Identification of four potential hematologic biomarkers for the diagnosis of Crohn's Disease

Minghui Wang
Medical College of Nanchang University
Yeyu Zhao
First Affiliated Hospital of Nanchang University
Qin-si Wan (ndyfy05036@ncu.edu.cn)
First Affiliated Hospital of Nanchang University

Research Article

Keywords: Crohn's disease, Diagnosis, Blood, Biomarkers, Bioinformatics analysis

Posted Date: May 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1612932/v1

Abstract

Objective

Diagnosis of Crohn's disease (CD) is often challenging and remains an urgent need for new diagnose options. This study aimed to identify new serological biomarkers and explored the role of these markers in the diagnosis of CD .

Method

Gene expression profiles in CD and normal for tissue and peripheral blood were downloaded from the Gene Expression Omnibus (GEO) database, and differentially expressed genes (DEGs) were identified by R software. Gene set enrichment analysis (GSEA), gene ontology(GO) enrichment and kyoto encyclopedia of genes and genomes(KEGG) analysis were performed to inspect the functional annotation of these DEGs. The overlapping DEGs that were up-/down-regulated both in the tissue and blood sample were used to develop diagnose model. ROC curve analysis was used to evaluate the model's performance.

Result

A total of 224 DEGs were identified between CD and normal tissue samples from the datasets GSE126124, GSE95095 and GSE16879. Enrichment analysis indicated that DEGs are mostly enriched in humoral immune response, regulation of inflammatory response, granulocyte migration, cell chemotaxis, cytokine-cytokine receptor interaction, chemokine signaling pathway, IL-17 signaling pathway, NOD-like receptor signaling pathway, TNF signaling pathway, complement coagulation cascades. Six haematologic/tissular-specific expressed genes(PDZK1IP1, STAT1, PI3, VAV1, PTGDS and LAIR2) were identified from the GSE119600 datasets. A proposed model with STAT1, PI3 ,VAV1 and PTGDS achieved 86.3% accuracy, 91.0% sensitivity and 77.1% specificity in the training set. The model also displayed good discriminative power with 96.4% sensitivity and 83.3% specificity in the validation set.

Conclusion

This work identified four genes as potential hematologic biomarkers of CD. The proposed model showed good performance which would be helpful for the early diagnosis of CD.

Introduction

Crohn's disease(CD) is a chronic relapsing inflammatory disease of the entire digestive tract frequently accompanied by cumulative tissue damage and complications such as stenosis, fistula, or abscess[1]. The life-long, incurable, disabling disease may occur in every group of ages, races, gender and

socioeconomic[2]. It should not be underestimated the heavy economic burden and psychological stress on the patients, their family, and the society with the rapidly increasing in the worldwide[3, 4].

The diagnosis of CD relies on a combination of symptoms, endoscopical, radiological, histological and biological features[5]. However, the initial symptoms, such as recurrent abdominal pain or diarrhea, are often mistaken for irritable bowel syndrome[6]. Due to unspecific symptoms and limited test accuracies, the span of delayed diagnosis can range from 5 to 24months or longer[7–9]. Accumulating evidence suggests that diagnostic delay is associated with lower quality of life, poor responses to drug therapy, more complications and consequent bowel resection[10, 11].

A timely and accurate diagnosis of CD is the prerequisite for effective treatment[12]. So, it is a considerable challenge to improve the early recognition of CD. Therefore,the aim of this study was to explore serum markers as potential and supplementary means for CD diagnosis.

Methods

Data acquisition and identification of DEGs

Four gene expression profiles (GSE126124, GSE95095, GSE16879 and GSE119600) were downloaded from the Gene expression omnibus (GEO) database using the “GEOquery” package[13]. The probe would be removed when there was more than one gene corresponding to the same probe. The largest expression value for the probe was retained when there was more than one probe corresponding to the same gene. The “limma” package[14]was used to conduct gene analysis on the normalized data to identify the DEGs between the CD and control samples with the threshold set as $|log2FC| > 1.0$ and adjusted $P < 0.05$. The differences in gene expression between the control and CD subjects were assessed using volcano plot and principal component analysis (PCA).

