MEDS Clinical Medicine (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/medsc.2023.040617
ISSN 2616-1907 Vol. 4 Num. 6

Analysis of genes and immune cell infiltration related to the diagnosis of ulcerative colitis based on machine learning

Yan He[1], Bao Xin[1,*]

[1]*School of Public Health, Shaanxi University of Traditional Chinese Medicine, Xianyang, Shaanxi, China*

[*]*Corresponding author: 89561185@qq.com*

Keywords: Ulcerative colitis, Machine learning, Diagnosis, Immune infiltration

Abstract: At present, the diagnosis of ulcerative colitis mainly relies on endoscopic methods, and the diagnostic results are often difficult to distinguish from Crohn's disease. This study aims to mine gene expression data at the molecular level to determine the factors related to the diagnosis of ulcerative colitis. Characteristic genes and immune infiltration analysis provide new directions for the diagnosis and treatment of ulcerative colitis. We downloaded the ulcerative colitis gene expression data sets GSE38713 and GSE87466 from the GEO database as training sets for differentially expressed gene analysis. We used three machine learning methods: random forest, XGB, and LASSO regression to analyze the differentially expressed genes. The integrated analysis results determined that Characteristic genes related to the diagnosis of ulcerative colitis and validated in the GSE47908 data set. Immune infiltration analysis was performed on normal samples and ulcerative colitis samples using the CIBERSOR algorithm, and the correlation between signature genes and immune cell infiltration levels was evaluated. It was finally determined that the differential genes related to the diagnosis of ulcerative colitis are: PDZK1IP1, SERPINA1, and TRIM29, which showed good diagnostic ability (AUC>0.8) in both the training set and the validation set. Moreover, PDZK1IP1, SERPINA1, and TRIM29 are positively correlated with dendritic cell resting, monocytes, macrophages, dendritic cell activation, and regulatory T cells, and are positively correlated with T cell CD4+ memory cell activation, natural killer cells, and M1 giant cells. Phagocytes were negatively correlated. In summary, PDZK1IP1, SERPINA1, and TRIM29 may be involved in the occurrence and development of ulcerative colitis through a variety of immune cells, and can be used as diagnostic biomarkers for ulcerative colitis.

1. Introduction

Ulcerative colitis (UC) is a common inflammatory bowel disease that starts in the rectum and extends to the entire colon in a continuous manner, with the characteristics of recurring relapses and remissions [1]. In recent years, with the development of economic level, the national dietary structure has changed significantly, and the incidence and prevalence of UC in my country have

increased significantly [2]. The cause of UC is still unclear, but it is mainly believed to be related to genetic, immune, environmental, microbial and other factors [3]. Currently, clinical diagnosis of UC mainly relies on endoscopic and histological examination, but it is difficult to accurately distinguish UC from Crohn's disease, and there is a possibility of misdiagnosis, thus delaying patient treatment [4]. Therefore, exploring effective UC diagnostic biomarkers is very important for precision medicine of UC.

With the development of high-throughput sequencing technology, massive omics data continue to pour in, providing more possibilities for people to understand the development of diseases at the molecular level [5]. However, the characteristics of omics data such as high dimensionality, high noise, and complex integration bring challenges to traditional data mining methods. Machine learning algorithms have the advantages of automatically processing large amounts of data, strong model generalization capabilities, and avoiding overfitting, and are suitable for processing Omics data have been widely used in various medical research in recent years [6]. For example, Tao Xiong et al. used a variety of machine learning algorithms to determine GoS2 and HPSE as biomarkers for aortic aneurysms [7]. This study intends to use the gene expression data of UC in the Gene Expression Omnibus (GEO) database, integrate multiple machine learning algorithms to explore characteristic genes related to UC diagnosis, and conduct immune infiltration analysis to provide direction for further research on UC.

2. Materials and Methods

2.1. Data download and preprocessing

The gene expression sequencing data and associated platform files for datasets GSE38713 and GSE87466 were retrieved from the GEO database. GSE38713 contains 13 normal samples and 15 ulcerative colitis samples. GSE87466 contains 21 normal samples and 74 ulcerative colitis samples. The two data sets were integrated, and the "sva" package of R software was used to perform batch correction and normalization processing on the two data sets to obtain a data set with 34 normal samples and 89 UC samples as a training set. Dataset GSE47908 was additionally procured to serve as an external validation set.

