Identification and immunological characterization of cuproptosis-related molecular clusters in ulcerative colitis

Yunfei Pu[1], Xianzhi Meng[2*] and Zhichen Zou[1]

Abstract

Background Ulcerative colitis is one of the two main forms of inflammatory bowel disease. Cuproptosis is reported to be a novel mode of cell death.

Methods We examined clusters of cuproptosis related genes and immune cell infiltration molecules in 86 ulcerative colitis samples from the GSE179285 dataset. We identified the differentially expressed genes according to the clustering method, and the performance of the SVM model, the random forest model, the generalized linear model, and the limit gradient enhancement model were compared, and then the optimal machine model was selected. To assess the accuracy of the learning predictions, the nomogram and the calibration curve and decision curve analyses showed that the subtypes of ulcerative colitis have been accurately predicted.

Results Significant cuproptosis-related genes and immune response cells were detected between the ulcerative colitis and control groups. Two cuproptosis-associated molecular clusters were identified. Immune infiltration analysis indicated that different clusters exhibited significant heterogeneity. The immune scores for Cluster2 were elevated. Both the residual error and root mean square error of the random forest machine model had clinical significance. There was a clear correlation between the differentially expressed genes in cluster 2 and the response of immune cells. The nomogram and the calibration curve and decision curve analyses showed that the subtypes of ulcerative colitis had sufficient accuracy.

Conclusion We examined the complex relationship between cuproptosis and ulcerative colitis in a systematic manner. To estimate the likelihood that each subtype of cuproptosis will occur in ulcerative colitis patients and their disease outcome, we developed a promising prediction model.

Keywords Cuproptosis, Ulcerative colitis, Prediction model, Molecular clusters, Machine learning

Introduction

Ulcerative colitis (UC) was first discovered in 1859, and it is one of two main types of inflammatory bowel disease (IBD) [1]. It is a chronic immuno-mediated IBD, though its exact pathogenesis is not known. Prolonged inflammation will damage gastrointestinal function, such as causing abdominal pain, bleeding and other symptoms [2]. In addition, diagnosis of digestive system diseases and evaluation of the prognosis of patients through inflammatory indicators and infiltration levels of immune cells [3]. For

*Correspondence:
Xianzhi Meng
3195@hrbmu.edu.cn; mengxianzhi@hrbmu.edu.cn
[1] The First Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang, China
[2] Department of Minimally Invasive Biliary Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin 150000, Heilongjiang, China

Pu *et al. BMC Gastroenterology*   (2023) 23:221

Page 2 of 14

example, Posul et al. 2015found that the ratio of neutrophils to lymphocytes could predict whether ulcerative colitis was in the active phase [4].

A number of factors have been linked to this condition, including dysregulated immune response, altered microflora in the gut, genetic susceptibility, and environmental factors [5]. Globally, the prevalence is high in developed countries, and is rapidly increasing in newly industrialized countries [6, 7]. Conley et al., 2017 demonstrated by 2025, the population is expected to reach 30 million worldwide [8]. Healthcare and social costs are significant as a result of the excessive morbidity and mortality rate caused by UC [9–11]. Abraham and Kane., 2012 demonstrated in spite of having the same phenotype, extreme early-onset IBD may have a different pathophysiology from IBD of earlier onset [12, 13]. As IBD incidence increases in developing countries and phenotypes are similar between Asians and Westerners, ethnicity appears not to be an important factor in UC. Due to the phenotypic similarity of UC across the globe, environmental changes may also be responsible for the observed epidemiological trends. Most patients with UC suffer from hemorrhagic diarrhea, and the clinical presentation is heterogeneous. Only 30% and 15% of patients have aggressive or generalized colitis, respectively. Solberg et al., 2009 reported that half of the patients may not respond to pharmacological treatment, resulting in a more complicated disease course [14]. UC is therefore formed as a result of different etiologies, which influence the course and severity of subsequent illness [15–17]. It is unfortunate that UC lacks a satisfactory treatment due to its clinical heterogeneity and complexity of pathological types. Despite the increasing association of biomarkers with UC, these results aren't convincing. In order to enhance clinical credibility, we extended the study of UC to the molecular subtype.

Researchers have previously attempted to define heterogeneity and treat UC [18]. Copper ion homeostasis in the human body is maintained by copper absorption, transport, and excretion, according to a recent study. Tang et al., 2022 reported that copper was the cause of a new form of cell death, called copperocytosis, which was related to the tricarboxylic acid cycle metabolism disorder [19, 20]. This mode of death is not completely independent of other regulatory modes of death, suggesting a close connection. Increasing evidence suggests that unbalanced copper homeostasis can affect tumor growth and induce tumor cell death. Related studies suggest that inhibition of mitochondrial pyruvate carrier and electron transport chain activity can mitigate the damage caused by cuproptosis [21–23]. Moreover, the infiltration of multiple immune cells in inflammatory diseases may be regulated through cuproptosis. Zhao et al., 2022 suggested that copper death led to excessive survival or proliferation of multiple immune cells in the synovial tissue of rheumatoid arthritis [24]. While copper plays a vital role in tumor immunity and cancer therapy, the induction of cell death in non-tumor is less well studied. In order to provide more personalized treatment for UC, it is imperative to identify more appropriate molecular clusters. Various disease mechanisms have not been studied in depth for the mechanism of non-tumor cuproptosis. Thus, cuproptosis genetic characteristics were used to identify UC subtypes.

This study was the first to systematically examine the differential expression and immune signature of cuproptosis-related genes in normal and UC individuals. Using 19 cuproptosis-related genes (CRGs) expression profiles, we divided 86 UC patients among 254 IBD patients into two groups, and found differences in immune cells between the two groups. By using the WGCNA algorithm, the differentially expressed genes (DEGs) between them, and the functional pathways of the differentially expressed genes were further explained. Additionally, we compared multiple machine learning algorithms to construct a special learning prediction model that presents different molecular clusters of UC-associated patients. Validation of the predictive model was conducted using decision curve analysis, nomographic analysis and calibration curve analysis. And there are two random forest models for further verification.

Materials
Preprocessing and acquisition of data
Related microarray data sets are obtained from GEO(www.ncbi.nlm.nih.gov/geo) including one experimental group GSE179285 data set, two verification groups GSE107597 data set and GSE92415 data set [25]. There were 31 healthy samples and 55 UC colon tissue samples in the trial GSE179285 datasets (GPL6480 platform), while the validation group GSE107597 datasets (GPL15207 platform), which included rectal tissue from 44 normal subjects and 75 UC (13–56 years old) samples, and the GSE92415 datasets (GPL13158 platform), including 21 normal and 162 UC (19–77 years old) colon tissue samples. Once the samples were downloaded, these raw data were collated and summarized in the next step. We used the Robust Multiarray Average (RMA) method ("affy" R package). The resulting expression measurements of the corresponding gene were estimated by RMA for all probes of the gene. The model generated by this algorithm will produce an estimate of the gene signal that takes into account the probe effect.

Pu *et al. BMC Gastroenterology*   (2023) 23:221

Page 3 of 14

The infiltration of immune cells was analyzed

We use CIBERSORT algorithm (https://cibersort.stanford.edu/) and signature matrix of gene expression data to analyze 22 kinds of immune cells corresponding relations, and *P* values were calculated by the CIBERSORT for each. We considered immune cell fractions accurate when they had $P < 0.05$. The total number of the 22 immune cells per sample is identical [26].

CRG and immune cell infiltration correlation analysis

We analyzed the percentage relationship between CRG and immune cell expression, in order to further demonstrate the link between CRGs and UC-related immune cell properties. $P < 0.05$ indicates relevant research significance, according to the spearman correlation coefficient. Lastly, plot results used the "corrplot" R package (version 0.92).