Enrichment analysis of of the overlapping DEGs

We identify the overlapping DEGs that were up- and down-regulated genes in the GSE126124, GSE95095 and GSE16879 datasets. Then we assessed the distribution trend of the overlapping DEGs to determine their contribution to the phenoty pe using c5.bp.v7.2.symbols.gmt [Gene ontology] gene sets for GSEA enrichment analyses. The significant enriched functions was selected with $p.adjust < 0.05$ and false discovery rate (FDR) < 0.25. We used the clusterProfiler package[15] to perform Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis on DEGs.

Screening and verification of diagnostic markers in blood

We also identify the overlapping DEGs that were up-/down-regulated in the tissue and blood sample. Form the GSE119600 datasets, we acquired the gene expression profiles in whole blood samples from 95 CD and 47 health controls. The data were randomly split into a training set (75%) for developing a diagnostic model and a testing set (25%) for validation. Their areas under the receiver operating

characteristic curves (AUC) were determined to evaluate the predicted value of the key genes and model using the IBM SPSS Statistics 25.

Results

Identification of DEGs

The characteristics of the available profiles are presented in Table 1. After data preprocessing, we used R software to extract a total of 458 DEGs from the gene expression matrix, as shown in the volcano map The two-dimensional PCA cluster diagrams after normalization (Fig. 1B, D, F) showed that the clustering of two groups between CD and control samples in three datasets were obvious indicating that the sample source was reliable. After screening with the threshold of an adjusted Pvalue < 0.05 and |logFC|>1, 1802 DEGs (1195 upregulated and 607 downregulated) were identified in the GSE16879 dataset, 823 DEGs (227 upregulated and 596 downregulated) were identified in GSE95095, and 135 DEGs (99 upregulated and 36 downregulated) were identified in GSE126124, as shown in the volcano maps(Fig. 1). Among these DEGs, 22 up-/down-regulated DEGs(CFB, SLC16A9, CFI, TGM2, C2, REG1A, REG1B, CXCL1, CHI3L1, ENTPD5, AQP8, CYR61, REG3A, HMGCS2, TNFAIP6, S100A9, IL1RN, SLC17A4, SLC26A2, TNC, MMP1,DEFA5) were consistently changed in the three datasets (Fig. 2).

Table 1
Characteristics of datasets in this study

Record	Molecule	Sample Type	Normal Control	Crohn's Disease	Platfrom
GSE16879	Total RNA	Mucosal biopsies	19	6	GPL570
GSE95095	Total RNA	Surgery biopsies	12	48	GPL14951
GSE126124	Total RNA	Colon biopsies	19	37	GPL6244
GSE119600	Total RNA	Whole blood	47	95	GPL10558

Enrichment analysis

In total, 224 DEGs were shared at least two datasets as identified through Venn diagram analyses(Fig. 2). We analyzed the correlation between these DEGs levels and the involving biological processes by GSEA. The screening criterion for significant gene sets was p.adjust < 0.05 and FDR < 0.25. We observed that most of the enriched gene sets were related to the antimicrobial humoral response, defense response, antimicrobial humoral immune response mediated by antimicrobial peptide, response to biotic stimulus, humoral immune response, inflammatory response, granulocyte chemotaxis, neutrophil chemotaxis, response to molecule of bacterial origin, granulocyte migration(Fig. 3). To further investigate the biological functions of these 224 DEGs, functional enrichment analyses were performed and results were shown in Fig. 4. GO analysis results showed that DEGs were mainly related to humoral immune response, response to lipopolysaccharide, response to molecule of bacterial origin, regulation of inflammatory

response, cellular response to lipopolysaccharide, cellular response to molecule of bacterial origin, cellular response to biotic stimulus, antimicrobial humoral response, cell chemotaxis and leukocyte migration. These DEGs were also enriched in KEGG pathways including the cytokine-cytokine receptor interaction, chemokine signaling pathway, IL-17 signaling pathway, viral protein interaction with cytokine and cytokine receptor, NOD-like receptor signaling pathway, pertussis, staphylococcus aureus infection, TNF signaling pathway, complement and coagulation cascades, rheumatoid arthritis.