2.2. Differentially expressed gene screening and enrichment analysis

Differential gene expression analysis between normal and UC tissues was conducted using the "limma" package in R, adopting a threshold of $|\log2FC|>1$ and an adjusted P-value <0.05. Subsequent Gene Ontology (GO) enrichment analysis of the identified DEGs was performed using the "clusterProfiler" package in R, elucidating their roles in biological processes (BP), cellular components (CC), and molecular functions (MF). Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analysis was executed to delineate the functional pathways and regulatory mechanisms associated with these DEGs.

2.3. Machine learning identifies key genes

Xtreme Gradient Boosting (XGB) is an algorithm based on GBDT, which has the characteristics of high efficiency and flexibility and is widely used in biological information mining [8]. The random forest algorithm is an algorithm based on ensemble learning that merges multiple decision trees together. It can well avoid overfitting of the model and is not susceptible to the influence of collinearity among variables [9]. The least absolute shrinkage and selection operator (LASSO) algorithm continuously compresses the coefficients by introducing penalty terms to achieve the

purpose of streamlining the model, and at the same time effectively handles the problems of overfitting and multicollinearity. Each curve in the regression coefficient trajectory graph represents a variable. The complexity of the model is adjusted through λ. Under the λ value with the smallest deviation in the cross-validation curve, the LASSO regression fitting effect is the best [10].

In this study, whether suffering from UC was used as the outcome variable, and DGE was used as the dependent variable. Three machine learning algorithms, XGB, random forest, and LASSO regression, were used to predict the differentially expressed genes related to the occurrence of UC. The results of the three machine learning algorithms were further integrated. Final identification of signature genes associated with ulcerative colitis diagnosis.

2.4. Diagnostic value of characteristic genes in UC

Receiver operating characteristic (ROC) curves for key genes were plotted for both the training and validation sets. Optimal cutoff values were determined based on the AUC, with the guiding principle that a greater AUC indicates superior predictive accuracy of the model. The higher the AUC, the better the predictive performance of the model.

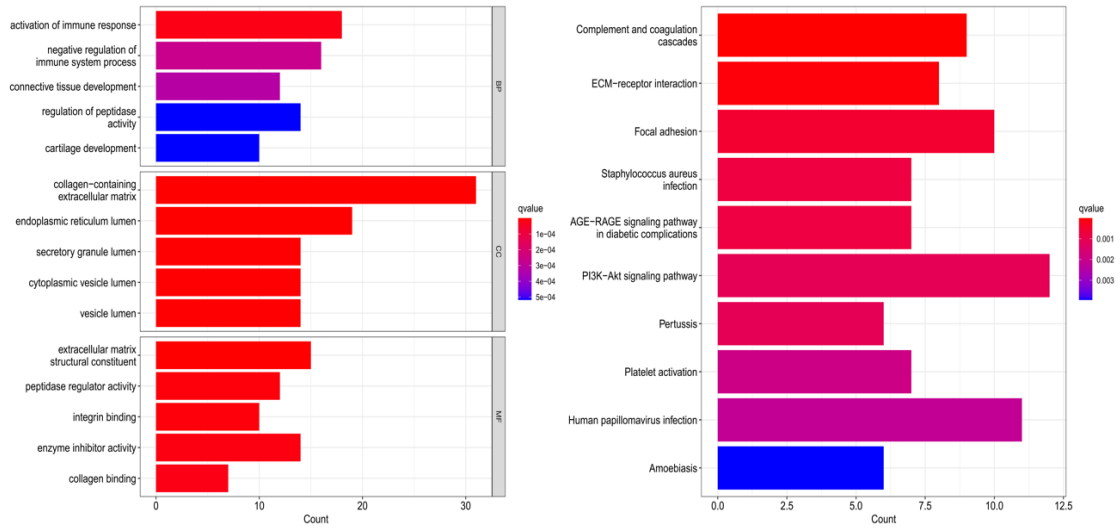
2.5. Immune infiltration analysis

CIBERSORT is a deconvolution algorithm based on linear support vector regression that is widely used to quantify immune cell infiltration levels in gene expression data. This study uses the CIBERSORT algorithm to evaluate immune cell infiltration in UC and normal tissue samples, and conducts correlation analysis between characteristic genes and immune cell infiltration levels.