Patients with UC are clustered unsupervised

As a result of the previous report, a total of 12 CRGs were obtained [19, 27]. We used the K-means algorithm for 1,000 iterations and the consensus cluster package for unsupervised cluster analysis to classify 55 UC samples into different clusters. The consensus matrix, consistent cluster score (> 0.9) and cumulative distribution function (CDF) curve were evaluated to select the optimal number of clusters, and the maximum number of clusters was $K = 9$.

Gene set variation analysis (GSVA) analysis

In order to show the differences of gene sets among different CRGs clusters,The method of Wu et al. 2021 reported that the R package of "GSVA" (version 2.11) was used for analysis [28]. For further GSVA analysis, we used the database known to be symbols from the MSigDB website "c2.cp.kegg.symbols" and "c5.go. Symbols". By comparing GSVA scores between different CRG clusters, we used the "limma" R package (version 3.52.1) to evaluate the score, where a score greater than 2 indicates that the study is meaningful.

Weighted gene co-expression network analysis (WGCNA)

We used the R package "WGCNA" (version 1,700.3) to identify co-expression modules [29]. In WGCNA analysis, 25% of the genes with the strongest correlation were selected for accuracy analysis. Firstly, a weighted adjacency matrix is set up using the optimal soft power, and then the matrix is transformed into a topological overlapping matrix (TOM). The next step is to analyze the module obtained when the minimum module value of the TOM dissimilarity of the hierarchical clustering tree algorithm is 100. The characteristic genes in each module represent overall gene expression in that module, and each module has a random color. The module significance (MS) index is what we need, which shows the relationship between a specific module and the disease. The term gene significance (GS) is used to describe the association between a gene and a clinical phenotype.

Multiple machine learning methods are used to construct a predictive model

Random forest model (RF), support vector machine model (SVM), generalized linear model (GLM), and extreme gradient enhancement (XGB) are the machine learning models we built on different clusters of CRGs. RF is a prediction and regression technique that uses randomization and substitution to represent the technique's precise predictions [30]. SVM can describe the value of our predictions as a learning model to improve the value of our predictions [31]. GLM consists of linear components and link functions, which can correctly evaluate the relationship between independent or mutually characteristic normal distributions [32]. The XGB has a predictable and enhanced set of numbers, increasing the accuracy of our prediction models. DEGs specific to each cluster were chosen as explanatory variables, and the corresponding variables are different clusters. We randomly classified the 55 UC samples into two training and validation sets (55%, $N = 30$). Through grid search, the insertion symbol package automatically adjusts the parameters in these machine learning models. All these models were evaluated via fivefold cross validation. To interpret and visualize the distribution of residuals and the importance of features between the aforementioned four machine learning models, the "DALEX" package (version 2.4.0) was used. Use the "pROC" R package (version 1.18.0) to visualize the area under the ROC curve [33]. Thus, the top five variables that were associated with UC were determined through a machine learning model. Liu et al. 2022 reported that a Nomogram was created through the "nomogramEx" R package [34].

Analyses of independent validation

Finally, the two validation group data sets GSE107597 and GSE92415 were further analyzed using the ROC curve of the "pROC" R package to evaluate the UC and no-UC predictions that we validated our experimental group model.

Results

Patients with UC have dysregulated cuproptosis regulators and activated immune responses

To clarify the biological functions of cuproptosis regulators in the occurrence and progression of UC. A
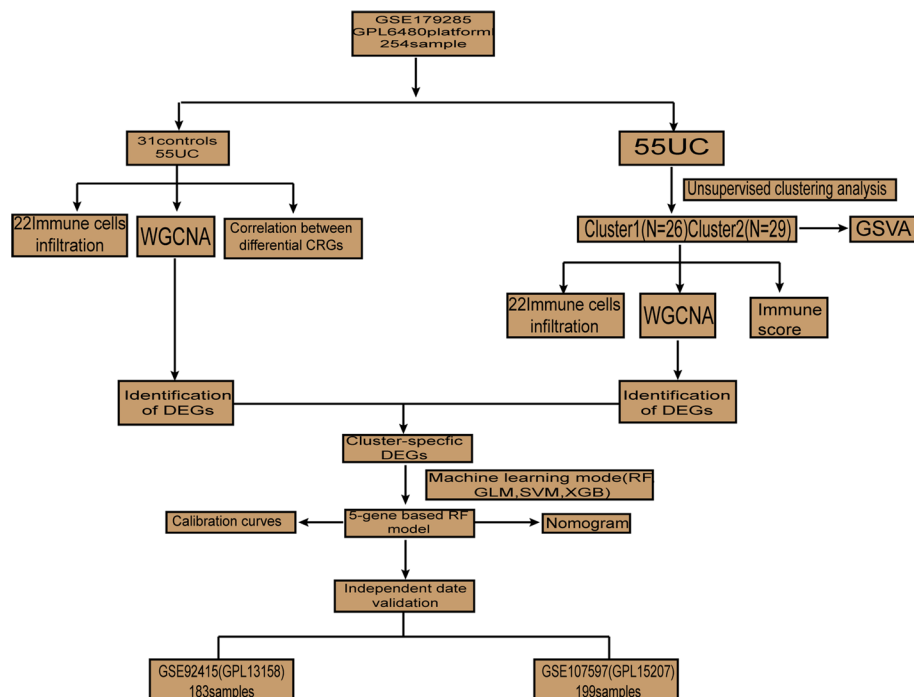
Pu *et al. BMC Gastroenterology*   (2023) 23:221

Page 4 of 14

Fig. 1 The flow-process diagram

detailed flow chart of the study process was shown in Fig. 1. We identified 12 CRGs as differentially expressed cuproptosis genes in the evaluation prediction model. These include DLST, DBT, GLS, FDX1, MTF1, NFE2L2, PDHA1, DLD, LIPT2, LIPT1, LIAS, PDHB expression levels that are higher than those in non-UC controls (Figs. 2A–C). Afterward, because we wanted to determine whether the cuproptosis-related regulatory genes have some clinical significance in UC patients, a correlation analysis was done to study the differential expression of CRGs (Fig. 2D). Surprisingly, there was a strong synergistic effect between some cuproptosis modulators, such as PDHA1 and DLD. Conversely, there is also some antagonism between LIPT 1 and MTF 1. Further study found that DBT and DLD were closely related to different regulators. In addition, a circle graph of a gene perfectly shows the differential graph of CRGs (Fig. 2E).

According to the CIBERSORT algorithm, we are shown the proportion of 22 immune cells, and we determine the difference in the immune system in UC and non-UC (Fig. 2F). As a result, UC patients presented higher levels of infiltrating CD4T cells, mast cells that were resting, mast cells that were activated, and neutrophils (Fig. 2G), indicating that UC may be caused by immune system alternations. Macrophage M2 showed a positive relationship with epidermal growth regulators (Fig. 2H). The CRGs appear to play

an important role in both the immune infiltration of UC as well as its molecular regulation.

The identification of clusters of cuproptosis in UC

We used a consensus clustering algorithm to group 55 UC samples by genes of 12 CRGs to identify cuproptosis-related expression patterns. When k=2, cluster numbers were most stable, and CDF curves ranged between 0.2 and 0.6 (Figs. 3A,B). The CDF curve shows the degree of difference in correlation between different k, while k has a minimum value of 2 and a maximum value of 9 (Fig. 3C). Moreover, Because the consistency score for each subtype is greater than 0.9, a k of 2 is the value of our best chosen molecular cluster (Fig. 3D). Based on the Principal Component Analysis (PCA) analysis, these two clusters differed significantly (Fig. 3E).