Screening and verification of diagnostic markers in blood

Using Venn diagram analysis(Fig. 5), 6 candidate genes in the intersection of the above three datasets and GSE119600 were selected for further analysis, including PDZK1IP1, STAT1, PI3, VAV1, PTGDS and LAIR2. Then we split the dataframe into a training sample (n = 102) and validation sample (n = 40). After backward stepwise selection, STAT1 (AUC = 0.728, 95%CI: 0.626–0.829), PI3 (AUC = 0.655, 95%CI:0.541–0.768), VAV1 (AUC = 0.866, 95%CI: 0.795–0.937) and PTGDS (AUC = 0.766, 95%CI: 0.660–0.872) were included in the final model for predicting CD from normal control. As is shown in Fig. 6, the AUC, sensitivity and specificity of proposed model was 0.942 (95% CI: 0.899, 0.984), 91.0%, 77.1% on the training set, respectively. In the validation set, the model also displayed good discriminative power of predicting the CD (AUC 0.973, 95% CI: 0.929-1.000) with a sensitivity of 96.4% and specificity of 83.3%.

were identified in mucosal biopsies from CD and N samples of the GSE16879, GSE95095 and GSE126124 datasets. 177 DEGs in blood between CD and N samples were identified from the GSE119600 datasets. DEGs, differentially expressed genes. CD, Crohn's Disease. N, control samples.

Discussion

CD is a typical immune-mediated inflammatory diseases, which leads to the damage of gastrointestinal tract but also the rest of the body. Early diagnosis and treatment of CD will effectively avoid complications from progression of the disease and improve quality of life by limiting complications. However, making an accurate diagnosis of CD at early stage is extremely difficult due to the lack of typical clinical symptoms and effective single diagnostic test. So, identifying reliable hematologic biomarkers are crucial for assisting the early detection of CD.

In our study, we identified 224 DEGs by comparing genes expressed in CD and normal tissue samples. The DEGs were enriched in biological processes such as hnumoral immune response, inflammatory response, neutrophil chemotaxis, granulocyte migration, and were enriched in biological functions such as regulation of inflammatory response, cellular response to biotic stimulus,cell chemotaxis and leukocyte migration. KEGG pathway enrichment analysis revealed that these DEGs mainly affected cytokine-cytokine receptor interaction, chemokine signaling pathway, IL-17 signaling pathway, viral protein interaction with cytokine and cytokine receptor, NOD-like receptor signaling pathway, TNF signaling pathway, complement coagulation cascades.

We identified six haematologic and tissular specific expressed genes using four GEO datasets. ROC curve analysis suggests that combined with four genes have favourable diagnostic value for CD both in training and validation set. Therefore, we hypothesize that STAT1, PI3 ,VAV1 and PTGDS may be biomarkers for diagnosis of CD.

STAT proteins are dormant cytoplasmic transcription factors involved in signal transduction of several cytokines and growth factors. Expression and activation of STAT1 play an important role in chronic inflammatory diseases, including asthma and rheumatoid arthritis[16, 17]. The increased expression and activation of STAT1 was also found in the intestinal lamina propria of IBD patients, which was correlate with the endoscopic and histological activity of the disease. A study by Gunther et al. reported that IFNL-induced death of Paneth cells at the crypt bottom in inflamed ileum samples via the activation of STAT1[18]. Therefore, STAT1 may play an important role in the disease progression of CD.

VAV1, a member of the Vav family, is expressed only in the immune system and plays a critical role in the development and activation of T-cells[19]. The epistasis between Themis1(a new actor of TCR signaling) and Vav1 controls the occurrence of Treg suppressive function and predisposes to inflammatory bowel disease development[20]. In addition, there are no reports about PI3, PTGDS and IBD. So, we considered these three markers may be novel and effective biomarkers for the diagnosis of CD.

Conclusion

Our work identified for haematologic specific expressed genes as potential biomarkers of CD. A diagnostic model including STAT1, PI3 ,VAV1 and PTGDS showed good performance which would be helpful for the early diagnosis of CD. Still, the biomarkers need to be refined in large sample sizes, and diagnostic model need to be further confirmed.

Declarations

Disclosure statement

The authors report there are no competing interests to declare. Ming-Hui Wang and Ye-Yu Zhao are joint co-fifirst authors.

1. Ethics approval and consent to participate

GEO belong to public databases. The patients involved in the database have obtained ethical approval. Users can download relevant data for free for research and publish relevant articles. Our study is based on open source data, so there are no ethical issues and other conflicts of interest.And the study was conducted in accordance with relevant guidelines.

