. 4d). The AUC values for all four genes exceeded 0.9, demonstrating their strong discriminative power.

Diagnostic performance validation of target genes

To validate the discriminative performance of the four target genes, analyses were conducted using the GSE38713 and GSE75214 datasets. Boxplot analysis revealed that in both external validation sets (Fig. 4b, c), the expression trends of the four target genes were consistent with those observed in the training set and differed significantly between UC samples and normal samples ($P < 0.05$). ROC curve analysis showed that the AUCs for the 4 target genes exceeded 0.8, demonstrating their strong discriminative performance (Fig. 4e).

Animal experiment validation

To validate the findings of the machine learning analysis, we established a DSS-induced UC mouse model and assessed disease severity using indicators such as colon length and body weight changes. The DSS intervention significantly shortened colon length, induced colonic congestion and edema, and markedly reduced body weight (Fig. 5a, b). Histological analysis via H&E staining revealed notable pathological features in the DSS group, including crypt loss, inflammatory cell infiltration, and severe mucosal damage (Fig. 5c). Furthermore, immunofluorescence staining showed a significant reduction in the expression of MUC2, a key intestinal mucosal barrier protein, in the colonic tissue of DSS-treated mice (Fig. 5d), indicating compromised intestinal barrier integrity. To further explore molecular mechanisms, we performed qPCR analysis to examine the mRNA expression levels of *Serpine1*, *Pea15a*, *Caspase-4*, and *Caspase-1* in colonic tissues (Fig. 5e). The results revealed that the expression levels of these genes were significantly upregulated in the DSS group compared with those

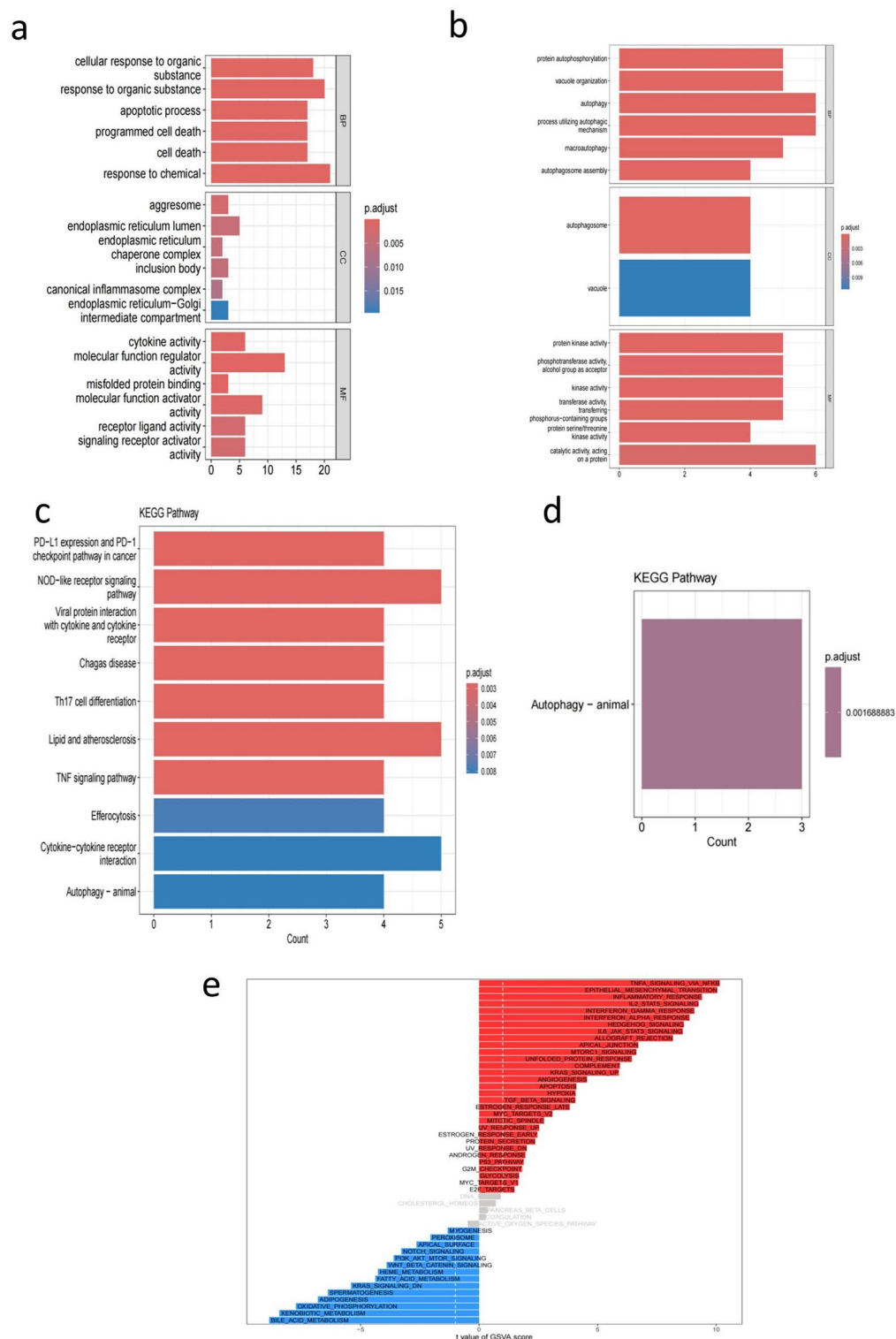
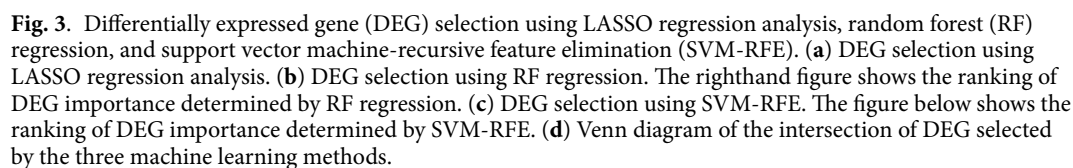