Comparing cuproptosis clusters based on epidermal cell-related factors, immune cell infiltration characteristics, and pathway annotation

Our first step was to assess the differential expression of the 12CRGs between Cluster1 and Cluster2 to determine their molecular characteristics. Both cuproptosis patterns exhibit distinct expression landscapes of CRGs (Fig. 4A). Cluster 1 of cuproptosis exhibited high expression levels of PDHA1, NFE2L2, FDX1, LIAS, DLD, PDHB, DBT, GLS, and DLST (Fig. 4B). Further, we
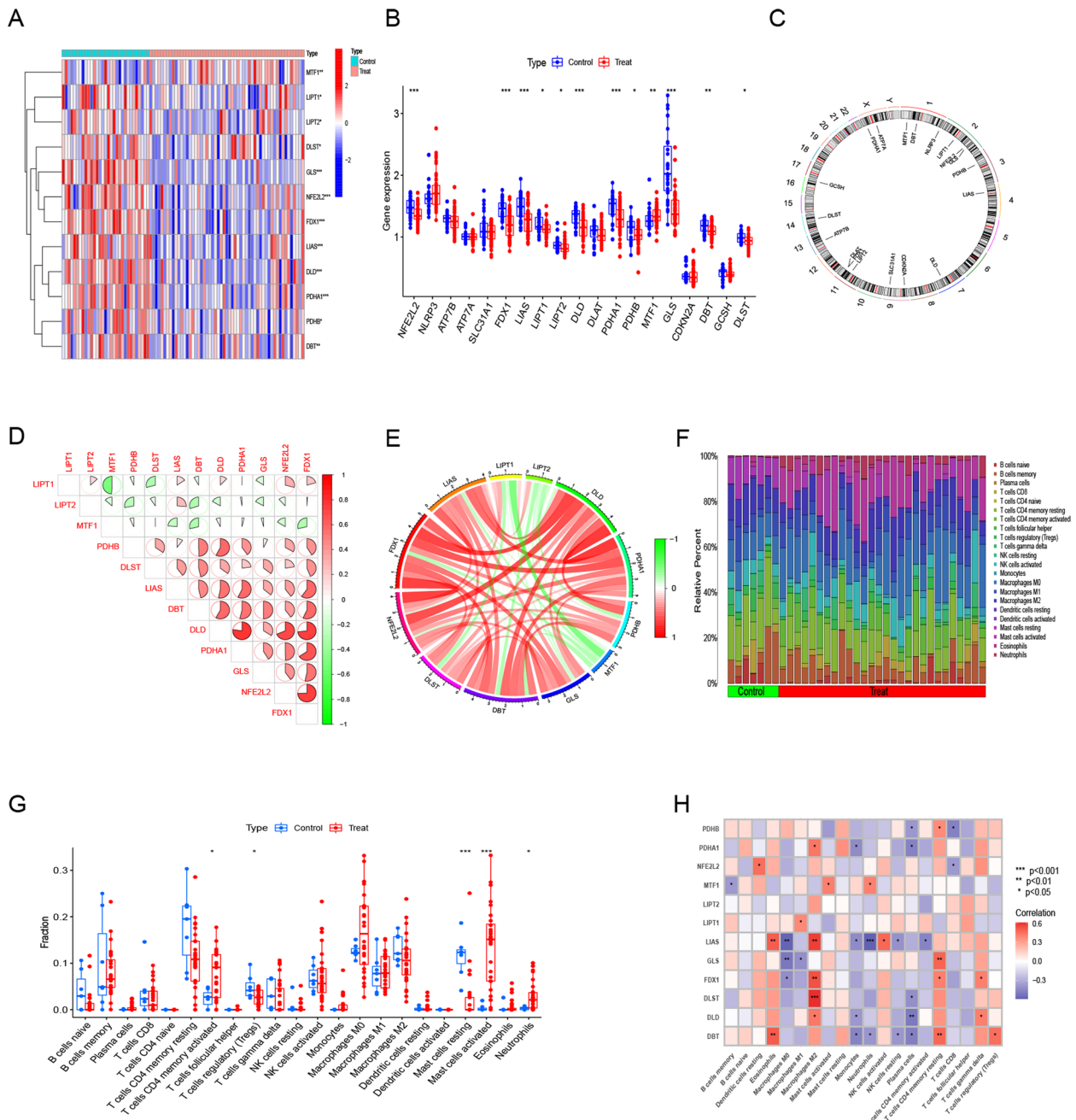
Pu *et al. BMC Gastroenterology*   (2023) 23:221

Page 5 of 14

Fig. 2 Differential expression of the CRGs in UC patients, heatmap of the expression correlation of the 12 CRGs **A**. Boxplots showed the expression of 12 CRGs between UC and no-UC, * *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001 **B**. The position of the 12 CRGs on the chromosome was shown **C**, correlation analysis the 12 CRGs using the area of the pie chart **D**. Network diagram of the relationships of the 12 genes in the CRGs **E**. Showed the richness of 22 immune cell infiltration between UC and no-UC **F**. Boxplots showed differences between UC and no-UC, * *p* < 0.05, *** *p* < 0.001 **G**. Correlation analysis of immune cell infiltration in 12 CRGs **H**

examined the immune-infiltrating cells and the immune environment between the two molecular clusters (Fig. 4C). While Cluster1 showed a greater proportion of resting and gamma delta T cells, Cluster2 showed a greater proportion of monocytes, macrophages M0,

and mast cells activated (Fig. 4D). We used GSVA to determine the pathway differential expression levels of cluster-specific DEGs between Cluster1 and Cluster2. In Cluster2, oxidative phosphorylation, citrate cycle, TCA cycle, limonene, pinene degradation, peroxisome
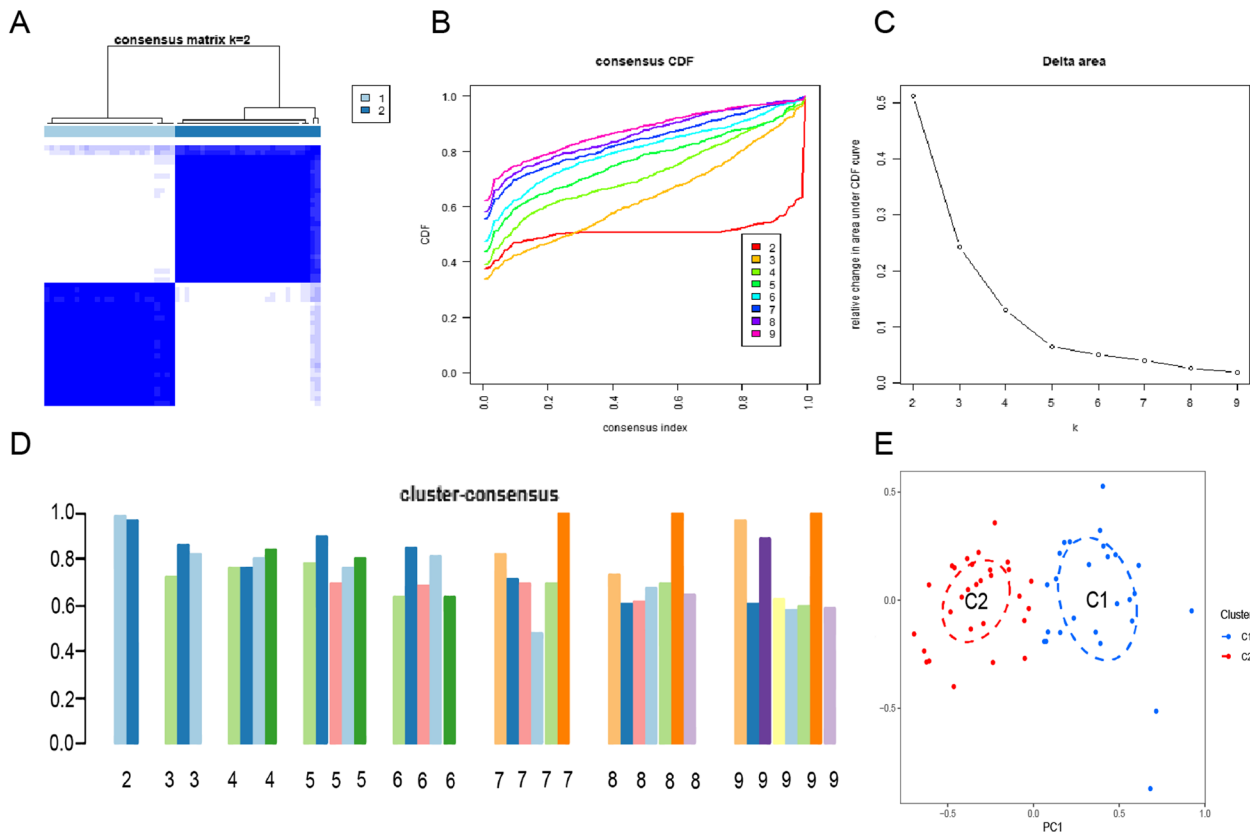
Fig. 3 For the identification of cuproptosis-related molecular clusters in UC, a consensus clustering matrix with k=2 was chosen and was the best molecular cluster classification **A**. CDF delta area curves **B**. Scores for consensus clustering **C**. A heatmap of non-negative matrix **D**. A PCA analysis of the distribution of the two molecular clusters **E**,

and metabolism signaling were enhanced, while Cluster1 had an upregulation of glycan biosynthesis, immune responses, cytokine receptors and Notch signaling activity. (Fig. 4E).

