Research Paper

Development of a susceptibility gene based novel predictive model for the diagnosis of ulcerative colitis using random forest and artificial neural network

Hanyang Li[1,2,3], Lijie Lai[1,2,3], Jun Shen[1,2,3]

[1]Division of Gastroenterology and Hepatology, Key Laboratory of Gastroenterology and Hepatology, Ministry of Health, Inflammatory Bowel Disease Research Center, Shanghai 200127, China
[2]Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China
[3]Shanghai Institute of Digestive Disease, Shanghai 200127, China

Correspondence to: Jun Shen; **email:** shenjun@renji.com

Keywords: ulcerative colitis, predictive model, susceptibility genes, random forest, artificial neural network
Received: May 7, 2020 **Accepted:** July 21, 2020 **Published:** October 24, 2020

ABSTRACT

Ulcerative colitis is a type of inflammatory bowel disease characterized by chronic and recurrent nonspecific inflammation of the intestinal tract. To find susceptibility genes and develop a novel predictive model of ulcerative colitis, two sets of cases and a control group containing the ulcerative colitis gene expression profile (training set GSE109142 and validation set GSE92415) were downloaded and used to identify differentially expressed genes. A total of 781 upregulated and 127 downregulated differentially expressed genes were identified in GSE109142. The random forest algorithm was introduced to determine 1 downregulated and 29 upregulated differentially expressed genes contributing highest to ulcerative colitis occurrence. Expression data of these 30 genes were transformed into gene expression scores, and an artificial neural network model was developed to calculate differentially expressed genes weights to ulcerative colitis. We established a universal molecular prognostic score (mPS) based on the expression data of the 30 genes and verified the mPS system with GSE92415. Prediction results agreed with that of an independent data set (ROC-AUC=0.9506/PR-AUC=0.9747). Our research creates a reliable predictive model for the diagnosis of ulcerative colitis, and

INTRODUCTION

Ulcerative colitis (UC) is a subtype of inflammatory bowel disease (IBD) characterized by chronic and recurrent nonspecific inflammation of the intestinal tract, manifesting as diffuse inflammation confined to the mucosa and submucosa of the large intestine [1]. The lesions of UC are commonly located in the sigmoid colon and rectum, and may extend to the entire colon. Over the past decade, UC prevalence is increasing worldwide, especially in developing countries. Ulcerative colitis can occur at any age, and majority of cases occur in late adolescence or early adulthood. Although no research has reported that the overall mortality rate in patients with UC is substantially different from that of the general population, UC leads to diminished quality of life with symptoms of abdominal pain, diarrhea, bloody mucopurulent stool, and autoimmune-related complications [2]. Therefore, the establishment of an accurate clinical prediction model (CPM) is of great importance to guide diagnosis and treatment.

Scientific and technological progress in the research field has contributed to several systematic sets of molecular prognostic indicator related CPMs of UC. The mRNA of neutrophil gelatinase-associated lipocalin (NGAL) showed overexpression in an inflamed intestine. Therefore, Budzynska et al. [3] used NGAL to predict the clinical and endoscopic activity of UC (area under curve [AUC] = 0.758). Hart et al. [4] combined fecal calprotectin (FC) with Mayo Endoscopic Score (MES) to predict endoscopic and histological activity of UC patients in clinical remission (AUC = 0.743). However, these predictive models are not efficient enough in the screening and early diagnosis of UC, given that clinical outcome assessments are based on a series of laboratory tests and operations. An effective and universal predictive model for early identification of and intervention in UC is lacking.

Although UC still remains unclear, abnormal immunoregulation, and genetic and environmental factors are thought to be relevant for UC pathogenesis [5]. Accumulating current evidence suggests that the pathogenesis of early UC involves the interaction of the susceptibility gene and environmental factors [6–8]. Similarity comparison between symptoms and gene functions has been used to predict pathogenic genes in IBD [9]. With the development of genome-wide association studies for UC, several validated susceptibility genes have been uncovered, including hMLH1, vitamin D receptor gene, etc. [10, 11]. Therefore, investigation of UC susceptibility genes is becoming the focus of UC screening and diagnosis.

Unsatisfying accuracy, low efficiency and loss of early screening prompt us to develop a novel UC predictive model. With the development of technology, machine learning has become a new approach in medical data processing. The random forest (RF) algorithm has been maturely applied in the field of Alzheimer's disease and acute myeloid leukemia [12, 13]. Artificial neural network (ANN) also demonstrates powerful abilities in medical data processing [14]. Thus, our study aims to uncover UC susceptibility genes using RF and establish an ANN predictive model of UC. The predictive model can be used to provide early screening markers for diagnosis and treatment of UC.

RESULTS

The steps to our cohort study are listed in Figure 1, and information for the training dataset (GSE109142) and validation set (GSE92415) related to UC is listed in Table 1.

Screening of DEGs

The volcano map in Figure 2 shows the expression status of all DEGs in GSE109142. A clear demarcation can be identified between upregulated genes and downregulated genes. The heat map in Figure 3 shows the expression status of 908 DEGs, from which we observe the cases with higher levels of upregulated gene expression in comparison to the controls.

DEG enrichment analysis

The 908 DEGs underwent Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. Results of GO analysis (p-value cutoff = 0.01) suggested that DEGs were significantly enriched in bacterial response-related biological processes, such as "response to bacterium" and "defense response to bacterium". KEGG pathway analysis (p-value cutoff = 0.05) suggested that the DEGs were primarily involved in "cytokine–cytokine receptor interaction" and the "Jak-STAT signaling pathway" (Figures 4 and 5).