3. Result

3.1. Differentially expressed gene screening

After analyzing the gene expression data of GSE38713 and GSE87466, a total of 143 DEGs were obtained, including 98 up-regulated genes and 45 down-regulated genes. The GO enrichment analysis results of DEG showed that BP was mainly enriched in activating immune response, negative immune regulation, connective tissue development, regulation of peptidase activity, and cartilage development. CC was mainly enriched in inhibiting collagen extracellular matrix, The endoplasmic reticulum lumen, secretory granule lumen, cytoplasmic vesicle lumen, vesicle lumen, MF are mainly enriched in extracellular matrix structural components, peptide magnesium regulatory activity, integrins, enzyme inhibitor activity and other biological functions (see Figure 1A). KEGG functional enrichment analysis results showed that DEGs were mainly enriched in the complement and coagulation cascade signaling pathways, extracellular matrix receptor interactions, focal adhesion, Staphylococcus aureus infection, AGE-RAGE signaling pathway, P13K-Akt signaling pathway, Whooping cough, platelet activation and other signaling pathways (see Figure 1B).

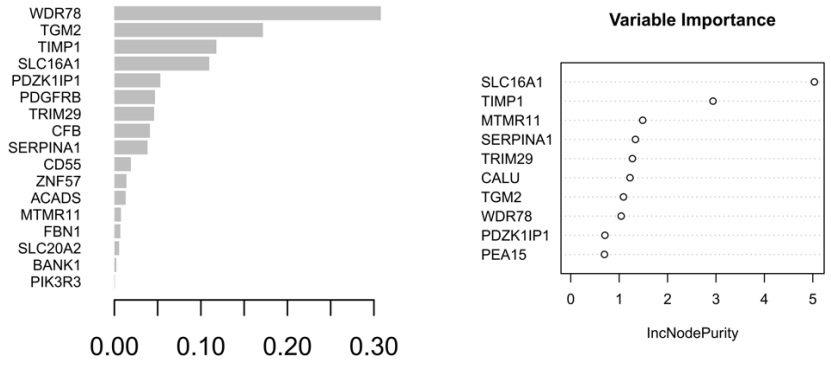


A B

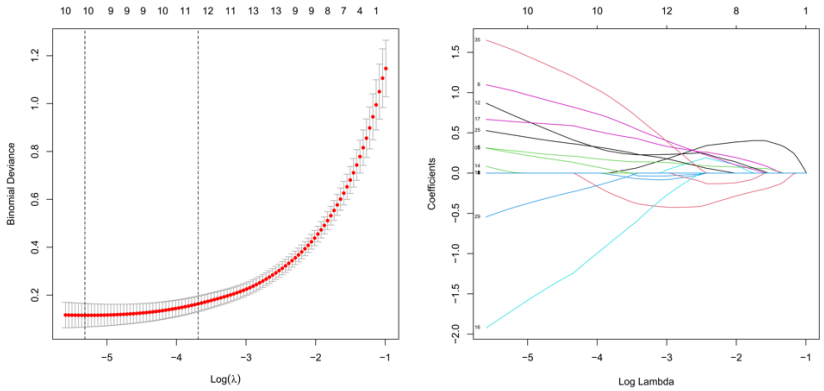
Figure 1(A): Results of GO enrichment analysis of differential genes. (B): Differential gene KEGG enrichment analysis results.

3.2. Machine learning to screen diagnostic signature genes

Use the R language "XGBoost" software package to screen 17 characteristic genes related to the diagnosis of ulcerative colitis (see Figure 2A); use the R language "Random Forest" software package to set the number of seeds to 2023, mytry=3, ntree=1000. Random forest feature selection retains the top ten most important genes (see Figure 2B). LASSO regression analysis was performed using the "glmnet" software package in R language, and 10-fold cross validation was used to finally obtain 10 differentially expressed genes related to diagnosis (see Figure 2C). Finally, integrating the characteristic gene selection results of three machine learning algorithms, it was found that there are three overlapping characteristic genes: PDZK1IP1, SERPINA1, and TRIM29. These three genes may be key genes in the occurrence of UC.



A B



C

Figure 2: (A. B. C) Machine learning algorithm screens diagnostic marker results.