2. Consent for publication

Not Applicable

3. Availability of data and materials

The datasets analysed during the current study are available in the [Gene expression omnibus (GEO) database] repository.

(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126124;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95095;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16879;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119600)

The datasets used and/or analyzed during the current study are available from the author named Ming-hui Wang(2798325754@qq.com) on reasonable request.

4. Competing interests

The authors report there are no competing interests to declare.

5. Funding

This study was supported by National Natural Science Foundation of Jiangxi Province (20192BAB215007) and Research Training Foundation for Young Teachers of Nanchang University (4209-16100009-PY201918).

6. Authors' contributions

Qin-si Wan and Ming-hui Wang wrote the main manuscript text and Ye-yu Zhao prepared figures 1-6. All authors reviewed the manuscript.

7. Acknowledgements

Not Applicable

References

1. Benitez, J.M., et al., Role of endoscopy, cross-sectional imaging and biomarkers in Crohn's disease monitoring. Gut, 2013. 62(12): p. 1806–16.
2. Moazzami, B., K. Moazzami and N. Rezaei, Early onset inflammatory bowel disease: manifestations, genetics and diagnosis. Turk J Pediatr, 2019. 61(5): p. 637–647.
3. Ng, S.C., 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): p. 2769–2778.
4. Gianluca, C., From good to bad fibroblasts: New promising targets to cure Crohn's disease. EBioMedicine, 2021. 70: p. 103483.

5. Ballester, F.M., M.M. Bosca-Watts and P.M. Minguez, Crohn's disease. Med Clin (Barc), 2018. 151(1): p. 26–33.
6. Qari, Y.A., Clinical Characteristics of Crohn's Disease in a Cohort from Saudi Arabia. Saudi J Med Med Sci, 2022. 10(1): p. 56–62.
7. Nahon, S., et al., Diagnostic delay in a French cohort of Crohn's disease patients. J Crohns Colitis, 2014. 8(9): p. 964–9.
8. Vavricka, S.R., et al., Systematic evaluation of risk factors for diagnostic delay in inflammatory bowel disease. Inflamm Bowel Dis, 2012. 18(3): p. 496–505.
9. Al-Mofarreh, M.A. and I.A. Al-Mofleh, Emerging inflammatory bowel disease in saudi outpatients: a report of 693 cases. Saudi J Gastroenterol, 2013. 19(1): p. 16–22.
10. Zaharie, R., et al., Diagnostic Delay in Romanian Patients with Inflammatory Bowel Disease: Risk Factors and Impact on the Disease Course and Need for Surgery. J Crohns Colitis, 2016. 10(3): p. 306–14.
11. Schoepfer, A.M., et al., Diagnostic delay in Crohn's disease is associated with a complicated disease course and increased operation rate. Am J Gastroenterol, 2013. 108(11): p. 1744-53; quiz 1754.
12. Fiorino, G. and S. Danese, Diagnostic Delay in Crohn's Disease: Time for Red Flags. Dig Dis Sci, 2016. 61(11): p. 3097–3098.
13. Davis, S. and P.S. Meltzer, GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics, 2007. 23(14): p. 1846–7.
14. Ritchie, M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 2015. 43(7): p. e47.
15. Yu, G., et al., clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS, 2012. 16(5): p. 284–7.
16. Yokota, A., et al., Preferential and persistent activation of the STAT1 pathway in rheumatoid synovial fluid cells. J Rheumatol, 2001. 28(9): p. 1952–9.
17. Lu, D., et al., IL27 suppresses airway inflammation, hyperresponsiveness and remodeling via the STAT1 and STAT3 pathways in mice with allergic asthma. Int J Mol Med, 2020. 46(2): p. 641–652.
18. Gunther, C., et al., Interferon Lambda Promotes Paneth Cell Death Via STAT1 Signaling in Mice and Is Increased in Inflamed Ileal Tissues of Patients With Crohn's Disease. Gastroenterology, 2019. 157(5): p. 1310–1322.e13.
19. Oberley, M.J., D.S. Wang and D.T. Yang, Vav1 in hematologic neoplasms, a mini review. Am J Blood Res, 2012. 2(1): p. 1–8.
20. Pedros, C., et al., An Epistatic Interaction between Themis1 and Vav1 Modulates Regulatory T Cell Function and Inflammatory Bowel Disease Development. J Immunol, 2015. 195(4): p. 1608–16.