Diagnosis-related DEGs with RF

We applied the expression data of 908 DEGs to a machine learning algorithm called RF classifier. The RF

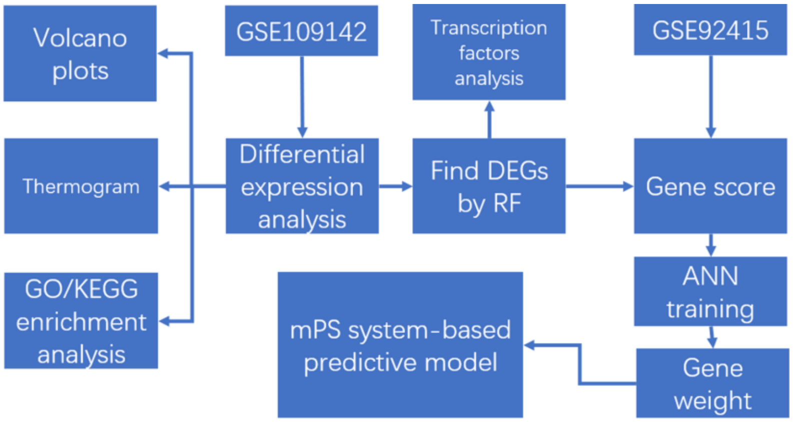
Volcano plots
GSE109142
Transcription factors analysis
GSE92415
Thermogram
Differential expression analysis
Find DEGs by RF
Gene score
GO/KEGG enrichment analysis
mPS system-based predictive model
ANN training
Gene weight

Figure 1. The flow chart in the study. DEGs analysis, gene set enrichment analysis, machine learning, and construction and validation of the mPS system-based predictive model.

Table 1. The information of training/validation sets (GSE109142/GSE92415).

Dataset ID	Platform	Ulcerative Colitis	Normal	Total
GSE109142	GPL16791	206	20	226
GSE92415	GPL13158	53	21	74

screening results are shown in Figure 6. From these sheets, we are able to identify the top-5 UC related genes in the 30 DEGs, including *FAM65C*, *CSF3R*, *CSF3*, *POM121L9P*, and *FER1L4*. The heat map showing the expression status of the top-30 DEGs can be seen in Figure 7.

Thereafter, the 30 DEGs were put into the TRRUST web server. Seven transcriptional regulators (SPI1, ETS2, CEBPB, ETS1, RELA, NFKB1, SP1) were identified to be related with five DEGs (*MMP1*, *CXCR1*, *COL7A1*, *CSF3R*, *PTGIR*). The gene regulatory network diagram is shown as Figure 8 and Table 2.

ANN-based establishment of molecular prognostic score

With expression data transformation of 30 DEGs into "Gene_Score", the weight of each gene was optimized with a neural network algorithm (Table 3). The mPS was calculated by summation of "Gene_Score" × "Gene_Weight" for all 30 DEGs [15]. Thereafter, we regarded the mPS of 226 samples as predicted values, and the clinical outcomes of UC as true values. Following calculation in the ROCR package in R version 3.5.3, the ROC-AUC of our predictive model was 0.9847, and the PR-AUC of our predictive model was 0.9444.

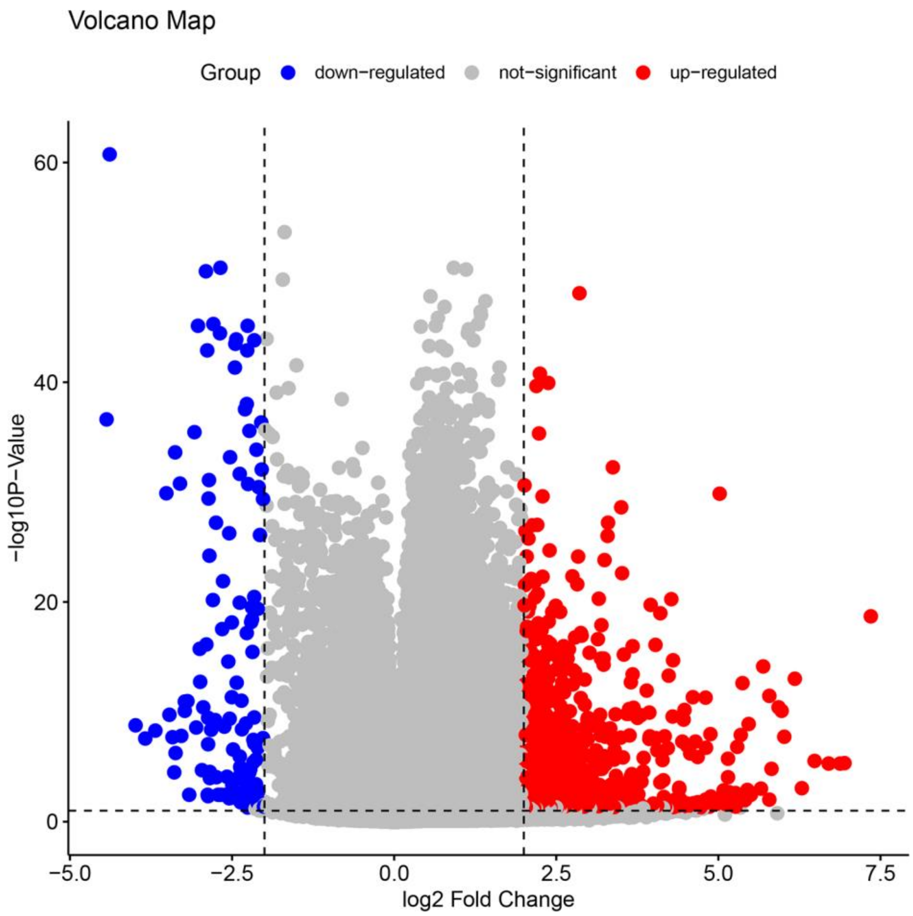

Figure 2. The volcano plots of all the DEGs in GSE109142. In the map, each blue spot represents a down-regulated gene, whereas each red spot represents an up-regulated gene. A clear demarcation can be identified between up-regulated genes and down-regulated genes.