Fig. 2. Gene ontology (GO) enrichment analysis, Kyoto Encyclopedia of Genes and Genomes^{20,22,23} (KEGG) pathway enrichment analysis, and gene set variation analysis (GSVA). **(a)** GO enrichment analysis of differentially expressed genes (DEGs). The color gradient from blue to red indicates significance of the differences. **(b)** GO enrichment analysis of downregulated DEGs. The color gradient from blue to red indicates increasing significance of the differences. **(c)** KEGG pathway enrichment analysis of upregulated DEGs. The color gradient from blue to red indicates increasing significance of the differences. **(d)** KEGG pathway enrichment analysis of downregulated DEGs. Enrichment of downregulated genes was significant only in the autophagy pathway. **(e)** GSVA of DEGs. The heatmap visualizes t -values derived from GSVA scores, representing the activity levels of various biological pathways. Positive t -values (right side, red color scale) indicate upregulated pathways in ulcerative colitis, while negative values (left side, blue color scale) denote downregulated pathways.



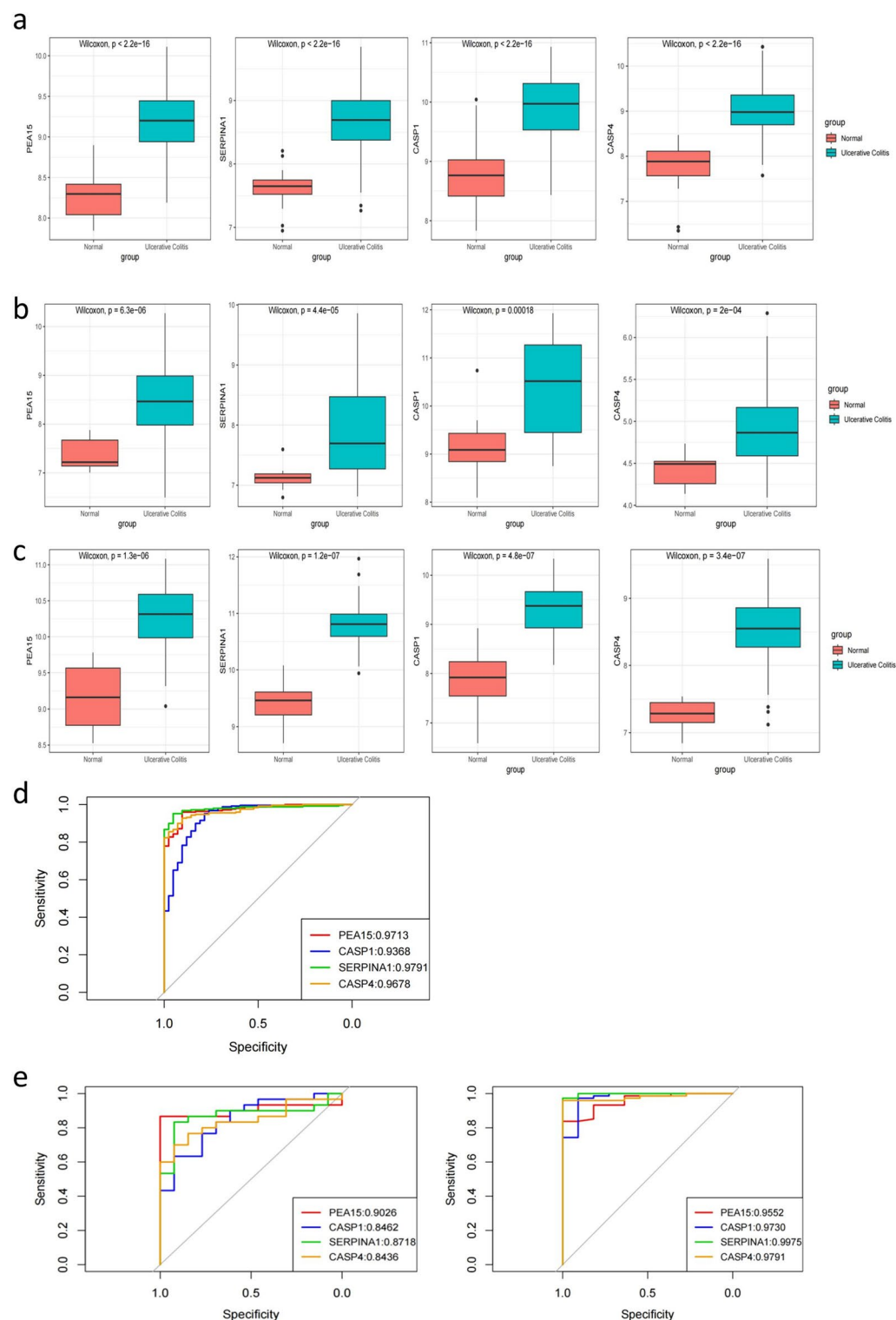


Fig. 4. Boxplots and receiver operating characteristic (ROC) curve of differentially expressed autophagy-related genes (DE-ARGs) from the training set (the merged dataset of GSE87466 and GSE92415) and test sets (GSE38713 and GSE75214). **(a)** Boxplot of DE-ARG (*PEA15*, *SERPINA1*, *CASP1*, and *CASP4*) expression levels in the training set, comparing the UC group (blue) and the normal group (red). **(b)** Boxplot of DE-ARG (*PEA15*, *SERPINA1*, *CASP1*, and *CASP4*) expression levels in the validation set GSE38713. **(c)** Boxplot of DE-ARG (*PEA15*, *SERPINA1*, *CASP1*, and *CASP4*) expression levels in the validation set GSE75214. **(d)** ROC curve illustrating the classification performance of the four DE-ARGs in the training set. **(e)** ROC curve illustrating the classification performance of the four DE-ARGs in the validation set GSE38713 and GSE75214, where a larger area under the curve (AUC) indicates better predictive performance and discriminative ability of the gene.