Screening of gene modules and construction of co-expression networks

We created gene co-expression networks and modules, selected the strongest 25% of genes in the WGCNA analysis for accuracy results according to the gene expression variance of GSE179285, and set the scale-free R2 to 0.9 and soft power set to 11, to select the co-regulation modules of the required genes we need (Fig. 5A). Co-expression networks and modules were established using

the WGCNA algorithm and also for the normal and UC subjects. A weighted adjacency matrix was constructed using the optimal soft power and transformed into TOM, where each module was randomly assigned a color, a total of 6 modules concerning coexpression of clinical features in the UC and no-UC studies (Figs. 5B–D). These genes in the 6 color modules were continuously applied for analyzing the similarity and adjacency of module-clinical features (UC and noUC) co-expression. The blue module is the most expressive, including 54 genes in the blue module (Fig. 5E). A weighted adjacency matrix was constructed using optimal soft power and transformed into a topological overlap matrix (TOM), where each module was randomly assigned a color. A total of 6 modules were

(See figure on next page.)

Fig. 4 It showed the molecular characteristics and immune characteristics between the two molecular clusters associated with cuproptosis. Heatmap of the expression signature between the 12 CRGs and the two molecular clusters **A**. Boxplot exhibited the expression of the 12 CRGs between the two molecular clusters, **$p < 0.01$, ***$p < 0.001$ **B**. Differences in richness of 22 immune cells in the two molecular clusters **C**. Boxplots showed the infiltration richness of the immune cells between the two molecular clusters * $p < 0.05$, ** $p < 0.01$ **D**. There are some differences in the hallmark pathway activities between the two molecular clusters ordered by the different t-value of the GSVA method **E**
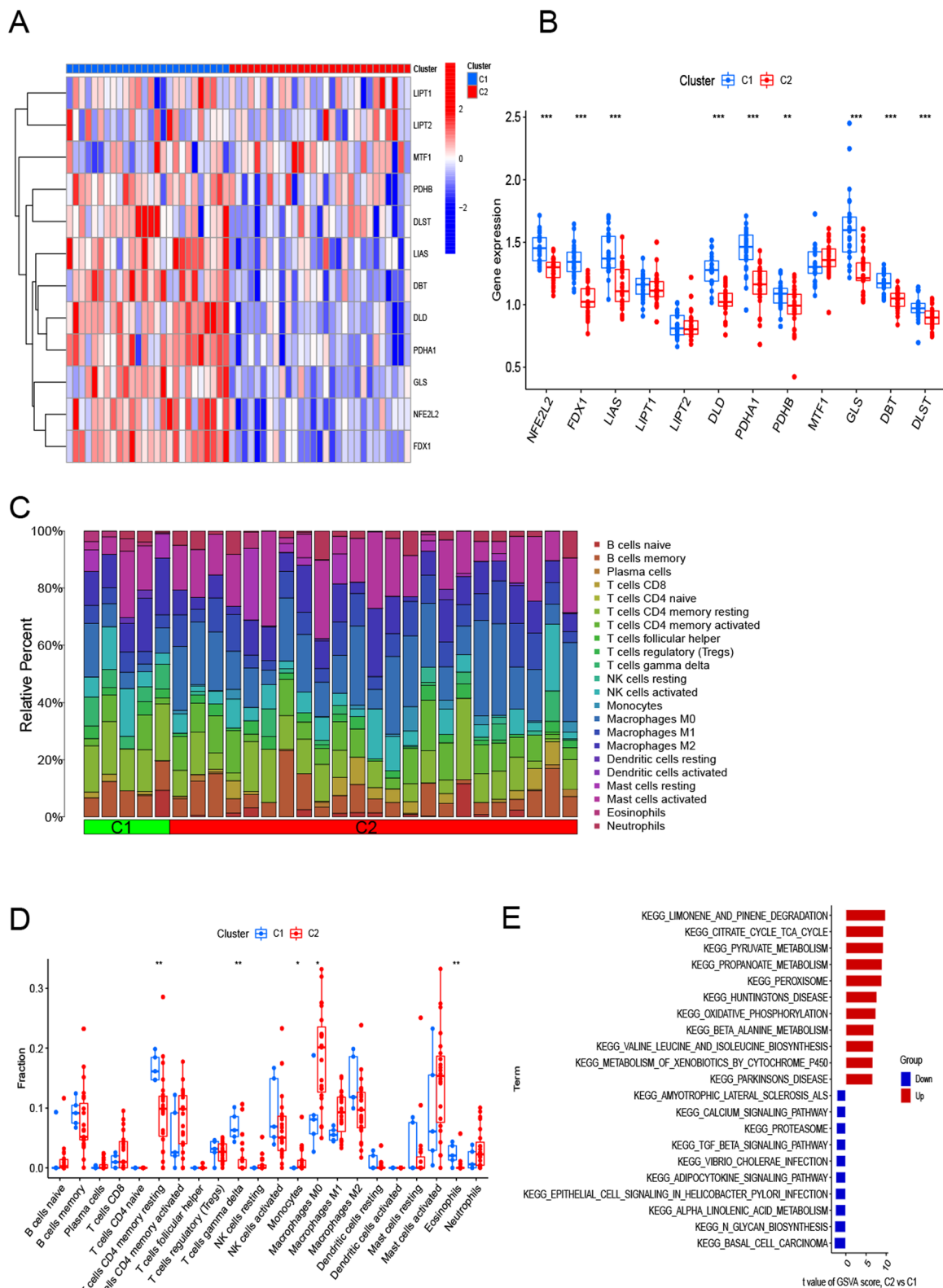
Fig. 4 (See legend on previous page.)