Validation of UC predictive model

To demonstrate whether the mPS system can predict the occurrence of UC, not only in the training set but also in other independent UC cohorts, we introduced an independent UC dataset (GSE92415) to validate the model. The heat map of 20 overlapped genes between GSE109142 and GSE92415 is shown in Figure 9. Similarly, the expression status of the 20 genes was transformed into a “Gene_Score” sheet containing 74 lines of samples, 20 columns of DEGs and 1 column of UC outcome variable (case/control). The “Gene_

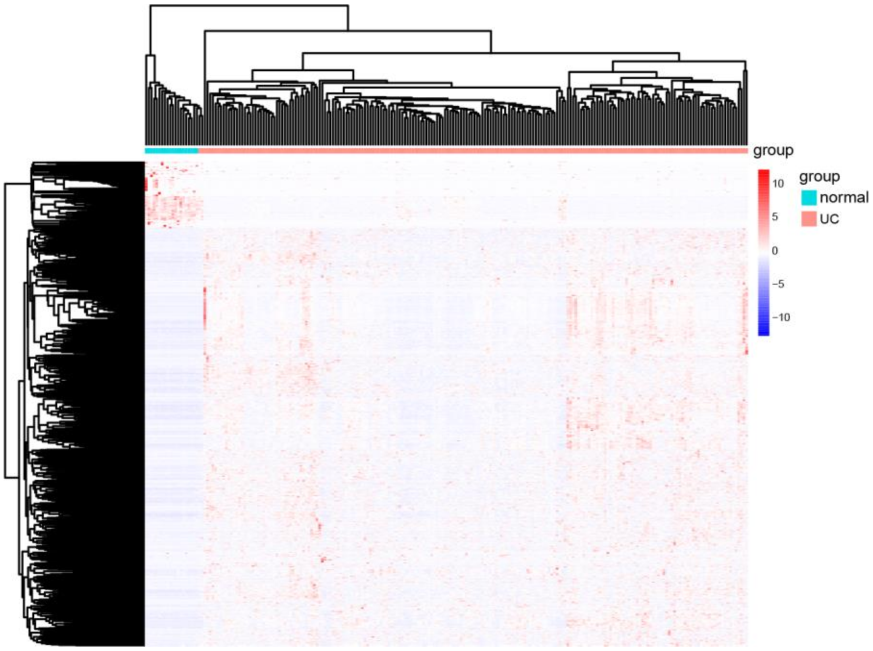

Figure 3. The heat map of all the DEGs in GSE109142. In the map, each list represents a gene and up-regulated genes and down-regulated genes also have a clear demarcation. Green/pink columns represents controls/cases (normal people/ UC patients).

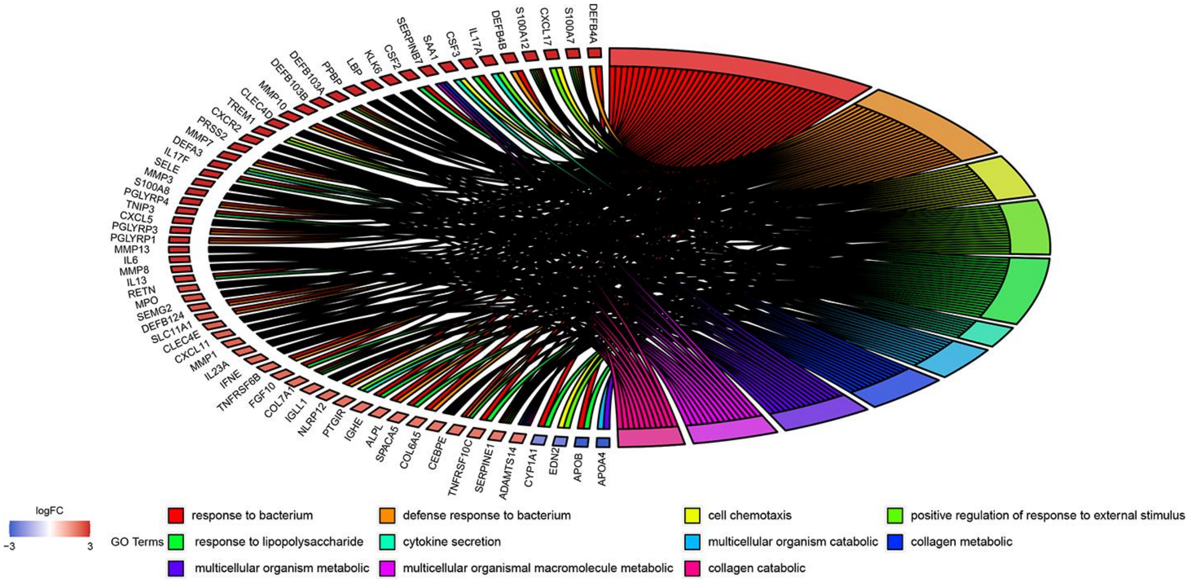

Figure 4. GO enrichment analysis of 908 DEGs. Genes are listed at the left side. Up-regulated genes are in red color, and down-regulated genes are in blue color conversely. The ligated bands between left and right side indicate that DEGs are related to the GO terms.

Weight" and mPS were calculated in the same way as GSE109142. Thereafter, we regarded the mPS of 74 samples as predicted values, and the clinical outcomes of UC as true values. The ROC-AUC of the predictive model in GSE92415 was 0.9506, and the PR-AUC was 0.9747.

DISCUSSION

UC has first emerged as a problem in industrialized countries and continued to rise in developing areas [16]. Besides the considerable cost, patients with active ulcerative colitis present high rates of fatigue, inferior