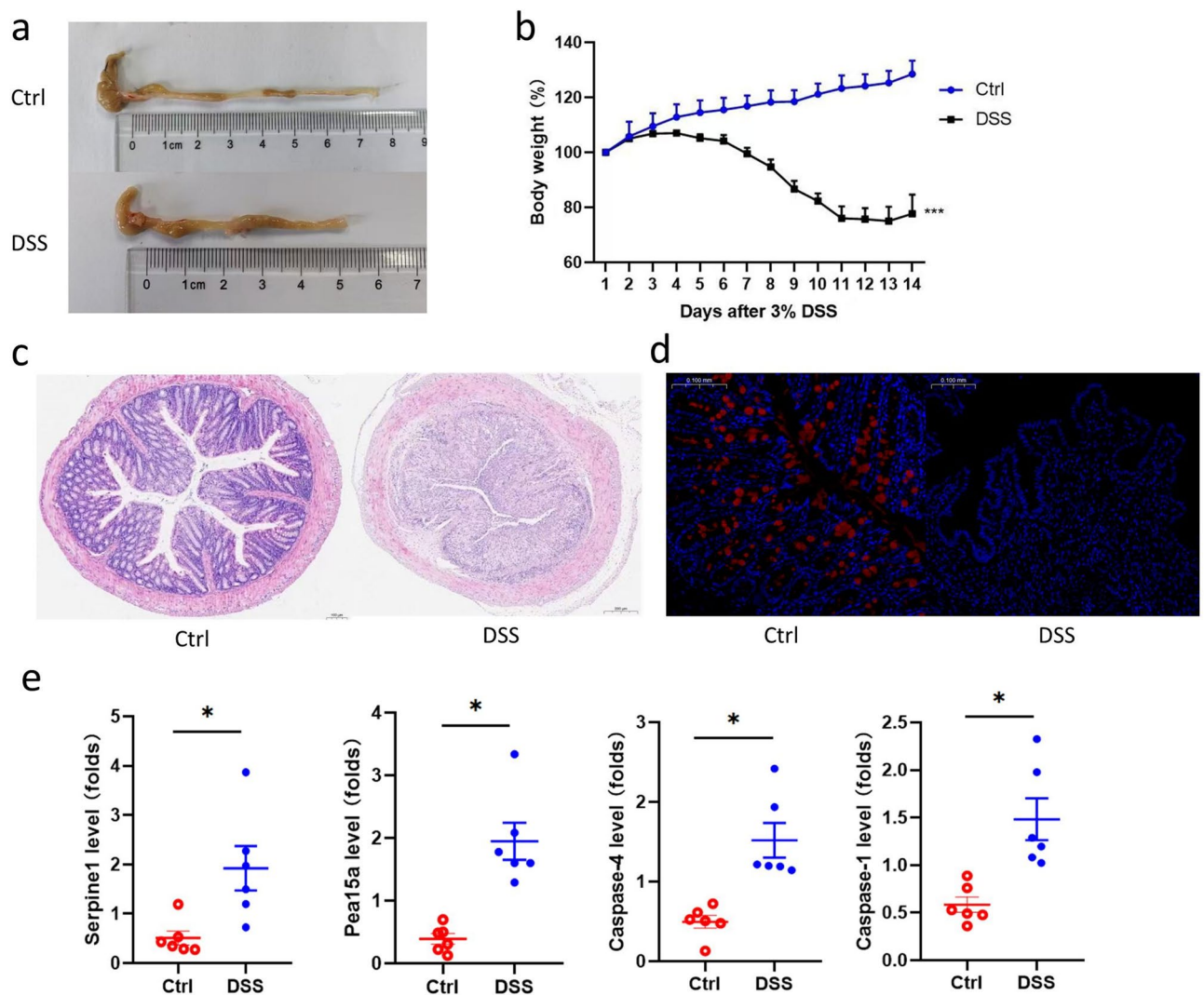


Fig. 5. Experimental validation of differential gene expression in mice. **(a)** Typical colon anatomy of mice in each group. **(b)** Changes in the weight of mice throughout the experiment. (two-way ANOVA). **(c)** H&E staining of colon tissue; scale bar, 100 μ m. **(d)** Immunofluorescence staining was used to show MUC2 expression in the colon. **(e)** The relative expression of *Serpine1*, *Pea15a*, *Casp4*, and *Casp1* in colon tissue were detected by real-time PCR. ($n \geq 6$). * $P < 0.05$ vs. DSS group, ** $P < 0.01$ vs. DSS group, *** $P < 0.001$ vs. DSS group, **** $P < 0.0001$ vs. DSS group.

in the control group, suggesting their potential involvement in the progression of UC and highlighting them as potential therapeutic targets.

PPI network construction

We used the STRING online database (<https://cn.string-db.org/>) to construct PPI network of DE-ARGs and the 11 target genes (Fig. 6a, b) and identify the interrelationships among the proteins encoded by the genes. In the PPI network of DE-ARGs, we removed genes that were not linked to any other genes, set a confidence level of 0.4, and clustered the genes. *SERPINA1*, *CASP4*, and *CASP1* were all associated with *HSPA5* in the network map of target genes. We hypothesized that *HSPA5* is involved in the key process of ulcerative junction genesis, which requires further investigation.