Pu *et al. BMC Gastroenterology*   (2023) 23:221
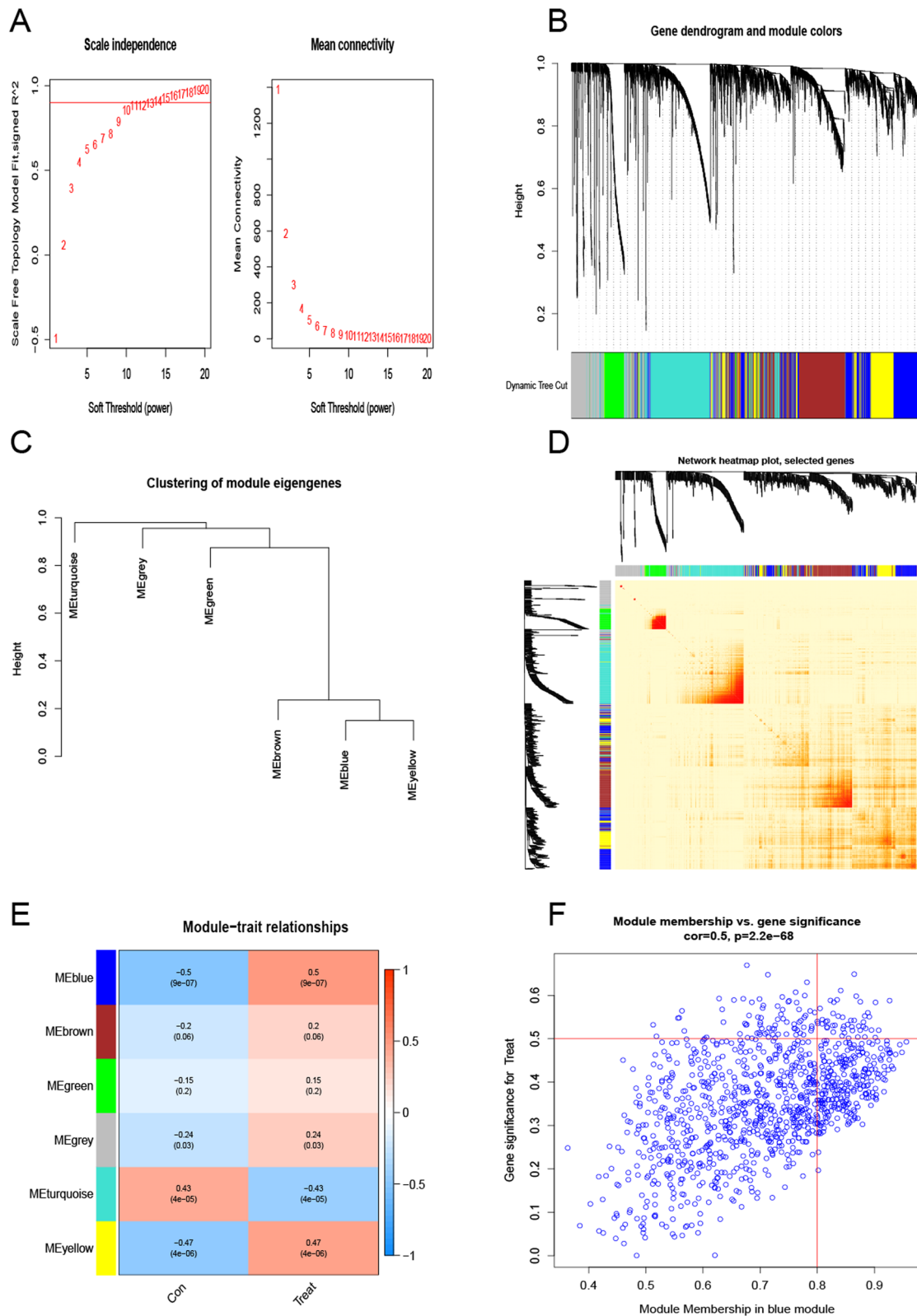
Page 8 of 14

Fig. 5 Gene co-expression network involved in gene expression in UC. Soft threshold power selection **A**. In the co-expression dendrogram, different colors represent different gene co-expression modules **B**. Showed the classification of gene module features **C**. Heatmap of the 6 genes representing the modules **D**. Rows represented modules, columns represented gene signature modules of clinical status and correlation analysis of clinical status **E**. Members of the module in blue were associated with genes in UC **F**
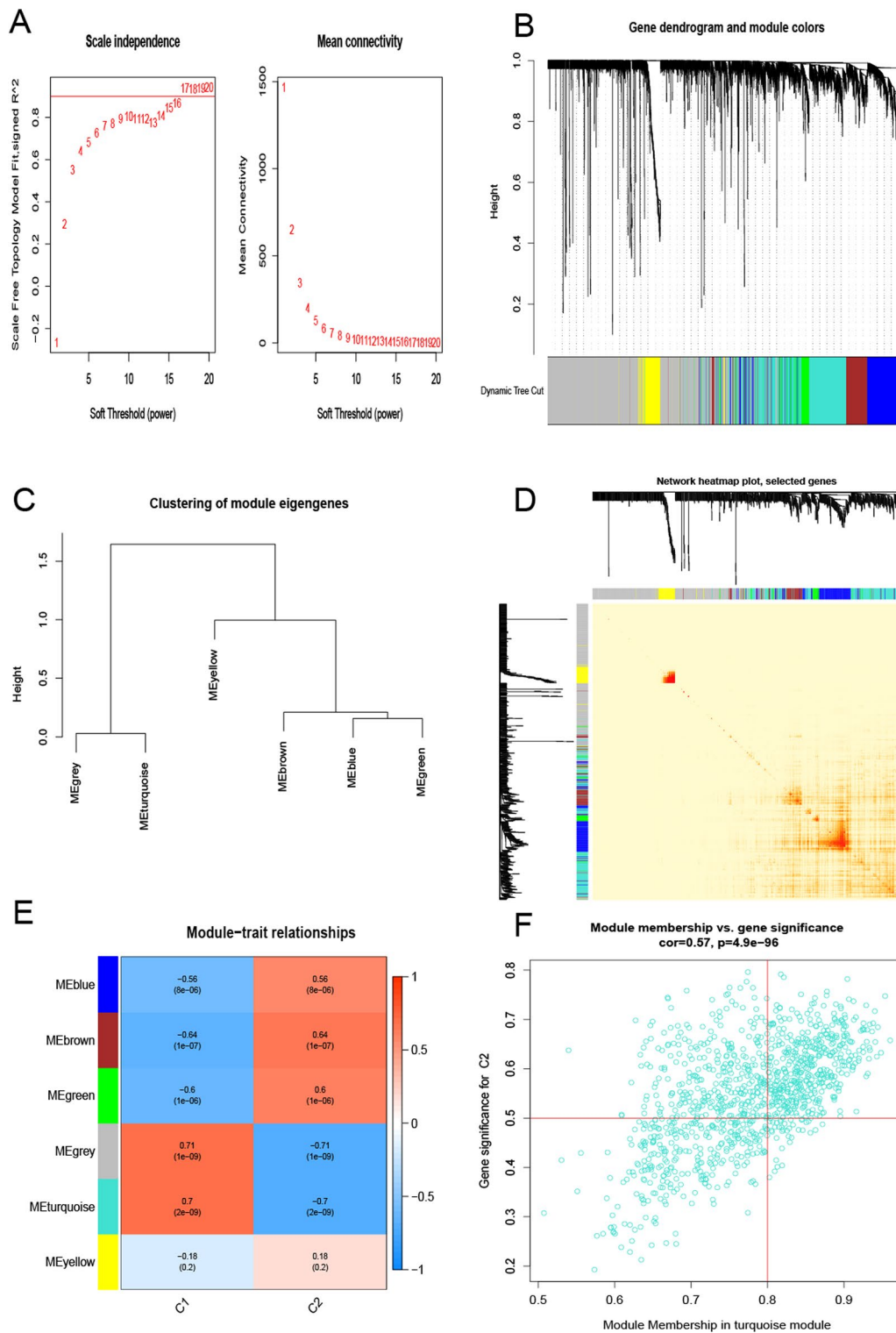
Fig. 6 Co-expression network of expressed genes between the two cuproptosis-related molecular cluster. Soft threshold power selection **A**. In the co-expression dendrogram, different colors represent different gene co-expression modules **B**. Showed the classification of gene module features **C**. Heatmap of the 6 genes representing the modules **D**. Rows represented modules, columns represented gene signature modules of clinical status and correlation analysis of clinical status **E**. Scatter plot of gene correlation of members of the turquoise module with molecular cluster 2 **F**

co-expressed clinical features. Furthermore, blue modules are associated with module-related genes (Fig. 5F).

Then we used the WGCNA algorithm to analyze the genes of the cuprotosis molecule clusters as β value of 16 and R 2 value of 0.9 (Fig. 6A), and also obtained 6 color modules containing a total of 4898 genes. The correlation heatmap concerning TOM also reveals significantly correlated genes (Figs. 6B–D). Between the two molecular clusters, in the relationship analysis of the relevant clinical characteristic modules, the UC cluster and the turquoise module (629 genes) were highly correlated (Fig. 6E). Correlation analysis showed that the turquoise color module gene was our best choice, and it had a significant correlation with the genes between the selected modules (Fig. 6F).