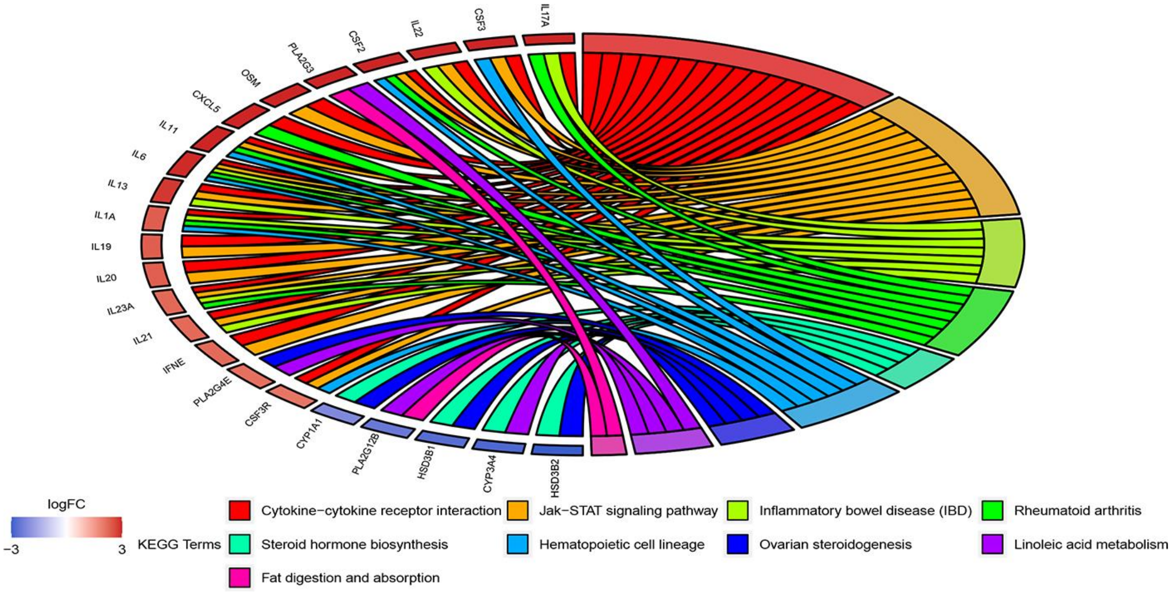

Figure 5. KEGG pathway enrichment analysis of 908 DEGs. The interpretation of Figure 5 is the same as Figure 4.

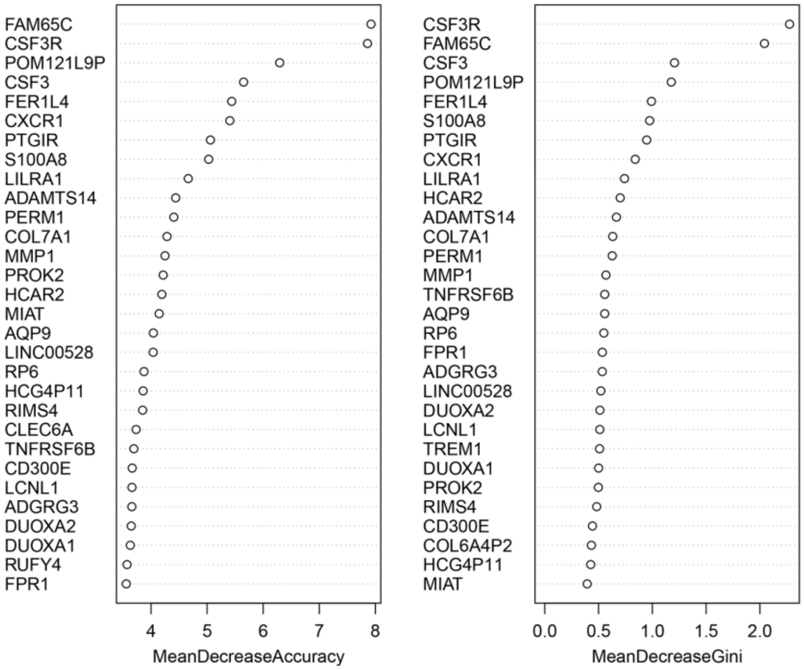

Figure 6. The screening results of the top-30 UC-related DEGs by random forest classifier. All the genes are sorted by the value of "Mean Decrease Accuracy" and "Mean Decrease Gini". The greater the two values are, the closer relationship with UC the DEG has.

health-related quality of life, and high disability [17]. Data also supports an association between UC and colorectal cancer [18]. Therefore, the necessity for UC management will be greater than ever. A prediction model based on valuable indices is intuitive and easy to use, which will facilitate decision making in clinical practice. In our study, we established an integrated toolkit of disease occurrence-related genes for UC, and developed a combination of machine learning algorithms to introduce a new clinical prediction score named mPS that is suitable for general UC patients. As a compatible and efficient scoring system, mPS proved useful in the evaluation of patient subpopulations in an independent platform.

The GO and KEGG enrichment analysis suggest that UC-related DEGs are involved in diversified GO terms and pathways, which reflects the dynamics and complexity of its pathogenesis. In practice, a link between alteration of intestinal flora and the occurrence of UC has been suggested previously [19]. Moreover, we observed that bacterial response-related biological processes are most prominent in GO terms. The best characterized pathway among the whole KEGG pathway is "cytokine–cytokine receptor interaction." As primary mediators of mucosal healing in IBD, cytokines including interleukin (IL)-1, IL-6, IL-10, and IL-13 occupy an important position in the Th2 atypical immune response of UC [20]. A marked difference in the levels of IL-4, IL-8, and IL-17A