3.3. Significance of key genes in ulcerative colitis

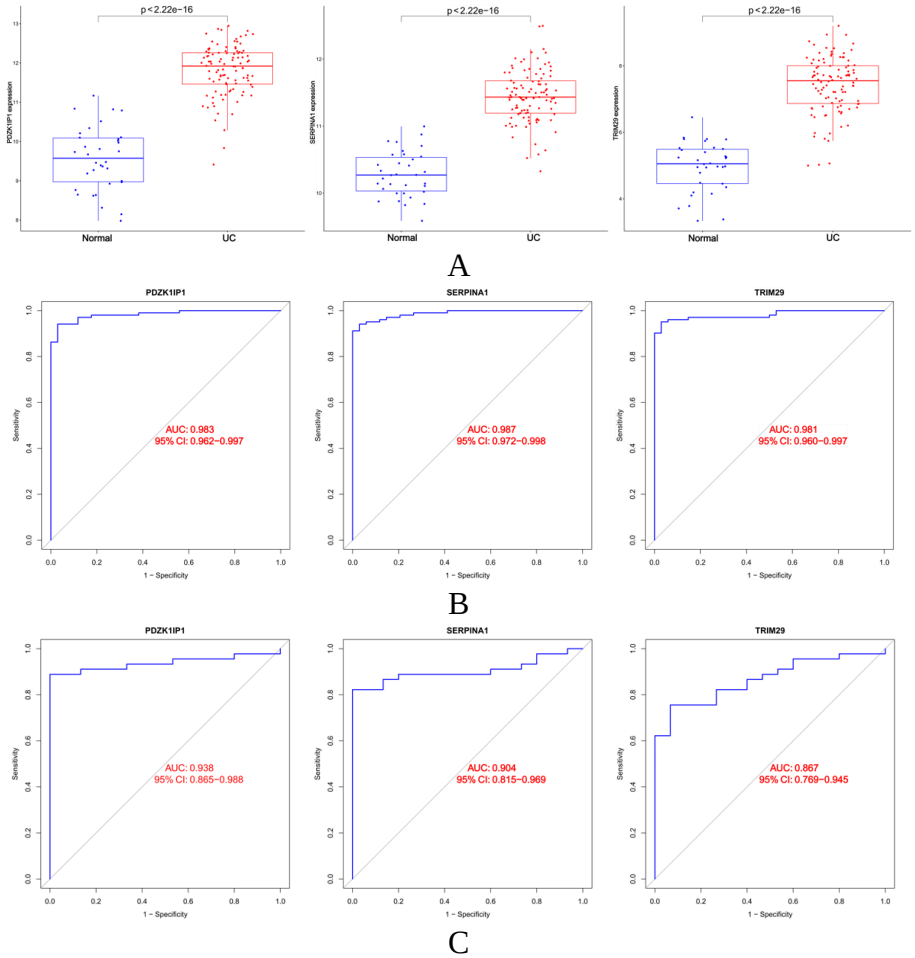


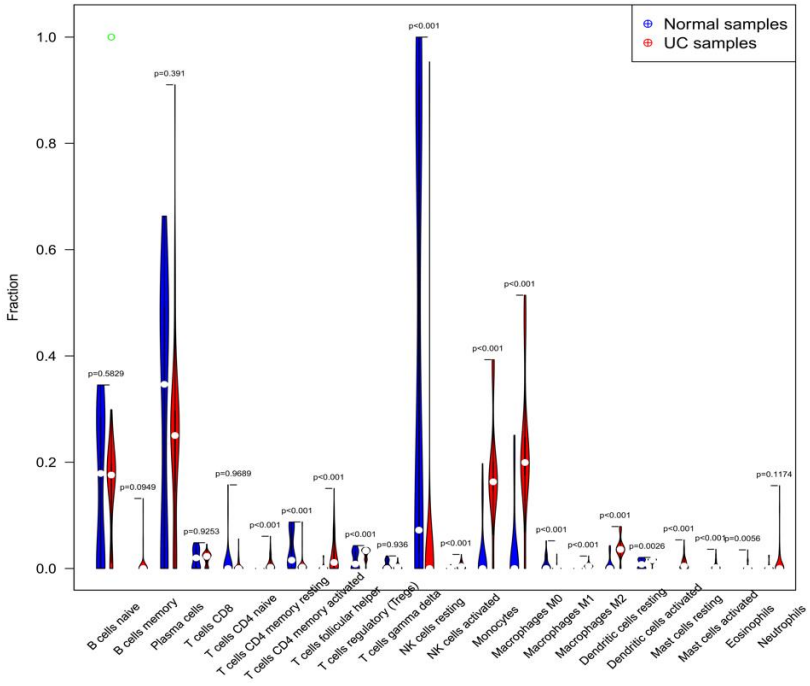
Figure 3 (A): Differential expression box plot of PDZK1IP1, SERPINA1, and TRIM29. (B): ROC curves of PDZK1IP1, SERPINA1, and TRIM29 in the training set. (C): ROC curves of PDZK1IP1, SERPINA1, and TRIM29 in the validation set.