Immune infiltration analysis

To investigate the differences in the enrichment levels of immune cells in different samples, we performed immune infiltration analysis using the GSVA package and drew a clustered heatmap to show the ssgsea results (Fig. 6c). According to the clustered heat map image, the levels of immune cells differed significantly between UC samples and normal samples. We plotted a grouped boxplot of the degree of immune cell infiltration (Fig. 6d) and found that, except for memory B cell levels, the levels of the other 27 immune cells differed significantly between UC and normal samples, and the levels of most immune cells were higher in UC samples. Finally, we constructed a correlation heatmap to assess correlations between the expression of the four target genes and the

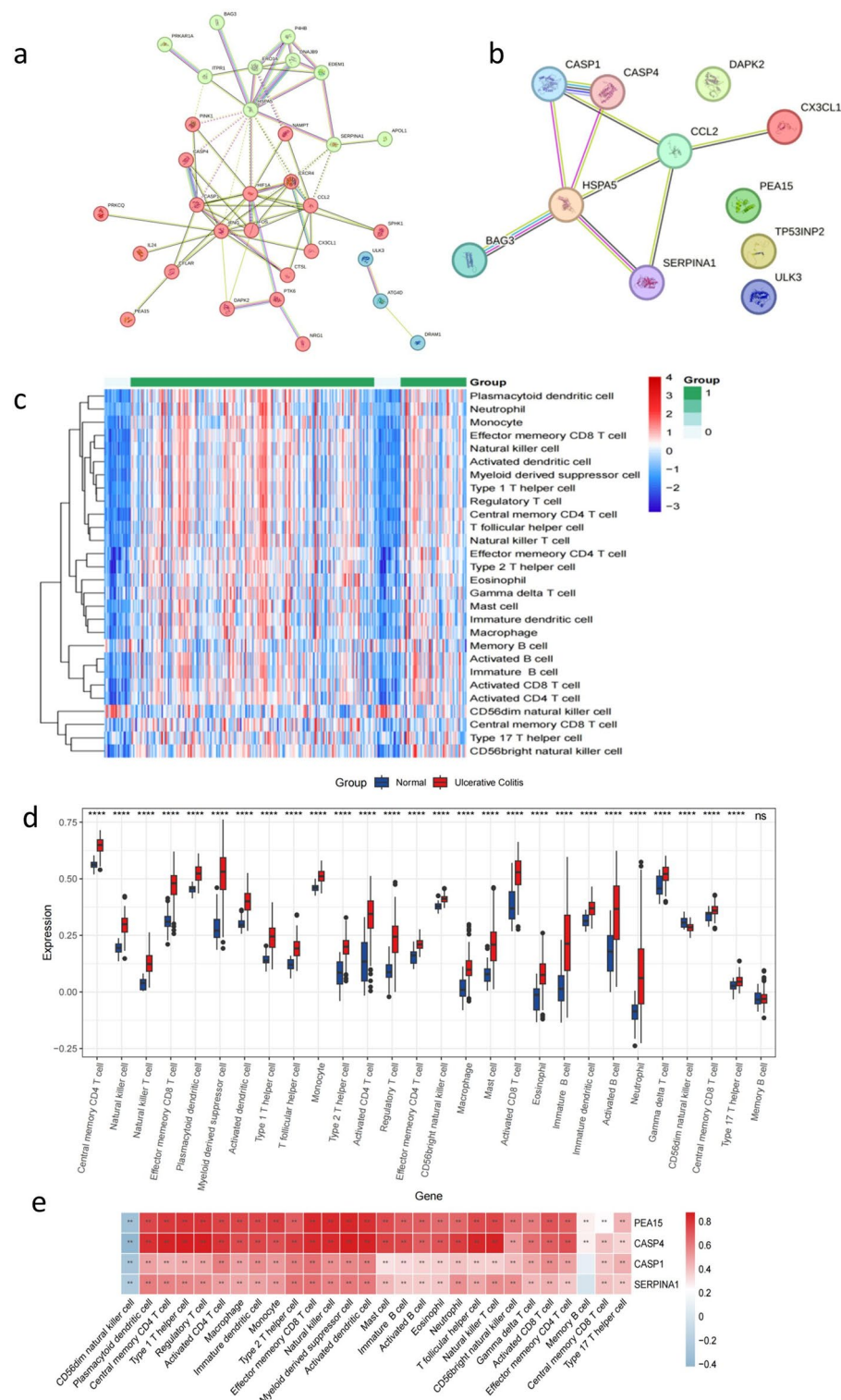


Fig. 6. Protein-protein interaction (PPI) network and immune infiltration analysis. **(a)** PPI network diagram illustrating the interactions among differentially expressed autophagy-related genes (DE-ARGs). The nodes represent genes or proteins, and differently colored nodes indicate distinct functional modules or clusters. **(b)** PPI Network of 11 target genes. **(c)** Clustering heatmap presenting immune cell infiltration levels across different samples. The color gradient from blue to red represents infiltration levels from low to high. **(d)** Boxplot comparing the expression levels of different immune cells between the normal group (blue) and the ulcerative colitis (UC) group (red). **(e)** Heatmap illustrating the correlation between the target genes (*PEA15*, *CASP4*, *CASP1*, and *SERPINA1*) and the infiltration levels of different immune cells. The color gradient from blue to red represents the degree of correlation from negative to positive, with deeper colors indicating stronger correlations.

levels of immune cells (Fig. 6e). The expression of four target genes was positively correlated with the levels of most immune cells (except for CD56dim natural killer [NK] cells), with a high degree of correlation.

Discussion

The incidence of UC has been rising steadily⁴, establishing it as a globally prevalent disease that poses a significant threat to health. However, current clinical treatments for UC, such as sulfasalazine, are associated with severe adverse effects²⁴, further compounding the treatment burden for patients. Thus, identifying novel targets for UC diagnosis and treatment is crucial for improving disease management and enhancing patients' quality of life.