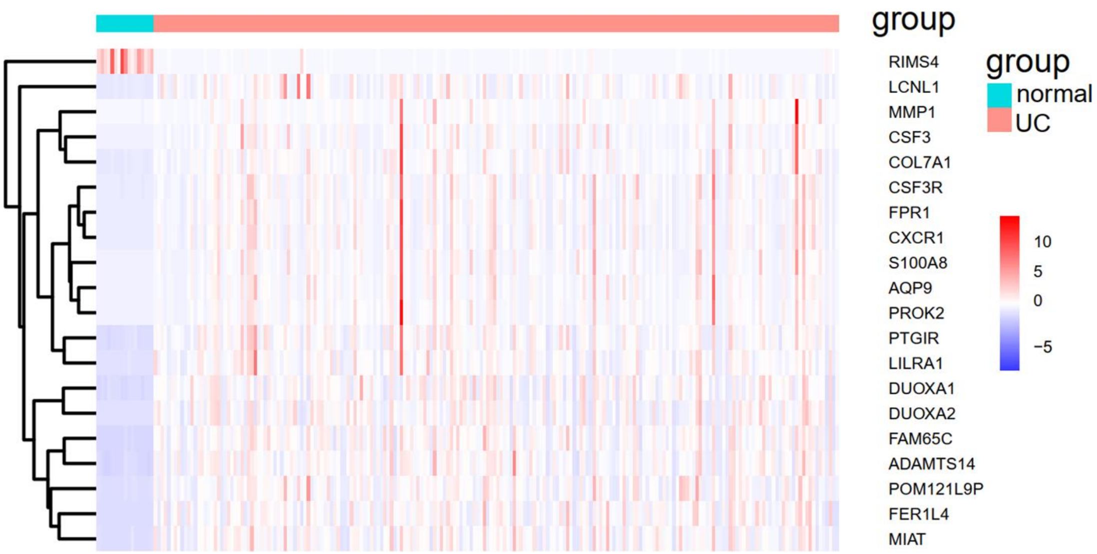

Figure 7. The heat map of the top-30 UC related DEGs in GSE109142.

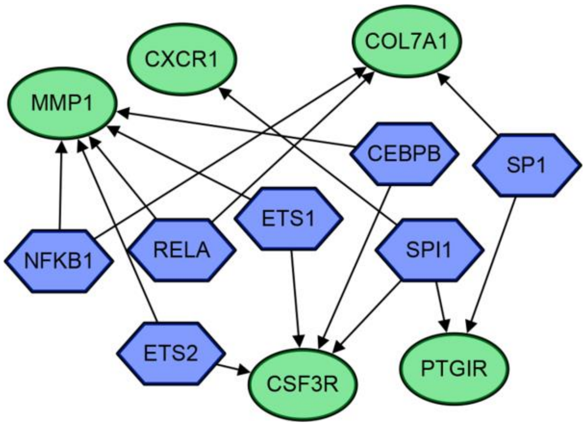

Figure 8. The gene regulatory network diagram of the 5 genes and 7 transcriptional regulators. The blue hexagons are transcriptional regulators, and the green circles are genes. The arrows indicate their regulatory relationships.

Table 2. The relationships between genes and transcriptional regulators predicted by TRRUST.

#	Key TF	# of overlapped genes	P value	Q value	List of overlapped genes
1	SPI1	3	0.0000497	0.000348	CXCR1, CSF3R, PTGIR
2	ETS2	2	0.00063	0.00221	MMP1, CSF3R
3	CEBPB	2	0.0022	0.00514	MMP1, CSF3R
4	ETS1	2	0.00379	0.00663	MMP1, CSF3R
5	RELA	2	0.0475	0.0561	MMP1, COL7A1
6	NFKB1	2	0.0481	0.0561	MMP1, COL7A1
7	SP1	2	0.104	0.104	COL7A1, PTGIR

Table 3. The "Gene_Weight" of 30 DEGs in GSE109142.

Gene symbol	Gene_Weight
DUOXA2	3.1869
S100A8	3.0181
TNFRSF6B	3.7603
MMP1	4.5608
CSF3R	3.0028
FER1L4	3.0258
COL7A1	3.6774
FPR1	3.0121
FAM65C	3.0217
AQP9	3.0115
MIAT	3.0328
PROK2	3.0119
RP6	4.5567
CSF3	4.0558
ADGRG3	3.0259
HCAR2	3.1662
CXCR1	3.0118
ADAMTS14	4.3368
PTGIR	3.0149
DUOXA1	3.2243
CD300E	4.3764
LILRA1	3.1524
HCG4P11	1.1187
LINC00528	3.7521
PERM1	4.2631
RUFY4	4.7713
POM121L9P	4.6224
LCNL1	4.5594
CLEC6A	3.0244
RIMS4	4.7128

were also reported among UC patients and control patients [21]. The correct interpretation of GO and KEGG enrichment analysis results supports the research of DEGs and their related metabolic pathways, leading to the discovery of new diagnostic indicators and therapeutic targets.

TRRUST analysis identified seven transcriptional regulators (SPI1, ETS2, CEBPB, ETS1, RELA, NFKB1, SP1) related to five DEGs (MMP1, CXCR1, COL7A1, CSF3R, PTGIR). Although it has been proved that the NFKB1 promoter polymorphism is not associated with UC [22, 23], the correlation between transcriptional regulators and DEGs are still worth exploring in further studies.

The innovative combination of machine learning approaches is the highlighted novelty of our study and has been used to improve the predictive ability of our UC predictive model, which has achieved good results in predictive ability creatively. As a form of an ensemble algorithm, RF has an outstanding performance on the processing of multiple-featured data with high accuracy and precision. The RF algorithm has been widely applied with success in the detection and prediction of clinical diseases such as Type 2 diabetes (AUC=0.89) and metabolic syndrome (accuracy>98%) [24, 25]. The ANN model has also been developed to predict the severity of IBD based on meteorological data and achieved high accuracy because of its self-learning ability and high efficiency [26]. Particularly, the cooperative machine learning approaches of RF and ANN was were reported to be efficient in many several data generation processes, including those with transcriptome profiles [27]. The development of mPS system has also been proved to be simple and cost-effective in the prediction of overall survival of brest cancer patients [15]. However, corresponding predictive models have not been used to predict the clinical outcome of UC patients. Previously, several systematic sets of molecular prognostic indicator related predictive models of UC have been established. Based on mRNA overexpression of NGAL in an inflamed intestine, Budzynska et al. used NGAL to predict the clinical and endoscopic activity of UC (AUC = 0.758) [3]. In a similar manner, Hart et al. combined FC with MES to predict endoscopic and histological activity of UC patients in clinical remission (AUC = 0.743) [4]. The mPS-based predictive model in our study shows higher overall accuracy and precision in both the training set and validation set (AUC = 0.9847 in GSE109142 and AUC = 0.9506 in GSE92415 respectively). In contrast to the two predictive models, machine learning and processing patterns are more reliable and accurate in data analysis. As data resources, we used downloaded UC gene expression profiles containing cases and controls to lower subjective errors during the calculation of the predictive model, such as the subjective nature of MES. Alternatively, the collection of UC gene expression profiles is much easier and more cost-effective than that of clinical patient information. With the AUC of the predictive model achieving 0.9506 in the independent validation set (GSE92415), the general applicability of the mPS system is also confirmed. RF classifier screening results identified CSF3R as the best-characterized DEG among all UC occurrence-related genes. It is worth noting that the upregulation of CSF3R is closely related to macrophage recruitment, which is regarded as a hallmark of IBD [28]. Therefore, investigations on the application of the mPS system in Crohn's disease is expected in future work.