Figures

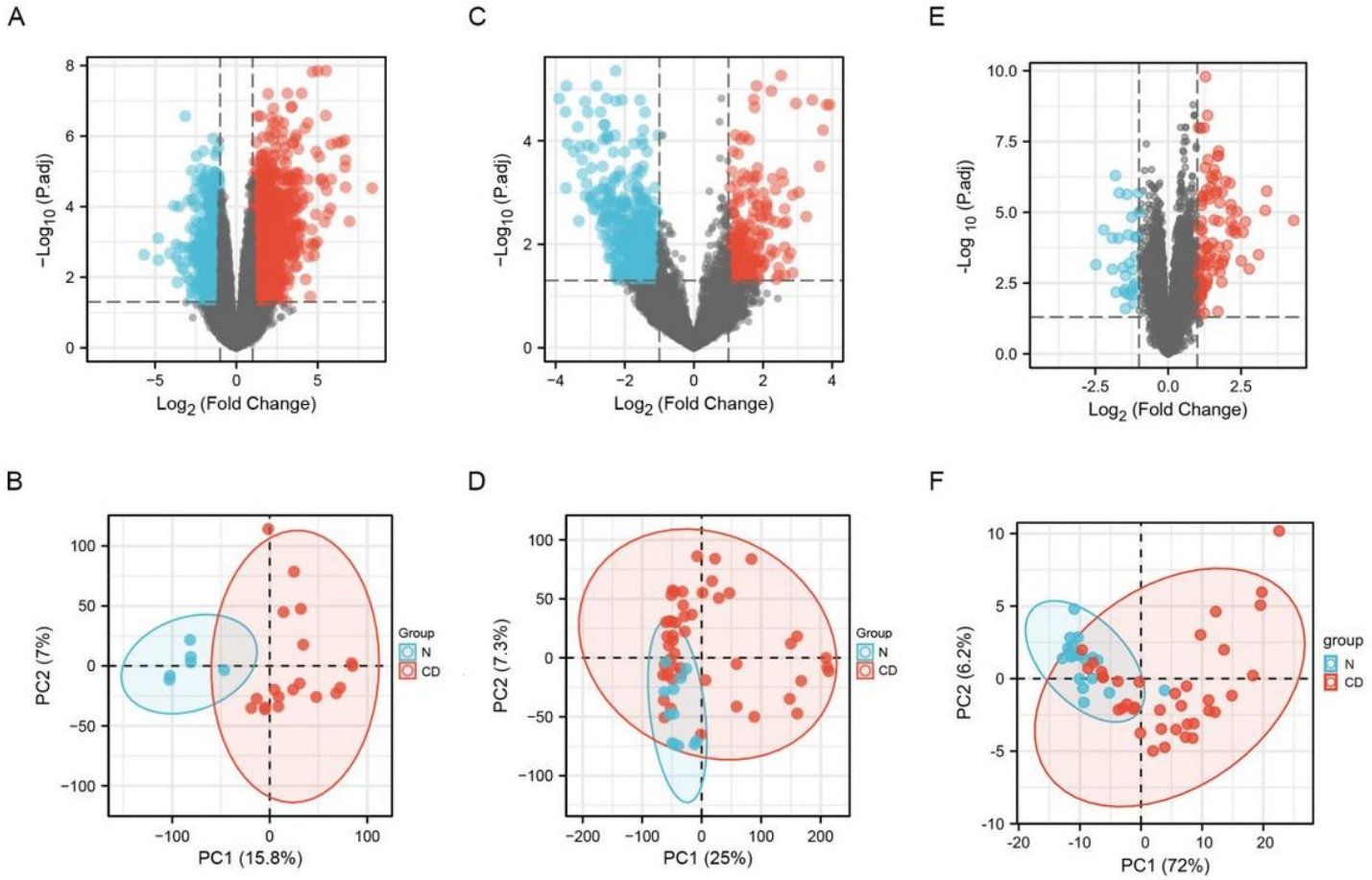

Figure 1

Differentially expressed genes. Volcano plot of DEGs between the CD and N samples of the GSE16879(A), GSE95095(C) and GSE126124(E) datasets. PCA plot of DEGs between the CD and N samples of the GSE16879(B), GSE95095(D) and GSE126124(F) datasets. DEGs, differentially expressed genes. CD, Crohn's Disease. N, control samples. PCA, principal component analysis.