A stable autophagic process is crucial for maintaining intestinal epithelial homeostasis, and the imbalance of intestinal epithelial homeostasis is considered a key factor in the development of UC. Therefore, the stability of the autophagic process may play an important role in the prevention and treatment of UC. Autophagy may undergo dynamic regulation in UC. For instance, one study showed that DSS-induced inflammation enhances autophagy in acute colitis, while curcumin and resveratrol exert protective effects on the intestinal mucosa by reducing autophagy²⁵. Conversely, another study demonstrated that resveratrol alleviates intestinal mucosal barrier dysfunction in mice with DSS-induced chronic colitis by enhancing autophagy²⁶. Further investigation into the bidirectional regulatory mechanisms of autophagy may not only provide deeper insights into the pathogenesis of UC but also offer novel molecular targets for its diagnosis and treatment.

In this study, genetic data of UC cases were obtained from the GEO public database. Using the GSE87466 and GSE92415 datasets, DEGs were identified based on the criteria $P < 0.05$ and $|\log FC| > 0.5$. These DEGs were then intersected with ARGs, resulting in 37 DE-ARGs. Further screening using three machine learning methods yielded 11 DE-ARGs, and based on importance ranking, 4 top-ranked genes (*PEA15*, *SERPINA1*, *CASP4*, and *CASP1*) were selected as target genes for subsequent analysis. Finally, immune infiltration analysis was performed on the four target genes to investigate their immune-related associations.

GO and KEGG pathway enrichment analyses were performed separately for the upregulated and downregulated DE-ARGs. Pathways enriched for the upregulated genes were predominantly associated with cell death and protein structural alterations, processes highly involved in immunity and cancer. It is well established that prolonged mucosal inflammation in UC patients increases their risk of developing colorectal cancer²⁷. Therefore, further analysis of these upregulated DE-ARGs may help uncover the mechanisms underlying the transition from inflammation to cancer and provide strategies for early intervention to reduce the incidence of colitis-associated colorectal cancer. The downregulated genes were found to primarily function in protein kinase and transferase activities, which phosphorylate amino acid residues, altering protein structure and interactions. Inhibition of protein phosphorylation has been shown to reduce inflammation in UC. For example, the FPR agonist Cmpd43 suppresses phosphorylation of cAMP-responsive element-binding proteins, reducing pro-inflammatory mediator secretion²⁸. Similarly, EFHD2 inhibits Cofilin phosphorylation, preventing TNF-induced epithelial apoptosis and protecting the intestines²⁹. As kinase pathways regulate inflammation and influence cancer progression³⁰, targeting these genes may offer novel therapeutic strategies for UC.

In addition, we performed GSVA of the entire gene set. The occurrence of UC was positively correlated with TNF α signaling via the NF- κ B pathway, epithelial-mesenchymal transition, and inflammatory response pathways, which were highly enriched in the UC group. These findings suggest that the activation of these pathways may play a significant role in the progression of UC. In contrast, compared with the UC group, the normal group showed enrichment in bile acid metabolism, xenobiotic metabolism, and oxidative phosphorylation pathways. The low expression of genes associated with these pathways in the UC group may be an important factor contributing to the pathogenesis of UC.

We applied three machine learning algorithms to develop a feature selection model for gene screening. The feature genes identified by the three algorithms were intersected to determine common candidates. The identification of the four target genes *PEA15*, *SERPINA1*, *CASP4*, and *CASP1* was primarily based on two key criteria: (1) their top-ranking positions in our analytical results and (2) existing literature evidence supporting the relevance of three of these genes in related studies, which further corroborates the reliability of our findings. The inclusion of *PEA15* was particularly warranted by limited existing research, prompting thorough investigation and discussion to potentially uncover novel biological significance.

To further validate the discriminatory performance of the target genes, we used the GSE38713 and GSE75214 datasets as external validation sets. The boxplots demonstrated that, in both validation datasets, the expression trends of the four target genes were consistent with those observed in the training set and showed significant differential expression between UC and normal samples ($P < 0.05$). ROC curve analysis further demonstrated that the AUC values for the target genes exceeded 0.8 in both validation datasets, confirming their robust discriminatory performance.

In addition, a DSS-induced UC mouse model was established to further validate the expression differences in the four identified genes. qPCR analysis revealed that *Pea15*, *Casp1*, *Serpina1*, and *Casp4* were significantly upregulated in the colonic tissues of model mice compared with those in the control group ($P < 0.05$). These findings not only corroborate the results obtained from prior analyses but also highlight the potential involvement of these genes in the pathophysiology of UC. Given their elevated expression levels in the disease state, these four genes may represent promising molecular targets for therapeutic intervention in UC. Further studies are warranted to elucidate their specific roles and underlying mechanisms in UC pathogenesis.

The four target genes analyzed in this study all participate in the occurrence and progression of the disease by influencing inflammatory responses. Søndergaard et al. found that levels of alpha-1 antitrypsin, encoded by *SERPINA1*, can be used to differentiate between mild, moderate, and severe ulcers³¹. *SERPINA1* is also linked to cancer progression, with high expression levels in colon cancer associated with poor prognosis due to enhanced STAT3 signaling³². We hypothesized that *SERPINA1* overexpression may exacerbate UC progression by enhancing the STAT3 signaling pathway, thereby promoting chronic inflammation and epithelial barrier

dysfunction. Its elevated expression in colorectal cancer suggests a potential role in inflammation-driven carcinogenesis associated with long-standing UC. It has been identified as a potential early diagnostic marker for colorectal cancer³³ and is also implicated in thyroid, gastric, and breast carcinomas^{34–37}.