Compared with healthy samples, the expression levels of PDZK1IP1, SERPINA1, and TRIM29 were significantly upregulated in UC patients (see Figure 3A). To further explore the diagnostic

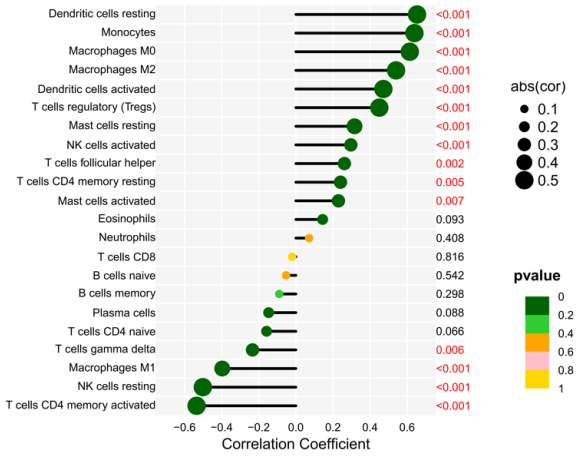
value of PDZK1IP1, SERPINA1, and TRIM29, the ROC curve was drawn in the training set and the AUC was calculated. The results showed that EMP2, CLIC5, and TNNC1 have strong ability to identify UC (AUC>0.9) (see Figure 3B), and show the same identification ability (AUC>0.8) in the validation set (see Figure 3C).

3.4. Immune infiltration analysis

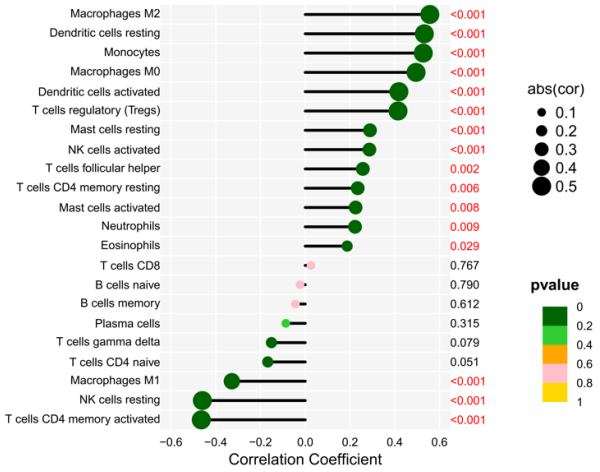
The relative abundance of various immune cells differed between normal tissue and ulcerative colitis samples, p<0.01 (see Figure 4A). Analysis of the correlation between characteristic genes and immune infiltration cell levels showed that PDZK1IP1, SERPINA1, and TRIM29 were all associated with resting dendritic cells, monocytes, macrophages, M2 macrophages, activated dendritic cells, and regulatory T cells. etc. are positively correlated with and negatively correlated with T cell CD4+ memory cell activation, natural killer cells, and M1 macrophages (see Figure 4B, Figure 4C, and Figure 4D).



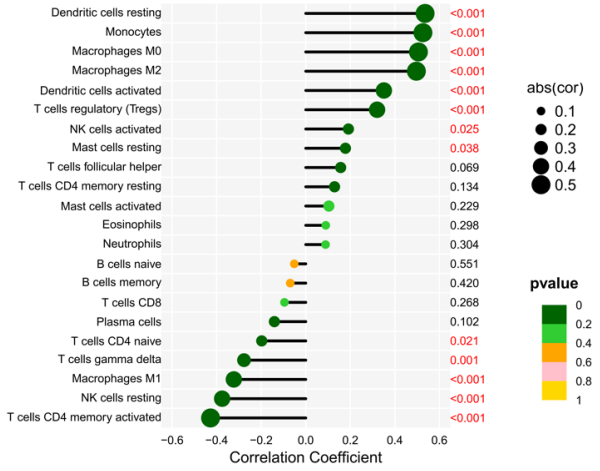
A



B



C



D

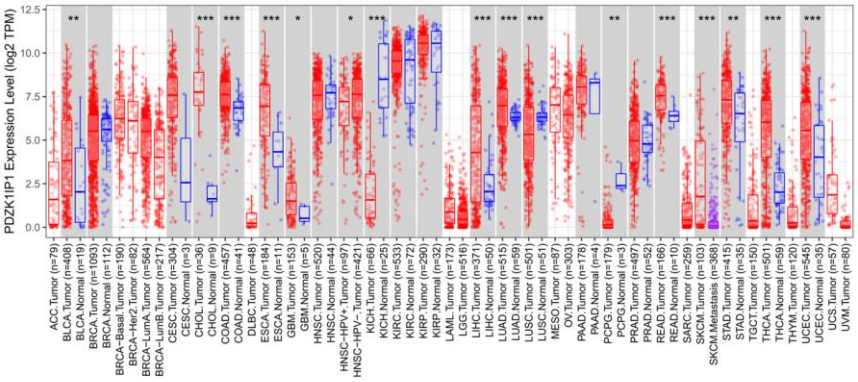
Figure 4(A): Results of immune infiltration analysis of 22 immune cell subtypes in normal samples and UC samples. (B): Correlation between PDZK1IP1 and immune infiltrating cells. (C): Correlation between SERPINA1 and immune infiltrating cells. (D): TRIM29 correlates with immune infiltrating cells.