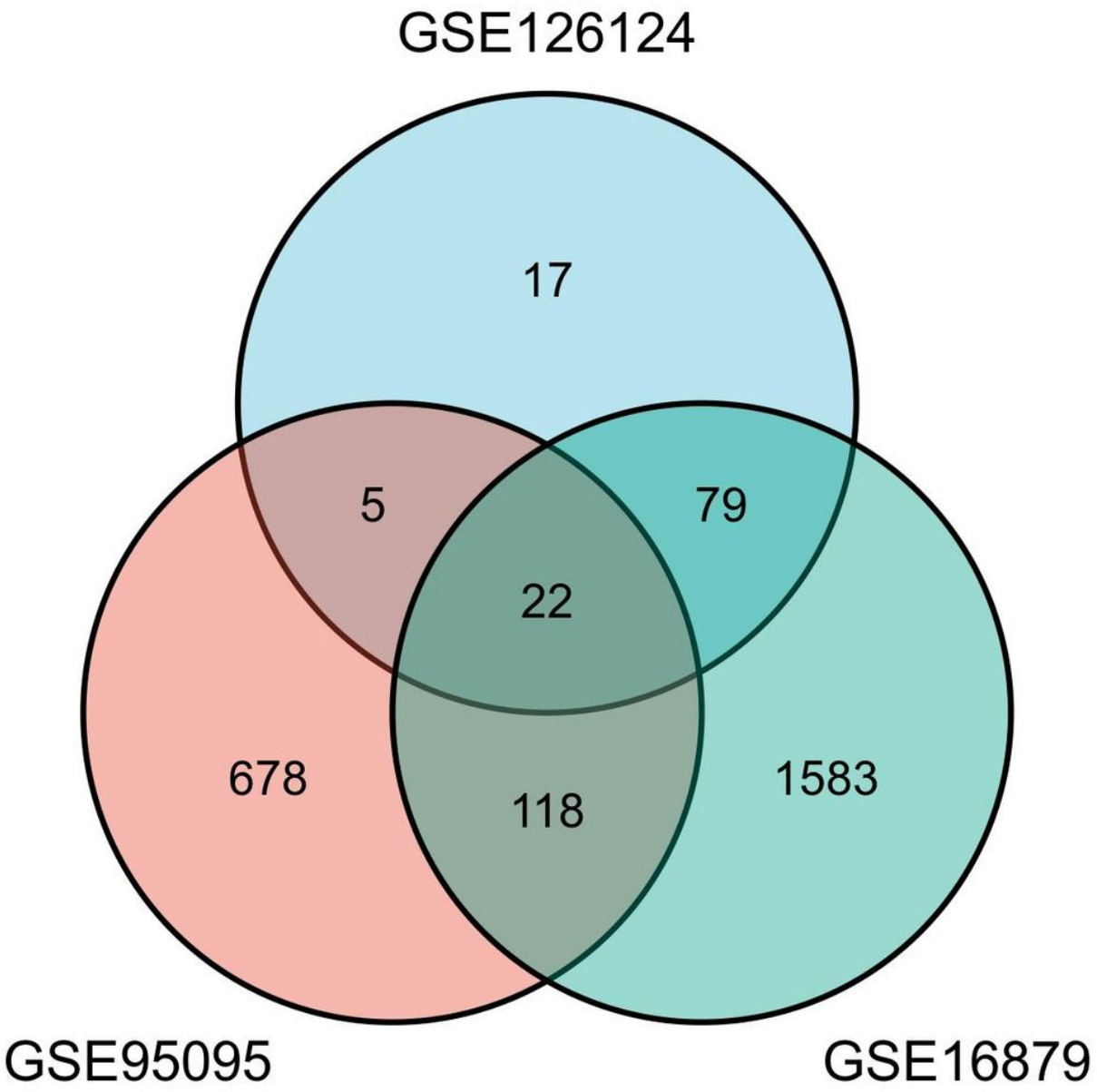

Figure 2

Co-expression of the differentially expressed genes of the GSE16879, GSE95095 and GSE126124 expression groups.