Identifying cluster-specific DEGs
We performed analysis of the analyzed UC and noUC modules and obtained the associated crossover genes with 49 specific DEGs (Fig. 7).

Machine learning model construction and assessment
Using the analysis to obtain the expression profiles of 49 cluster-specific DEGs, we used four machine learning models to obtain isoform genes called high predictive value, and these four learning models were RF, SVM, GLM, and XGB. The four models were constructed with the 'DALEX' package, and the residual expression distribution was plotted for each model. The RF learning model has the least residual expression (Figs. 8A,B). We selected the top 10 variables based on root mean square error (RMSE) (Fig. 8C). We obtained relevant

ROC curves based on the five-fold cross-validation of the learning algorithm of the machine learning model, with AUC = 0.910 in RF, AUC 0.861 in SVM, AUC 0.639 in GLM, and AUC 0.875 in XGB (Fig. 8D).

According to the results of AUC, we chose RF as the model that shows the UC of different clusters. As a result of the RF model, the top five most important variables were selected as predictor genes (BHLHE40, CDK20, SCIN, VIL1, and DDC). The RF model was further evaluated by developing a nomogram to predict the crush disease group of the 86 patients with UC (Fig. 9A). We used the calibration curve and the decision curve to analyze the prediction value of the nomogram model, and obtained that the UC cluster prediction and the actual risk error are very small (Fig. 9B), and the decision curve analysis further proves the accuracy of our nomogram model (Fig. 9C). We then used two validation groups, the UC tissue database and the no-UC tissue database, to validate our 5-gene prediction model. In the GSE92415 dataset, the AUC value for the 5-gene prediction model was 0.976, while in the GSE107597 dataset, it was 0.797 (Figs. 9D,E).

Discussion
We studied for the first time the expression of copper cuproptosis-related regulators in the colon tissues of UC and no-UC. At the same time, we also showed the specific role of cuproptosis-related gene phenotype and the immune microenvironment in UC. There were more dysregulated CRGs in UC patients than in normal individuals, suggesting that CRGs play an important role in UC. There were significant synergistic or antagonistic effects
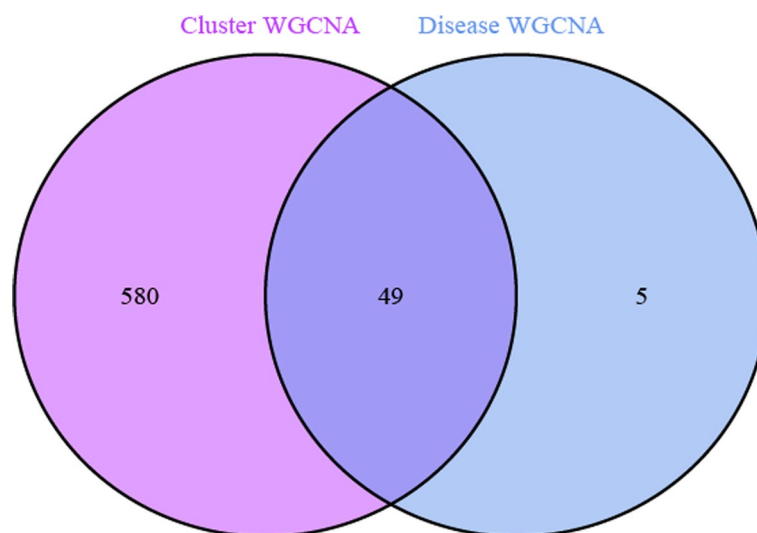

Fig. 7 DEGs and biological characteristic differences between the two molecular clusters. Intersection between cuproptosis-related molecular cluster genes and module-related genes in the GSE179285 dataset
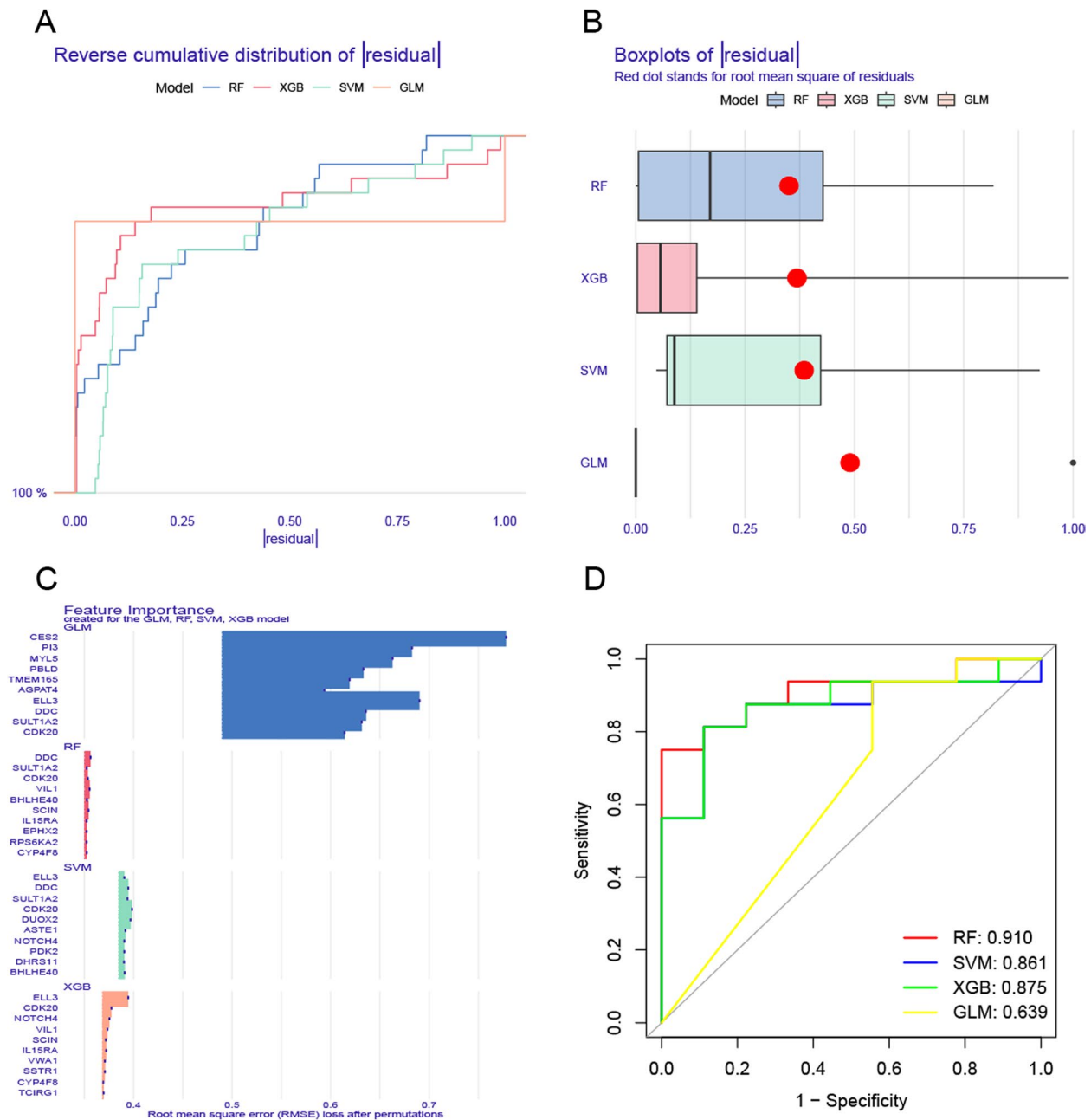
Fig. 8 Construction and evaluation of the four machine models: RF, SVM, GLM, and XGB. Cumulative residual distribution of the four machine learning models **A**. Boxplots exhibited the magnitude of the residuals across the respective machine learning models, red dots indicated the root mean square of residuals (RMSE) **B**. Feature expression of the four machine-learning models **C**. The fivefold cross ROC analysis curve validates the RF, SVM, GLM, and XGB machine learning models **D**

of some cuproptosis modulators in patients with UC, as evidenced by CRG interactions.