4. Discuss

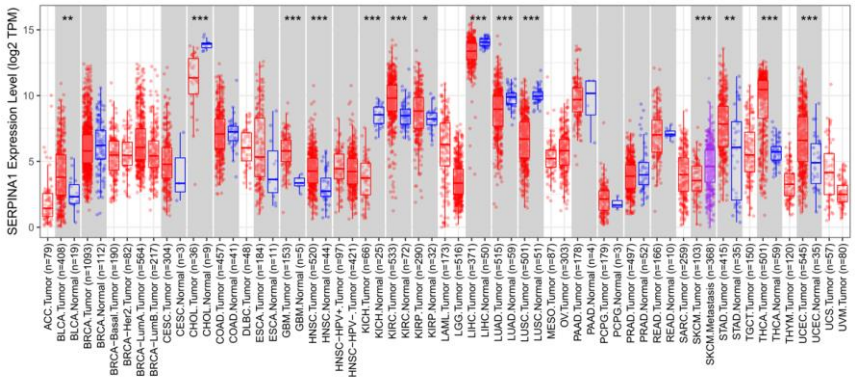
UC is a chronic non-specific disease. The main clinical manifestations are abdominal pain, diarrhea, mucus, pus and blood in the stool. It is difficult to cure and easy to relapse. It has been identified by the World Health Organization as one of the modern refractory diseases [11]. There are currently no diagnostic criteria, and exclusive diagnosis mainly relies on a combination of clinical manifestations, endoscopic and pathological examinations [12]. Therefore, exploring diagnostic biomarkers for UC is very beneficial to improving the clinical treatment of patients.

Studies have shown that UC patients have a higher risk of colon cancer compared with normal tissue [13]. To this end, we further explored the correlation between the three UC diagnostic biomarkers discovered in this study and the occurrence of colon cancer. Pan-cancer analysis of PDZK1IP1, SERPINA1, and TRIM29 was performed through the TIMER database (http://timer.cistrome.org/) and it was found that PDZK1IP1 and TRIM29 were significantly highly expressed in colon cancer (see Figure 5). By plotting the relationship between PDZK1IP1 and

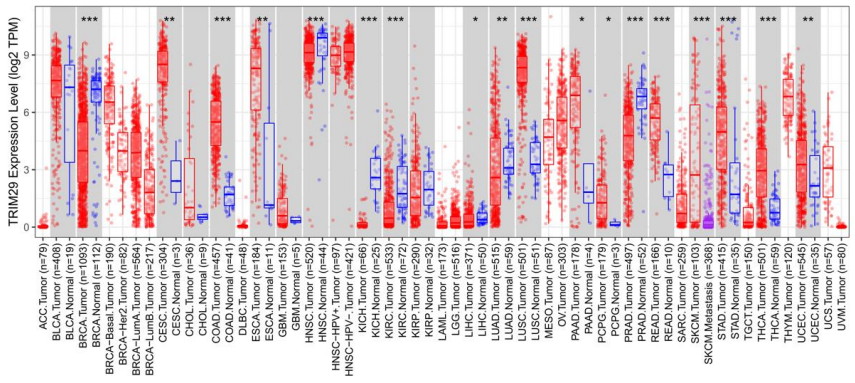
TRIM29 The overall survival curve quantifies the clinical value of PDZK1IP1 and TRIM29 in colon cancer. The results show that the expression of PDZK1IP1 has no significant impact on the survival of colon cancer patients p>0.05, while the expression of TRIM29 is significantly related to the overall survival rate of colon cancer (p<0.05), the overall survival rate of low-expression TRIM29 is higher than that of patients with high-expression colon cancer, which helps to improve the prognosis of patients and prolong the survival time of patients (see Figure 6). This is consistent with the study by Jiang T et al. who found that overexpression of TRIM29 can activate P13K/ This is consistent with the research conclusion that the AKT signaling pathway promotes the occurrence of colorectal cancer [14].