CASP1 and *CASP4*, which belong to the cysteine asparaginase family, are involved in inflammatory and immune responses. *CASP1*, upon activation via the NLRP3 inflammasome, promotes the maturation of cytokines such as IL-1 β and IL-18^{38,39}. Inhibiting *CASP1* activity with disulfiram and Cu²⁺ has been shown to reduce inflammation in UC⁴⁰. *CASP4* plays a key role in intestinal inflammation, is highly expressed in colorectal cancer, and may serve as a therapeutic target⁴¹. *CASP1* and *CASP4* trigger pyroptosis, a form of programmed cell death⁴², by cleaving GSDMD, thereby activating the pro-inflammatory cytokines IL-1 β /IL-18, exacerbating inflammatory responses, and further aggravating intestinal barrier damage⁴³. Another study confirms that *CASP1* serves as a key biomarker in pyroptosis of intestinal epithelial cells in inflammatory bowel disease⁴⁴. As both genes are involved in cellular pyroptosis, the role of this process in UC requires further study.

PEA15 is a phosphatidylethanolamine-binding protein implicated in multiple cancers. It has been integrated into a prognostic prediction model for gastric cancer, where it demonstrates strong predictive power for prognosis and treatment response⁴⁵. Additionally, *PEA15* has been identified as a potential biomarker for renal cell carcinoma and hepatocellular carcinoma^{46,47} and holds promise as a therapeutic target for various diseases⁴⁸. According to our findings, *PEA15* consistently ranked highly in importance scores across all three algorithms. However, its specific role in the development of UC remains unclear. Previous studies suggest that *PEA15* may be associated with the ERK1/2-related pathway⁴⁸. Therefore, we hypothesize that *PEA15* may modulate mucosal inflammation by altering the ERK1/2 signaling pathway in UC. *PEA15* merits further investigation and may represent a novel therapeutic target for UC.

These four genes likely contribute to UC through distinct yet interconnected pathways. Further functional studies are needed to validate their roles and therapeutic potential.

Finally, we performed immune infiltration analysis to evaluate differences in immune cell enrichment across different samples. Notably, CD56dim NK cell infiltration levels were lower in UC samples than in normal samples, whereas the levels of the other immune cells were higher in UC samples. The correlation heatmap further illustrated the relationships and strengths of association between the target genes and immune cells. Interestingly, the correlation between the target genes and CD56dim NK cells differed from those of the other 27 immune cell types, as the expression of the target genes was negatively correlated with the level of CD56dim NK cells. We hypothesize that in UC, *PEA15*, *CASP1*, *CASP4*, and *SERPINA1* may collectively contribute to an immunosuppressive microenvironment, impairing the immune surveillance function of CD56dim NK cells and thereby exacerbating disease progression.

In summary, we identified *PEA15* as a potential diagnostic marker for UC, representing a novel finding distinct from those of previous studies. This discovery provides new targets and insights for understanding the mechanisms of UC and developing treatment. Additionally, we identified three other target genes—*SERPINA1*, *CASP4*, and *CASP1*—using three machine learning methods, which reconfirmed the diagnostic value of these genes in UC. In the future, diagnostic models incorporating these markers could be developed to facilitate early diagnosis of UC and advance precision medicine.

Our study had certain limitations. First, the parameters used in this study were limited to genetic data, neglecting the potential influence of other clinical correlates on UC. The model could be improved by incorporating these additional factors. Second, the sample size of this study was relatively small, increasing the potential for error. Expanding the sample size in future studies will be necessary to validate the model's performance more accurately. Third, our study primarily analyzed gene expression at the mRNA level, and validation at the protein level was not performed. Further verification studies should be performed to confirm whether the DEGs exhibit corresponding changes at the protein level and verify the biological significance of the findings. Finally, as our data were derived from genetic datasets in the GEO database, the diagnostic potential of the identified markers and constructed model must be validated through in vitro experiments and clinical trials to confirm their applicability in clinical settings.

Data availability

The datasets used in this study are all sourced from the public database Gene Expression Omnibus (GEO). The specific datasets used in our research included GSE87466, GSE92415, GSE38713, GSE75214. All datasets can be found by searching on the GEO website (<https://www.ncbi.nlm.nih.gov/geo/>). The use of these datasets adheres to the terms and conditions of the GEO database. The processing and analysis of all data were conducted as described in the methods section of this report. We confirm that all data used in this research are publicly accessible through the provided link, allowing other researchers to verify and conduct further studies.

Received: 2 March 2025; Accepted: 3 November 2025

Published online: 20 December 2025

References

1. Le Berre, C., Honap, S. & Peyrin-Biroulet, L. Ulcerative colitis. *Lancet (London England)*. **402** (10401), 571–584. [https://doi.org/10.1016/S0140-6736\(23\)00966-2](https://doi.org/10.1016/S0140-6736(23)00966-2) (2023).
2. Talley, N. J. et al. An evidence-based systematic review on medical therapies for inflammatory bowel disease. *Am. J. Gastroenterol.* **106** (Suppl 1), S2–S26. <https://doi.org/10.1038/ajg.2011.58> (2011).
3. Ananthakrishnan, A. N. Epidemiology and risk factors for IBD. *Nature reviews. Gastroenterol. Hepatol.* **12** (4), 205–217. <https://doi.org/10.1038/nrgastro.2015.34> (2015).
4. , G. B. D. I. B. D. C. The global, regional, and National burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet Gastroenterol. Hepatol.* **5** (1), 17–30. [https://doi.org/10.1016/S2468-1253\(19\)30333-4](https://doi.org/10.1016/S2468-1253(19)30333-4) (2020).