However, our study still has several limitations. Firstly, numerous environmental factors have been linked with UC and may either increase the risk of or protect against developing this condition and can also become modifiers of UC course [29]. Our susceptibility gene based novel predictive model may has limited value in predictive power as others to some extent. Secondly, though we have validated the UC-related DEGs in the large expression profile GSE109142, the validation set GSE92415 is relatively small. The mPS system development and the predictive model was based on the dataset from the GEO, thus our predictive model should be practiced and verified in laboratory experiments and clinical work. In fact, there are a variety of protocols for the preservation of UC samples, which may present challenges to our original intention of building a "platform-independent" score system suitable to data acquired from any method. Comparison of different protocols and establishment of a universal and accurate method for the detection of the expression level of the 30 DEGs are still challenges that must be overcome before the application of the mPS system in clinical trials. Indeed, all analyses in our study were conducted in a retrospective manner. Regarding methodological rigor, prospective studies are also indispensable to validate our findings. We believe that the development of the mPS-based predictive model can not only optimize the screening and early

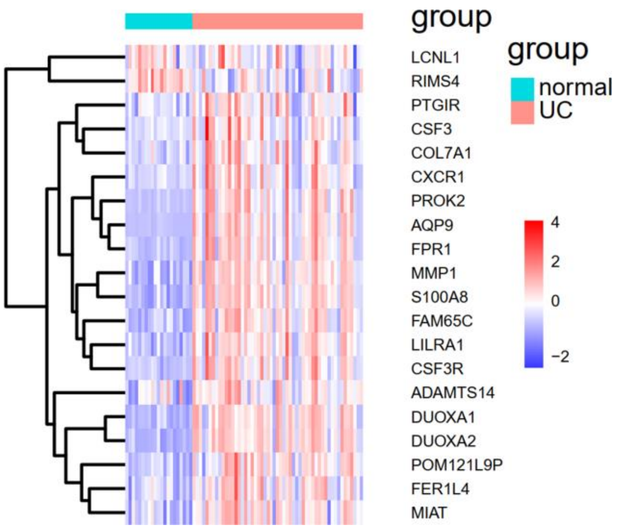

Figure 9. The heat map of the 20 overlapped genes between GSE109142 and GSE92415.

diagnosis of UC, but also prompt further biochemical research and provide new ideas for its clinical treatment.

MATERIALS AND METHODS

Study design and process

We downloaded two sets of UC gene expression profiles containing cases and controls (GSE109142 and GSE92415). Both sets are independent UC cohorts used multiple times and have proved to be accurate and reliable in recent UC-related research. The *t* test package in R version 3.5.3 was used to screen DEGs between cases and controls according to their specific threshold. Further analysis of DEGs included volcano plots, heat maps, gene ontology (GO) functions and KEGG pathway enrichment analyses. The RF algorithm was introduced to determine DEGs contributing the highest occurrence of UC. The analysis of transcription factors related to the regulation of these DEGs was also performed. Expression data of DEGs from all samples were transformed into gene expression scores. Based on these, we developed an ANN model to calculate the respective weight of DEGs to UC. Thereafter, we established an mPS system based on the expression data of these DEGs and verified it with GSE92415 to evaluate the accuracy of our predictive model.

Data resources

GSE109142 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109142) and GSE92415 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92415) were downloaded from the GEO database (Table 1). The data preprocessing was implemented as follows. The expression estimates in GSE109142 were measured in transcripts per million. The expression estimates in GSE92415 were measured in log2-transformed quantile-normalized signal intensity. We mapped probes to genes and removed the unloaded probes. If more than one probe corresponded to a certain gene, we considered the median probe value as the final expression value of the specific gene.

Screening for DEGs

The t test package in R version 3.5.3 was used to identify DEGs of GSE109142. We performed the analysis with set threshold: |log2(FC)| >2 and Benjamini-Hochberg adjusted $p < 0.05$. Finally, we gotacquired a list of 908 DEGs, including 781 upregulated DEGs and 127 downregulated DEGs.

Volcano plots and heat map

Volcano plots of all the genes in GSE109142 were drawn byusing the ggpubr package ("ggplot2" based publication ready plots) in R version 3.5.3. The heat map of DEGs in GSE109142 was drawn byusing the pheatmap package in R version 3.5.3. Among the DEGs, the red color indicatesd upregulation, and the blue color indicatesed downregulation.

GO function and KEGG pathway enrichment analysis

GO and KEGG pathway enrichment analysis were used to identify the functions of DEGs related to UC in GSE109142. These were performed on the DAVID website (http://david.abcc.ncifcrf.gov). The GOplot package in R version 3.5.3 was used for the visualization of the results.

Screening of diagnosis-related DEGs

RF is a general technique for the training and prediction of samples based on the classification tree. In our study, the randomForest package in R version 3.5.3 was used to determine the top-30 DEGs according to their contribution to the occurrence of UC. The screening threshold was set as mean decrease of accuracy ≥ 3.56 and mean decrease of gini index > 0.53. In this manner, the top-30 DEGs were analyzed on the TRRUST web server (https://www.grnpedia.org/trrust/) to find related transcriptional regulators that can be used to explain observed gene expression changes and identify potential signaling pathways involved in the pathogenesis of UC.