A



B



C

Figure 5(A): PDZK1IP1 pan-cancer analysis results. (B): SERPINA1 pan-cancer analysis results. (C): TRIM29 pan-cancer analysis results.

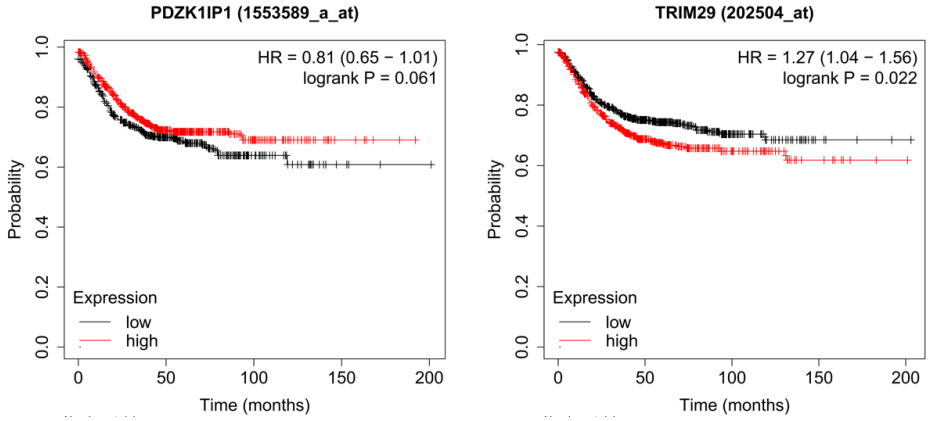


Figure 6: PDZK1IP1, TRIM29 overall survival curve.

This study combined bioinformatics methods and machine learning algorithms to analyze UC gene expression data, and finally identified three characteristic genes, PDZK1IP1, SERPINA1, and TRIM29, that are related to the diagnosis of UC. Through pan-cancer analysis, it was found that low-expression TRIM29 can prolong life in patients with colon cancer. Survival time. PDZK1IP1, also known as MAP17, plays an important role in both inflammation and cancer [15]. For example, Wei Zhang et al. found that PDZK1IP1 can promote the proliferation, migration and invasion of papillary thyroid cancer [16]. The research results of J TBjerrum et al. showed that, PDZK1IP1 has the ability to distinguish UC from Crohn's colitis [17], but the pathway by which PDZK1IP1 induces UC needs further experiments to determine. SERPINA1 is the main member of the serine protease inhibitor family and is related to the occurrence of various diseases such as anemia and chronic obstructive pulmonary disease [18]. Peishan Qiu et al. have found that SERPINA1 is a central gene related to active UC autophagy and is related to active UC autophagy. There is a positive correlation with the occurrence of sexual UC [19], which is consistent with the conclusion of this study. TRIM29 is a member of the TRIM protein family. Studies have shown that overexpressed TRIM29 can regulate the NF-kB pathway through PKC activation and play an important role in carcinogenesis, autophagy and immunity of the disease [20]. However, regarding the role of TRIM29 in UC Research progress has rarely been reported. This study found that the expression of PDZK1IP1, SERPINA1, and TRIM29 was significantly upregulated in UC, which provided a theoretical basis for the accessibility and feasibility of clinical application of PDZK1IP1, SERPINA1, and TRIM29 as diagnostic markers for UC.

UC is considered a complex, immune-mediated disease [21]. This study used CIBERSOTR to evaluate the types of immune cell infiltration in UC patients and normal samples. The results showed that T cell CD4+ memory cell resting, T cell CD4+ memory cell activation, filter Alveolar helper T cells, regulatory T cells, resting natural killer cells, activated natural killer cells, monocytes, macrophages, and dendritic cell activation may be related to the occurrence of UC. At the same time, correlation analysis between characteristic genes and immune infiltration cell levels showed that PDZK1IP1, SERPINA1, and TRIM29 interact with different immune cells and may be key factors in regulating the molecular and immune infiltration status of UC patients. The above immune infiltration analysis studies show that infiltrating immune cells play an important role in UC, and related immunotherapy may become the focus of future research.

5. Conclusions

In summary, this study found that PDZK1IP1, SERPINA1, and TRIM29 are closely related to the occurrence of UC and may be diagnostic biomarkers for UC, and further found that high

expression of TRIM29 may promote the occurrence of colon cancer. However, this study still has some limitations. First, the sample size used in the public database is small, and the research results lack further experimental verification. In summary, the three key genes unearthed by this study combined with bioinformatics technology and machine learning algorithms play an important role in the occurrence and immune infiltration of UC, and have the potential to become targets for early diagnosis and treatment of UC.