5. Piovani, D. et al. Environmental Risk Factors for Inflammatory Bowel Diseases: An Umbrella Review of Meta-analyses. *Gastroenterology*, 157(3), 647–659. (2019). <https://doi.org/10.1053/j.gastro.2019.04.016>
6. Danese, S. et al. Unmet medical needs in ulcerative colitis: an expert group consensus. *Dig. Dis. (Basel Switzerland)*. **37** (4), 266–283. <https://doi.org/10.1159/000496739> (2019).
7. Lee, J. M. & Lee, K. M. Endoscopic diagnosis and differentiation of inflammatory bowel disease. *Clin. Endoscopy*. **49** (4), 370–375. <https://doi.org/10.5946/ce.2016.090> (2016).
8. Mizushima, N. & Komatsu, M. Autophagy: renovation of cells and tissues. *Cell* **147** (4), 728–741. <https://doi.org/10.1016/j.cell.2011.10.026> (2011).
9. Mizushima, N. A brief history of autophagy from cell biology to physiology and disease. *Nat. Cell Biol.* **20** (5), 521–527. <https://doi.org/10.1038/s41556-018-0092-5> (2018).
10. Tang, X., Nagayach, A. & Wang, C. Microglial autophagy in neurogenesis: a new player in alzheimer's disease. *Neural Regeneration Res.* **19** (12), 2573–2574. <https://doi.org/10.4103/NRR.NRR-D-23-01962> (2024).
11. Jahng, J. W. S. et al. Iron overload inhibits late stage autophagic flux leading to insulin resistance. *EMBO Rep.* **20** (10), e47911. <https://doi.org/10.15252/embr.201947911> (2019).
12. Mokarram, P. et al. New frontiers in the treatment of colorectal cancer: autophagy and the unfolded protein response as promising targets. *Autophagy* **13** (5), 781–819. <https://doi.org/10.1080/15548627.2017.1290751> (2017).
13. Matsuzawa-Ishimoto, Y. et al. Autophagy protein ATG16L1 prevents necroptosis in the intestinal epithelium. *J. Exp. Med.* **214** (12), 3687–3705. <https://doi.org/10.1084/jem.20170558> (2017).
14. Larabi, A., Barnich, N. & Nguyen, H. T. T. New insights into the interplay between autophagy, gut microbiota and inflammatory responses in IBD. *Autophagy* **16** (1), 38–51. <https://doi.org/10.1080/15548627.2019.1635384> (2020).
15. Li, K. et al. Molecular comparison of adult and pediatric ulcerative colitis indicates broad similarity of molecular pathways in disease tissue. *J. Pediatr. Gastroenterol. Nutr.* **67** (1), 45–52. <https://doi.org/10.1097/MPG.0000000000001898> (2018).
16. Sandborn, W. J. et al. Subcutaneous golimumab induces clinical response and remission in patients with moderate-to-severe ulcerative colitis. *Gastroenterology* **146** (1), 85–e15. <https://doi.org/10.1053/j.gastro.2013.05.048> (2014).
17. Planell, N. et al. Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. *Gut* **62** (7), 967–976. <https://doi.org/10.1136/gutjnl-2012-303333> (2013).
18. Vancamelbeke, M. et al. Genetic and transcriptomic bases of intestinal epithelial barrier dysfunction in inflammatory bowel disease. *Inflamm. Bowel Dis.* **23** (10), 1718–1729. <https://doi.org/10.1097/MIB.0000000000001246> (2017).
19. The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* **47** (D1), D330–D338. <https://doi.org/10.1093/nar/gky1055> (2019).
20. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** (1), 27–30. <https://doi.org/10.1093/nar/28.1.27> (2000).
21. Hänzelmann, S., Castelo, R. & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7. <https://doi.org/10.1186/1471-2105-14-7> (2013).
22. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53** (D1), D672–D677. <https://doi.org/10.1093/nar/gkac909> (2025).
23. Kanehisa, M. Toward Understanding the origin and evolution of cellular organisms. *Protein Science: Publication Protein Soc.* **28** (11), 1947–1951. <https://doi.org/10.1002/pro.3715> (2019).
24. Fan, Y. et al. Edible bird's nest ameliorates dextran sulfate Sodium-Induced ulcerative colitis in C57BL/6J mice by restoring the Th17/Treg cell balance. *Front. Pharmacol.* **12**, 632602. <https://doi.org/10.3389/fphar.2021.632602> (2021).
25. Zhang, L. et al. Curcumin and Resveratrol suppress dextran sulfate sodium-induced colitis in mice. *Mol. Med. Rep.* **19** (4), 3053–3060. <https://doi.org/10.3892/mmr.2019.9974> (2019).
26. Pan, H. H. et al. Resveratrol alleviates intestinal mucosal barrier dysfunction in dextran sulfate sodium-induced colitis mice by enhancing autophagy. *World J. Gastroenterol.* **26** (33), 4945–4959. <https://doi.org/10.3748/wjg.v26.i33.4945> (2020).
27. Eaden, J. Review article: colorectal carcinoma and inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **20** (Suppl 4), 24–30. <https://doi.org/10.1111/j.1365-2036.2004.02046.x> (2004).
28. Li, T. et al. Formyl peptide receptor 1 mitigates colon inflammation and maintains mucosal homeostasis through the inhibition of CREB-C/EBP β -S100a8 signaling. *Mucosal immunology*, S1933-0219(24)00038–2. Advance online publication. (2024). <https://doi.org/10.1016/j.mucimm.2024.04.001>
29. Wu, J. et al. EFHD2 suppresses intestinal inflammation by blocking intestinal epithelial cell TNFR1 internalization and cell death. *Nat. Commun.* **15** (1), 1282. <https://doi.org/10.1038/s41467-024-45539-x> (2024).
30. Temby, M., Boye, T. L., Hoang, J., Nielsen, O. H. & Gubatan, J. Kinase signaling in Colitis-Associated colon cancer and inflammatory bowel disease. *Biomolecules* **13** (11), 1620. <https://doi.org/10.3390/biom13111620> (2023).
31. Soendergaard, C., Nielsen, O. H., Seidelin, J. B., Kvist, P. H. & Bjerrum, J. T. Alpha-1 antitrypsin and granulocyte colony-stimulating factor as serum biomarkers of disease severity in ulcerative colitis. *Inflamm. Bowel Dis.* **21** (5), 1077–1088. <https://doi.org/10.1097/MIB.0000000000000348> (2015).
32. Ma, Y. et al. CEBPB-mediated upregulation of SERPINA1 promotes colorectal cancer progression by enhancing STAT3 signaling. *Cell. Death Discovery*. **10** (1), 219. <https://doi.org/10.1038/s41420-024-01990-9> (2024).
33. Hua, H. et al. A proteomic classifier panel for early screening of colorectal cancer: a case control study. *J. Translational Med.* **22** (1), 188. <https://doi.org/10.1186/s12967-024-04983-5> (2024).
34. Pan, L., Zhang, L., Fu, J., Shen, K. & Zhang, G. Integrated transcriptome sequencing and weighted gene co-expression network analysis reveals key genes of papillary thyroid carcinomas. *Heliyon* **10** (7), e27928. <https://doi.org/10.1016/j.heliyon.2024.e27928> (2024).
35. Li, Q. et al. Identification of Serpin peptidase inhibitor clade A member 1 (SERPINA1) might be a poor prognosis biomarker promoting the progression of papillary thyroid cancer. *Life Sci.* **329**, 121938. <https://doi.org/10.1016/j.lfs.2023.121938> (2023).
36. Kwon, C. H. et al. Serpin peptidase inhibitor clade A member 1 is a biomarker of poor prognosis in gastric cancer. *Br. J. Cancer.* **111** (10), 1993–2002. <https://doi.org/10.1038/bjc.2014.490> (2014).
37. Lei, C. et al. Unraveling breast cancer prognosis: a novel model based on coagulation-related genes. *Front. Mol. Biosci.* **11**, 1394585. <https://doi.org/10.3389/fmolb.2024.1394585> (2024).
38. Chiu, H. W., Chen, C. H., Chang, J. N., Chen, C. H. & Hsu, Y. H. Erratum to: Far-infrared promotes burn wound healing by suppressing NLRP3 inflammasome caused by enhanced autophagy. *J. Mol. Med.* **94** (7), 821. <https://doi.org/10.1007/s00109-016-1436-x> (2016).
39. Halle, A. et al. The NALP3 inflammasome is involved in the innate immune response to amyloid-beta. *Nat. Immunol.* **9** (8), 857–865. <https://doi.org/10.1038/ni.1636> (2008).
40. Zhou, W. et al. Disulfiram with Cu2+ alleviates dextran sulfate sodium-induced ulcerative colitis in mice. *Theranostics* **13** (9), 2879–2895. <https://doi.org/10.7150/thno.81571> (2023).
41. Flood, B. et al. Altered expression of caspases-4 and -5 during inflammatory bowel disease and colorectal cancer: diagnostic and therapeutic potential. *Clin. Exp. Immunol.* **181** (1), 39–50. <https://doi.org/10.1111/cei.12617> (2015).
42. Shi, J. et al. Cleavage of GSDMD by inflammatory caspases determines pyroptotic cell death. *Nature* **526** (7575), 660–665. <https://doi.org/10.1038/nature15514> (2015).
43. Wang, K. et al. Structural mechanism for GSDMD targeting by autoprocessed caspases in pyroptosis. *Cell* **180** (5), 941–955e20. <https://doi.org/10.1016/j.cell.2020.02.002> (2020).