In this study, we found that there were higher infiltration levels in UC patients with T cells CD4 memory activated, mast cells activated, and neutrophils. UC is a common disease of the immune-mediated digestive system. The immune system influenced the disease

development by modulating the infiltration of immune cells in the diseased area [35]. T cell immunotherapy had an outstanding advantage in the treatment of ulcerative colitis [36]. Furthermore, the important role of mast cells in resistance to bacterial infection and inflammation. The presence of microbial-gut-brain axis disease with mast cells as the main component in UC patients [37].
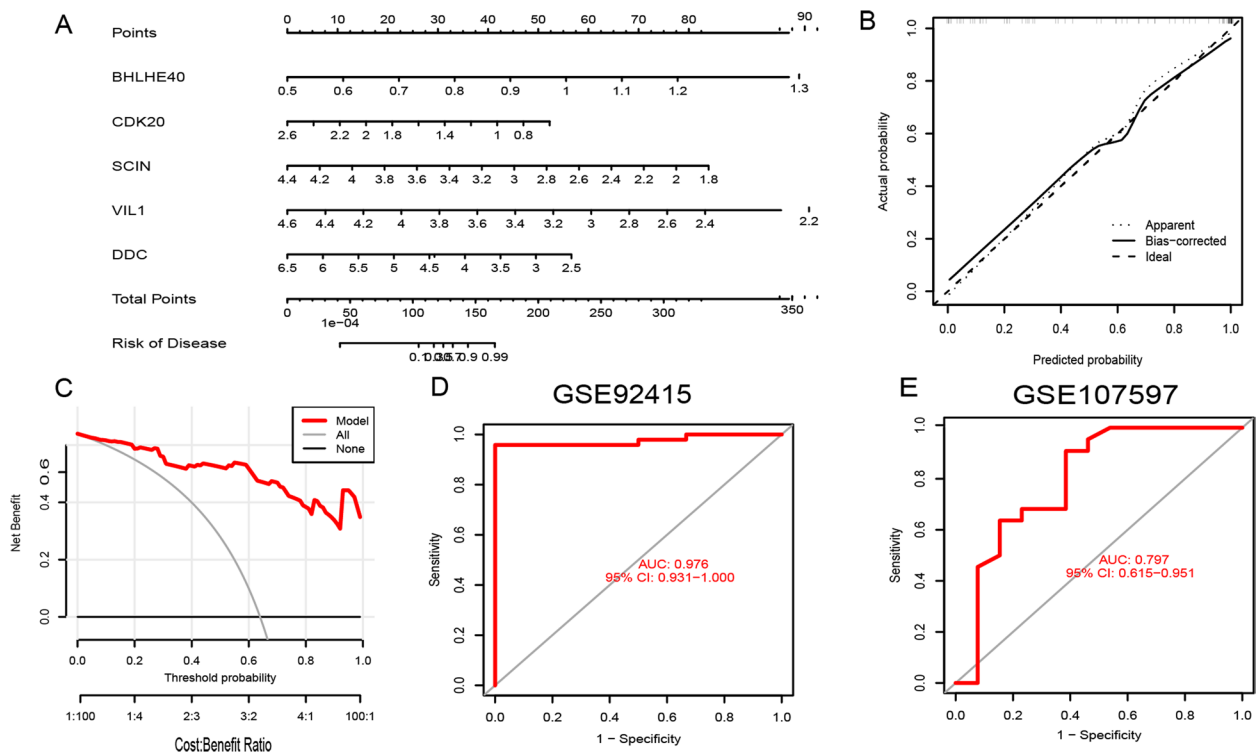
Pu *et al. BMC Gastroenterology*   (2023) 23:221

Page 12 of 14

Fig. 9 Further validation of the 5-gene-based RF model. Risk prediction of UC clusters used a nomogram of the 5-gene-based RF model **A**. To assess the predictive ability of the nomogram model, calibration curves **B** and DCA were made **C**. In the GSE92415 **D** and GSE107597 **E** datasets. ROC analysis used the final model genes based on fivefold cross-validation

Neutrophils participated in disease progression through the secretion of proinflammatory cytokines and it was an abnormal infiltrating immune cell in UC [38].

Furthermore, we used unsupervised cluster analysis to analyze expression landscapes of CRGs in UC patients to identify two distinct cuproptosis-related clusters. The accuracy of our typing was further determined by PCA analysis. In Cluster1, immune scores were elevated and immune infiltration was relatively higher. The DEGs identified by Cluster 2 were primarily enriched in limonene and pinene degradation, citrate cycle, TCA cycle, and pyruvate metabolism processes, while The DEGs identified by Cluster 1 were primarily enriched in infection and immune-related pathways. For example,there were helicobacter pylori infection, vibrio cholerae infection, amyotrophic lateral sclerosis als. This indicates that having a higher degree of immune infiltration occurs in the Cluster1.

With the growing population and increasing databases, machine learning models of demographic and imaging indicators are increasingly used to analyze the clinic [39]. We therefore compared the predictive performance of the four machine learning models in UC, using the expression of the DEGs concerning UC, and selected RF as our optimal model. RF (AUC=0.9829)

indicates a high clinical significance for the machine learning model that we chose in the prediction study of UC. As a next step, we constructed an RF model based on five genes and five relevant variables (BHLHE40, CDK20, SCIN, VIL1, and DDC). Zhang et al., 2020 demonstrated that a gene called BHLHE40 was expressed exclusively in patients with UC, which acted as a crucial regulator of colon cancer cell growth. This suggestted that BHLHE40 might be a useful treatment for UC patients. However, its role in UC had not been studied or mentioned [40]. Lai et al., 2020 demonstrated that there was increasing research suggesting CDK20 was a key controlled of cell cycle checkpoints controlling cell proliferation and involved in the development of multiple cancers. He further speculated that this gene was a potential therapeutic target for inflammatory diseases, but it needed further verification [41, 42]. Chen et al., 2021 reported that PDE4D was crucial to the development of intestinal diseases by phosphorylating and activating the intestinal cell kinase.Given the high level of expression of SCIN in cancer tissues, we speculate that its expression level may also affect the diagnosis and prognosis of UC.Studies have found that SCIN also affects the colon's cellular immune infiltration. He did not do any further research [43, 44]. We

further studied and improved the predictability for UC. Salewski et al., 2022 reported that VIL 1 involved in the tumor-associated immune infiltratio [45]. However, they did not mention the non-tumor UC patients, and we initially verified that colon inflammation was associated with the expression of VIL 1 protein. Artemaki et al., 2020 reported that positive correlation existed between DDC protein expression levels and unfavorable CRC prognoses [46]. We further verified the clinical correlation between DDC and UC.

Two validation data sets (AUC = 0.976 and 0.797) based on the 5-gene subject operating characteristic curve (ROC) provide new ideas for the diagnosis of UC prediction. In addition, we developed a nomogram model for diagnosing UC subtypes using BHLHE40, CDK20, SCIN, VIL1, and DDC. Based on the results of our study, we concluded that this prediction model is capable of making excellent predictions, indicating its value for clinical applications.

In this study, some limitations need to be highlighted. This paper is a full-coverage bioinformatics study for the analysis, so we need more clinical studies and experimental studies to validate the expression level of CRG involved in UC. More clinical datas and samples are required to support the accuracy of our model, the correlation between CRGs and immune infiltration.

In spite of the growing knowledge of the effects of environmental exposures, genetics, and gut microbes on disease, the exact mechanisms by which diseases develop remain unclear. We currently investigate the correlation between CRGs and immune infiltrates and the immunological variability of different cuproptosis-related molecular clusters in UC patients. A RF model for 5 genes is considered the best machine learning model for detecting UC tissue subtypes and the predictive model for UC patients. In our study, we identify for the first time the role of cuproptosis in UC and reveal the unknown underlying molecular mechanisms of UC.