References

[1] Ordas I, Eckmann L, Talamini M, et al. Ulcerative colitis [J]. Lancet, 2012, 380(9853): 1606-1619.

[2] Lin, S., et al. Fusobacterium nucleatum aggravates ulcerative colitis through promoting gut microbiota dysbiosis and dysmetabolism[J]. Journal of Periodontology, 2023, 94(3):405-418.

[3] Eisenstein M. Ulcerative colitis: towards remission [J]. Nature, 2018, 563(7730): S33.

[4] Kaenkumchorn T, Wahbeh G. Ulcerative Colitis: Making the Diagnosis [J]. Gastroenterol Clin North Am, 2020, 49(4): 655-669.

[5] Kwoji, I. D., et al. 'Multi-omics' data integration: applications in probiotics studies[J]. npj Science of Food, 2023, 7(1):25.

[6] Handelman G S, Kuan K H, Chandra R V, et al. eDoctor, Machine Learning and the Future of Medicine [J]. Journal of Internal Medicine, 2018, 284.

[7] Xiong T, Lv X S, Wu G J, et al. Single-Cell Sequencing Analysis and Multiple Machine Learning Methods Identified G0S2 and HPSE as Novel Biomarkers for Abdominal Aortic Aneurysm [J]. Front Immunol, 2022, 13: 907309.

[8] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System; proceedings of the Knowledge Discovery and Data Mining, F, 2016 [C].

[9] Ghosh D, Cabrera J. Enriched Random Forest for High Dimensional Genomic Data [J]. IEEE/ACM Trans Comput Biol Bioinform, 2022, 19(5): 2817-2828.

[10] Dai P, Chang W, Xin Z, et al. Retrospective Study on the Influencing Factors and Prediction of Hospitalization Expenses for Chronic Renal Failure in China Based on Random Forest and LASSO Regression [J]. Frontiers in Public Health, 2021, 9: 748-759.

[11] Sun Y, Zhang Z, Zheng C Q, et al. Mucosal lesions of the upper gastrointestinal tract in patients with ulcerative colitis: A review [J]. World J Gastroenterol, 2021, 27(22): 2963-2978.

[12] Sato, Y., et al. Inflammatory Bowel Disease and Colorectal Cancer: Epidemiology, Etiology, Surveillance, and Management[J]. Cancers, 2023, 15(16):4154.

[13] Yao D, Dong M, Dai C, et al. Inflammation and Inflammatory Cytokine Contribute to the Initiation and Development of Ulcerative Colitis and Its Associated Cancer [J]. Inflamm Bowel Dis, 2019, 25(10): 1595-1602.

[14] Jiang T, Wang H, Liu L, et al. CircIL4R activates the PI3K/AKT signaling pathway via the miR-761/TRIM29/PHLPP1 axis and promotes proliferation and metastasis in colorectal cancer [J]. Mol Cancer, 2021, 20(1): 167.

[15] Garcia-Heredia J M, Carnero A. Dr. Jekyll and Mr. Hyde: MAP17's up-regulation, a crosspoint in cancer and inflammatory diseases [J]. Mol Cancer, 2018, 17(1): 80.

[16] Zhang W, Zheng D, Jin L, et al. PDZK1IP1 gene promotes proliferation, migration, and invasion in papillary thyroid carcinoma [J]. Pathol Res Pract, 2022, 238: 154091.

[17] Bjerrum J T, Nyberg C, Olsen J, et al. Assessment of the validity of a multigene analysis in the diagnostics of inflammatory bowel disease [J]. J Intern Med, 2014, 275(5): 484-493.

[18] Sangeetha T, Nargis Begum T, Balamuralikrishnan B, et al. Influence of SERPINA1 Gene Polymorphisms on Anemia and Chronic Obstructive Pulmonary Disease [J]. J Renin Angiotensin Aldosterone Syst, 2022, 2022: 2238320.

[19] Qiu P, Liu L, Fang J, et al. Identification of Pharmacological Autophagy Regulators of Active Ulcerative Colitis [J]. Front Pharmacol, 2021, 12: 769718.

[20] Liang C, Dong H, Miao C, et al. TRIM29 as a prognostic predictor for multiple human malignant neoplasms: a systematic review and meta-analysis [J]. Oncotarget, 2018, 9(15): 12323-12332.

[21] Porter R J, Kalla R, Ho G T. Ulcerative colitis: Recent advances in the understanding of disease pathogenesis [J]. F1000Res, 2020, 9.