Calculation of gene_weight using ANN

The expression data of the 30 DEGs were first transformed into "Gene_Score" based on their expression level. In the case of a certain sample, the expression value of a specific gene was compared to the median of all sample expression values of that sample. If the expression value of the upregulated gene is higher, it will be valued as 1, otherwise 0. Similarly, if the expression value of downregulated gene is higher, it will be valued as 0, otherwise 1. The outcome variable was the occurrence of UC: cases were valued as 1, controls were valued as 0. Finally, we received a "Gene_Score" sheet containing 226 lines of samples, 30 columns of DEGs and 1 column of UC outcome variable (case/control).

The establishment of ANN was achieved by the Keras package in R version 3.5.3 (UC outcome variable = y, the expression values of DEGs = x). The rectified linear unit (RELU) was used as an activation function for thr hidden layers [30, 31]; correspondingly, softmax functions were used for the output layers [32]. Cross entropy was used as the error function [33]. Adam algorithm was introduced to optimize each "Gene_weight," and speed up training [34]. Following ANN training, the "Gene_Weight" of a specific DEG was valued as the maximum weight of that DEG in the hidden layer [15].

Construction and validation of the UC predictive model

The mPS scoring system played an important role in the construction of the UC predictive model. As a new type of scoring system, mPS has obtained success in precise prediction of overall survival of breast cancer patients [15]. "Gene_weight" of a specific DEG was achieved with the help of a combination of RF and ANN. The mPS of a certain sample was calculated by summation of "Gene_Score" × "Gene_Weight" for all 30 DEGs [15].

The independent dataset GSE92415 was used for the validation of the mPS scoring system based on GSE109142. We took the intersection of the 30 DEGs in GSE109142 and all genes in GSE92415 to obtain final lists of 20 overlapped genes. Within the intersection, the expression levels of 20 genes were transformed to binary status (above the median or below the median). As follows, "Gene_Score" and "Gene_weight" were valued according to the data processing methods mentioned above. The mPS of the validation set was also calculated by the summation of "Gene_Score" × "Gene_Weight". The ROCR package in R version 3.5.3 was used to calculate the AUC, which was regarded as an indicator to evaluate the predictive performance of our model.

AUTHOR CONTRIBUTIONS

Hanyang Li performed the statistical procedure and draft the article. Lijie Lai performed the statistical procedure and data proofreading. Jun Shen designed this study. Qi Feng helped edit the manuscript. The final version of the manuscript was approved by all authors.

CONFLICTS OF INTEREST

These authors declare no conflicts of interest.

FUNDING

Supported by grants from National Natural Science Foundation of China (No. 81770545) and MDT Project of Clinical Research Innovation Foundation, Renji Hospital, School of Medicine, Shanghai Jiaotong University (PYI-17-003)

REFERENCES

1. Ambrosini R, Barchiesi A, Di Mizio V, Di Terlizzi M, Leo L, Filippone A, Canalis L, Fossaceca R, Carriero A. Inflammatory chronic disease of the colon: how to image. Eur J Radiol. 2007; 61:442–48. https://doi.org/10.1016/j.ejrad.2006.07.028 PMID:17197146

2. Jess T, Gamborg M, Munkholm P, Sørensen TI. Overall and cause-specific mortality in ulcerative colitis: meta-analysis of population-based inception cohort studies. Am J Gastroenterol. 2007; 102:609–17. https://doi.org/10.1111/j.1572-0241.2006.01000.x PMID:17156150

3. Budzynska A, Gawron-Kiszka M, Nowakowska-Dulawa E, Spiewak J, Lesinska M, Kukla M, Waluga M, Hartleb M. Serum neutrophil gelatinase-associated lipocalin (NGAL) correlates with clinical and endoscopic activity in ulcerative colitis but fails to predict activity in Crohn's disease. J Physiol Pharmacol. 2017; 68:859–865. PMID:29550798

4. Hart L, Chavannes M, Kherad O, Maedler C, Mourad N, Marcus V, Afif W, Bitton A, Lakatos PL, Brassard P, Bessissow T. Faecal calprotectin predicts endoscopic and histological activity in clinically quiescent ulcerative colitis. J Crohns Colitis. 2020; 14:46–52. https://doi.org/10.1093/ecco-jcc/jjz107 PMID:31314884

5. Abraham C, Cho JH. Inflammatory bowel disease. N Engl J Med. 2009; 361:2066–78. https://doi.org/10.1056/NEJMra0804647 PMID:19923578

6. Zhang H, Massey D, Tremelling M, Parkes M. Genetics of inflammatory bowel disease: clues to pathogenesis. Br Med Bull. 2008; 87:17–30. https://doi.org/10.1093/bmb/ldn031 PMID:18753178

7. Chowers Y. Interaction of genetic, environmental and immune factors in the pathogenesis of inflammatory bowel diseases. Isr Med Assoc J. 2002; 4:815–17. PMID:12389349

8. Man Z, Feng GJPiMB. New progress in the pathogenesis of ulcerative colitis. Progress in Modern Biomedicine. 2010; 10:3160–65. https://www.cabdirect.org/cabdirect/abstract/20113085707

9. Eguchi R, Karim MB, Hu P, Sato T, Ono N, Kanaya S, Altaf-Ul-Amin M. An integrative network-based approach to identify novel disease genes and pathways: a case study in the context of inflammatory bowel disease. BMC Bioinformatics. 2018; 19:264. https://doi.org/10.1186/s12859-018-2251-x PMID:30005591

10. Vietri MT, Riegler G, De Paola M, Simeone S, Boggia M, Improta A, Parisi M, Molinari AM, Cioffi M. I219V polymorphism in hMLH1 gene in patients affected with ulcerative colitis. Genet Test Mol Biomarkers. 2009; 13:193–97. https://doi.org/10.1089/gtmb.2008.0088 PMID:19371218

11. Pei FH, Wang YJ, Gao SL, Liu BR, Du YJ, Liu W, Yu HY, Zhao LX, Chi BR. Vitamin D receptor gene polymorphism and ulcerative colitis susceptibility in han Chinese. J Dig Dis. 2011; 12:90–98. https://doi.org/10.1111/j.1751-2980.2011.00483.x PMID:21401893