Declarations

References

1. Mulder DJ, Noble AJ, Justinich CJ, Duffin JM. A tale of two diseases: the history of inflammatory bowel disease. J Crohns Colitis. 2014;8(5):341–8.
2. Buse Balci S, Aktas G. A comprehensive review of the role of hemogram derived inflammatory markers in gastrointestinal conditions. Iran J Colorectal Res. 2022;10(3):75–86.
3. Aktaş G, Duman TT, Atak B, et al. Irritable bowel syndrome is associated with novel inflammatory markers derived from hemogram parameters. Fam Med Prim Care Rev. 2020;22(2):107–10.
4. Posul E, Yilmaz B, Aktas G, Kurt M. Does neutrophil-to-lymphocyte ratio predict active ulcerative colitis? Wien Klin Wochenschr. 2015;127(7–8):262–5.
5. Ng SC, Shi HY, Hamidi N, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet. 2017;390(10114):2769–78.
6. Jones G-R, Lyons M, Plevris N, et al. IBD prevalence in Lothian, Scotland, derived by capture-recapture methodology. Gut. 2019;68(11):1953–60.
7. Kaplan GG. The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol. 2015;12(12):720–7.
8. Conley S, Proctor DD, Jeon S, Sandler RS, Redeker NS. Symptom clusters in adults with inflammatory bowel disease. Res Nurs Health. 2017;40(5):424–34.
9. Bombardier CH, Buchwald D. Chronic fatigue, chronic fatigue syndrome, and fibromyalgia. Disability and health-care use Med Care. 1996;34(9):924–30.
10. Nocerino A, Nguyen A, Agrawal M, et al. Fatigue in inflammatory bowel diseases: etiologies and management. Adv Ther. 2020;37(1):97–112.
11. Cohen RD, Yu AP, Wu EQ, et al. Systematic review: the costs of ulcerative colitis in Western countries. Aliment Pharmacol Ther. 2010;31(7):693–707.
12. Abraham BP, Kane S. Fecal markers: calprotectin and lactoferrin. Gastroenterol Clin North Am. 2012;41(2):483–95.
13. Ng SC, Bernstein CN, Vatn MH, et al. Geographical variability and environmental risk factors in inflammatory bowel disease. Gut. 2013;62(4):630–49.
14. Solberg IC, Lygren I, Jahnsen J, et al. Clinical course during the first 10 years of ulcerative colitis: results from a population-based inception cohort (IBSEN Study). Scand J Gastroenterol. 2009;44(4):431–40.
15. Ho G-T, Chiam P, Drummond H, et al. The efficacy of corticosteroid therapy in inflammatory bowel disease: analysis of a 5-year UK inception cohort. Aliment Pharmacol Ther. 2006;24(2):319–30.
16. Farrell RJ. Biologics beyond Anti-TNF Agents for Ulcerative Colitis - Efficacy, Safety, and Cost? N Engl J Med. 2019;381(13):1279–81.
17. Singh S, George J, Boland BS, Vande Casteele N, Sandborn WJ. Primary non-response to tumor necrosis factor antagonists is associated with inferior response to second-line biologics in patients with inflammatory bowel diseases: a systematic review and meta-analysis. J Crohns Colitis. 2018;12(6):635–43.
18. Kirsner JB. Historical aspects of inflammatory bowel disease. J Clin Gastroenterol. 1988;10(3):286–97.

19. Tsvetkov P, Coy S, Petrova B, et al. Copper induces cell death by targeting lipoylated TCA cycle proteins. Science. 2022;375(6586):1254–61.

20. Tang D, Chen X, Kroemer G. Cuproptosis: a copper-triggered modality of mitochondrial cell death. Cell Res. 2022;32(5):417–8.

21. Cobine PA, Brady DC. Cuproptosis: Cellular and molecular mechanisms underlying copper-induced cell death. Mol Cell. 2022;82(10):1786–7.

22. Tsukumo S-I, Yasutomo K. Regulation of CD8+ T cells and antitumor immunity by notch signaling. Front Immunol. 2018;9:101.

23. Garis M, Garrett-Sinha LA. Notch signaling in B cell immune responses. Front Immunol. 2020;11: 609324.

24. Zhao J, Guo S, Schrodi SJ, He D. Cuproptosis and cuproptosis-related genes in rheumatoid arthritis: implication, prospects, and perspectives. Front Immunol. 2022;13: 930278.

25. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846–7.

26. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.

27. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572–3.

28. Wu D, Yin Z, Ji Y, et al. Identification of novel autophagy-related lncRNAs associated with a poor prognosis of colon adenocarcinoma through bioinformatics analysis. Sci Rep. 2021;11(1):8069.

29. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

30. Rigatti SJ. Random forest. J Insur Med. 2017;47(1):31–9.

31. Gold C, Sollich P. Model selection for support vector machine classification. Neurocomputing. 2003;55(1):221–49.

32. Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society: Series A (General). 1972;135(3):370–84.

33. Zhang X, Lu Z, Ren X, et al. Genetic comprehension of organophosphate flame retardants, an emerging threat to prostate cancer. Ecotoxicol Environ Saf. 2021;223: 112589.

34. Liu Y, Wang J, Li L, et al. AC010973.2 promotes cell proliferation and is one of six stemness-related genes that predict overall survival of renal clear cell carcinoma. Sci Rep. 2022;12(1):4272.

35. Bastida G, Mínguez A, Nos P, Moret-Tatay I. Immunoepigenetic regulation of inflammatory bowel disease: current insights into novel epigenetic modulations of the systemic immune response. Genes (Basel). 2023;14(3):554.

36. Li F, Ouyang J, Chen Z, et al. Nanomedicine for T-cell mediated immunotherapy. Adv Mater. 2023;25: e2301770.

37. Shao M, Yuan F, Liu J, Luo H. Mast cell specific receptor mrgprb2 regulating experimental colitis is associated with the microbiota-gut-brain axis. J Inflamm Res. 2022;15:6137–51.

38. Tang S, Zhong W, Li T, Li Y, Song G. Isochlorogenic acid A alleviates dextran sulfate sodium-induced ulcerative colitis in mice through STAT3/NF-κB pathway. Int Immunopharmacol. 2023;118: 109989.

39. Falahati F, Westman E, Simmons A. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimers Dis. 2014;41(3):685–708.

40. Zhang L, Li Z, Skrzypczynska KM, et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. Cell. 2020;181(2):442-459.e29.

41. Lai L, Shin GY, Qiu H. The role of cell cycle regulators in cell survival-dual functions of cyclin-dependent kinase 20 and p21Cip1/Waf1. Int J Mol Sci. 2020;21(22):8504.

42. Satoda Y, Noguchi T, Fujii T, et al. BROMI/TBC1D32 together with CCRK/CDK20 and FAM149B1/JBTS36 contributes to intraflagellar transport turnaround involving ICK/CILK1. Mol Biol Cell. 2022;33(9):ar79.

43. Chen K, Zeng Z, Ma C, Dang Y, Zhang H. Commentary on: screening of immunosuppressive cells from colorectal adenocarcinoma and identification of prognostic markers. Biosci Rep. 2021;41(12):BSR20211096.

44. Aschtgen M-S, Bernard CS, De Bentzmann S, Lloubès R, Cascales E. SciN is an outer membrane lipoprotein required for type VI secretion in entero-aggregative Escherichia coli. J Bacteriol. 2008;190(22):7523–31.

45. Salewski I, Henne J, Engster L, et al. CDK4/6 blockade provides an alternative approach for treatment of mismatch-repair deficient tumors. Oncoimmunology. 2022;11(1):2094583.

46. Artemaki PI, Papatsirou M, Boti MA, et al. Revised exon structure of l-DOPA Decarboxylase (DDC) reveals novel splice variants associated with colorectal cancer progression. Int J Mol Sci. 2020;21(22):8568.

Publisher's Note