44. Xu, X. et al. Elevated levels of IRF1 and CASP1 as pyroptosis-related biomarkers for intestinal epithelial cells in crohn's disease. *Front. Immunol.* **16**, 1551547. <https://doi.org/10.3389/fimmu.2025.1551547> (2025).
45. Yao, Y. et al. Comprehensive analysis of autophagy-related clusters and individual risk model for immunotherapy response prediction in gastric cancer. *Front. Oncol.* **13**, 1105778. <https://doi.org/10.3389/fonc.2023.1105778> (2023).
46. Chu, G. et al. Identification of a novel Protein-Based signature to improve prognosis prediction in renal clear cell carcinoma. *Front. Mol. Biosci.* **8**, 623120. <https://doi.org/10.3389/fmolb.2021.623120> (2021).
47. Dong, F. et al. Identification of survival-related predictors in hepatocellular carcinoma through integrated genomic, transcriptomic, and proteomic analyses. *Biomed. pharmacotherapy = Biomedecine Pharmacotherapie.* **114**, 108856. <https://doi.org/10.1016/j.biopha.2019.108856> (2019).
48. Greig, F. H. & Nixon, G. F. Phosphoprotein enriched in astrocytes (PEA)-15: a potential therapeutic target in multiple disease States. *Pharmacol. Ther.* **143** (3), 265–274. <https://doi.org/10.1016/j.pharmthera.2014.03.006> (2014).

Author contributions

All authors contributed to the study conception and design. Data acquisition, bioinformatics analysis, and the drafting of this section were completed by X.X.C., Z.S., M.Z., L.H. and M.S. Experimental validation and the corresponding write-up were performed by S.M. L.Z. and H.S. were involved in the final development of the project and manuscript preparation. All authors reviewed and commented on previous versions of the manuscript. All authors have read and approved the final manuscript.

Funding

This study was supported by the Jiangsu Leading Talents in Traditional Chinese Medicine (SLJ0306) and the Jiangsu Provincial Administration of Traditional Chinese Medicine (2022WTZX04). The funding body had no influence on the study design, data collection, analysis, and manuscript writing.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

We confirm that all methods in this study were conducted in accordance with relevant guidelines and regulations. The Animal Ethics Committee of Nanjing University of Chinese Medicine approved the animal experiments (Application Number: 2023DW-082-01).

Consent to participate

All datasets used in this study were obtained from the Gene Expression. Omnibus public database, which provides anonymized data that had been collected with the necessary ethical approvals and participant consent in the original studies. As such, this study complies with all ethical guidelines, and no further consent was needed from participants.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-27363-5>.

Correspondence and requests for materials should be addressed to H.S. or L.Z.

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

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

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

© The Author(s) 2025