12. Lebedev AV, Westman E, Van Westen GJ, Kramberger MG, Lundervold A, Aarsland D, Soininen H, Kłoszewska I, Mecocci P, Tsolaki M, Vellas B, Lovestone S, Simmons A, and Alzheimer's Disease Neuroimaging Initiative and the AddNeuroMed consortium. Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. Neuroimage Clin. 2014; 6:115–25. https://doi.org/10.1016/j.nicl.2014.08.023 PMID:25379423

13. Shi M, Xu G. Development and validation of GMI signature based random survival forest prognosis model to predict clinical outcome in acute myeloid leukemia. BMC Med Genomics. 2019; 12:90. https://doi.org/10.1186/s12920-019-0540-5 PMID:31242922

14. Grobman WA, Stamilio DM. Methods of clinical prediction. Am J Obstet Gynecol. 2006; 194:888–94. https://doi.org/10.1016/j.ajog.2005.09.002 PMID:16522430

15. Shimizu H, Nakayama KI. A 23 gene-based molecular prognostic score precisely predicts overall survival of breast cancer patients. EBioMedicine. 2019; 46:150–59. https://doi.org/10.1016/j.ebiom.2019.07.046 PMID:31358476

16. Kaplan GG. The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol. 2015; 12:720–27. https://doi.org/10.1038/nrgastro.2015.150 PMID:26323879

17. Fumery M, Singh S, Dulai PS, Gower-Rousseau C, Peyrin-Biroulet L, Sandborn WJ. Natural history of adult ulcerative colitis in population-based cohorts: a systematic review. Clin Gastroenterol Hepatol. 2018; 16:343–56.e3. https://doi.org/10.1016/j.cgh.2017.06.016 PMID:28625817

18. Bopanna S, Ananthakrishnan AN, Kedia S, Yajnik V, Ahuja V. Risk of colorectal cancer in Asian patients with ulcerative colitis: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol. 2017; 2:269–76. https://doi.org/10.1016/S2468-1253(17)30004-3 PMID:28404156

19. Sasaki M, Klapproth JM. The role of bacteria in the pathogenesis of ulcerative colitis. J Signal Transduct. 2012; 2012:704953. https://doi.org/10.1155/2012/704953 PMID:22619714

20. Roda G, Marocchi M, Sartini A, Roda E. Cytokine networks in ulcerative colitis. Ulcers. 2011; 2011:391787. https://doi.org/10.1155/2011/391787

21. Pearl DS, Shah K, Brown J, Shute JK, Trebble TM. Active ulcerative colitis is associated with downregulation of the TH1, TH2 and TH17 cytokine response and elevated IL-8 levels in mucosal biopsies. Gut. 2011; 60:A214. https://doi.org/10.1136/gut.2011.239301.450

22. Mirza MM, Fisher SA, Onnie C, Lewis CM, Mathew CG, Sanderson J, Forbes A. No association of the NFKB1 promoter polymorphism with ulcerative colitis in a british case control cohort. Gut. 2005; 54:1205–06. https://doi.org/10.1136/gut.2005.070029 PMID:16009698

23. Oliver J, Gómez-García M, Paco L, López-Nevot MA, Piñero A, Correro F, Martín L, Brieva JA, Nieto A, Martín J. A functional polymorphism of the NFKB1 promoter is not associated with ulcerative colitis in a spanish population. Inflamm Bowel Dis. 2005; 11:576–79. https://doi.org/10.1097/01.mib.0000161916.20007.76 PMID:15905705

24. López B, Torrent-Fontbona F, Viñas R, Fernández-Real JM. Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. Artif Intell Med. 2018; 85:43–49. https://doi.org/10.1016/j.artmed.2017.09.005 PMID:28943335

25. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. ScientificWorldJournal. 2015; 2015:581501. https://doi.org/10.1155/2015/581501 PMID:26290899

26. Peng JC, Ran ZH, Shen J. Seasonal variation in onset and relapse of IBD and a model to predict the frequency of onset, relapse, and severity of IBD based on artificial neural network. Int J Colorectal Dis. 2015; 30:1267–73. https://doi.org/10.1007/s00384-015-2250-6 PMID:25976931

27. Kong Y, Yu T. A deep neural network model using

random forest to extract feature representation for gene expression data classification. Sci Rep. 2018; 8:16477. https://doi.org/10.1038/s41598-018-34833-6 PMID:30405137

28. Degagné E, Pandurangan A, Bandhuvula P, Kumar A, Eltanawy A, Zhang M, Yoshinaga Y, Nefedov M, de Jong PJ, Fong LG, Young SG, Bittman R, Ahmedi Y, Saba JD. Sphingosine-1-phosphate lyase downregulation promotes colon carcinogenesis through STAT3-activated microRNAs. J Clin Invest. 2014; 124:5368–84. https://doi.org/10.1172/JCI74188 PMID:25347472

29. Dutta AK, Chacko A. Influence of environmental factors on the onset and course of inflammatory bowel disease. World J Gastroenterol. 2016; 22:1088–100. https://doi.org/10.3748/wjg.v22.i3.1088 PMID:26811649

30. Lomuscio A, Maganti LJaAI. An approach to reachability analysis for feed-forward ReLU neural networks. 2017. https://arxiv.org/abs/1706.07351

31. Agarap AFMJaN, Computing E. Deep Learning using Rectified Linear Units (ReLU). 2018. https://arxiv.org/abs/1803.08375

32. Rueckauer B, Lungu IA, Hu Y, Pfeiffer M, Liu SC. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. Front Neurosci. 2017; 11:682. https://doi.org/10.3389/fnins.2017.00682 PMID:29375284

33. Aurelio YS, De Almeida GM, De Castro CL, Braga AD. Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. 2019; 50:1937–49. https://doi.org/10.1007/s11063-018-09977-1

34. Kingma DP, Ba JJaL. Adam: A Method for Stochastic Optimization. 2014. https://arxiv.org/abs/1412.6980