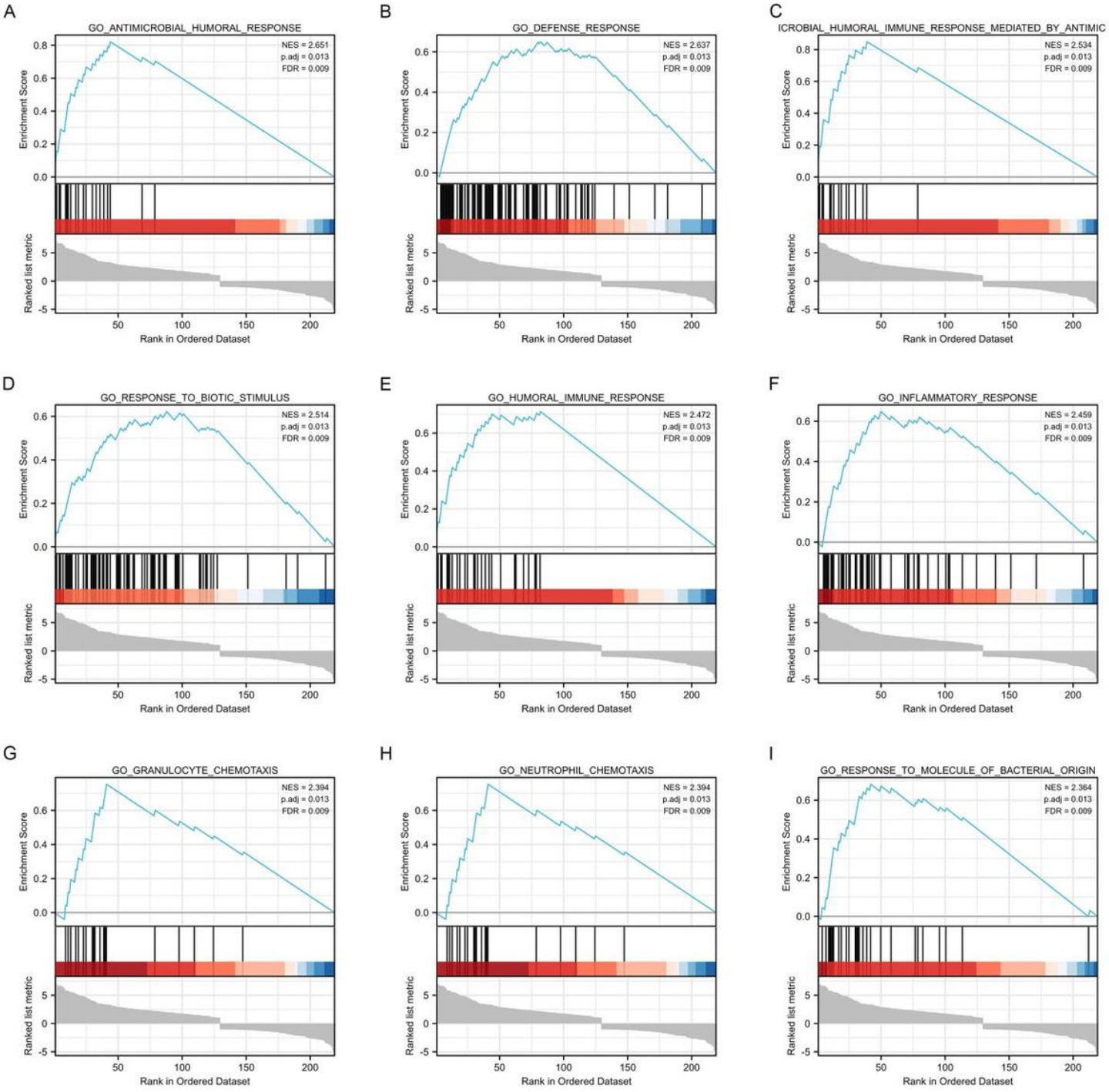

Figure 3

Partial display of the GSEA analysis results. NES, normalized enrichment score. p.adjust, adjust p value. FDR, false discovery rate. CD, Crohn's Disease. N, control samples.

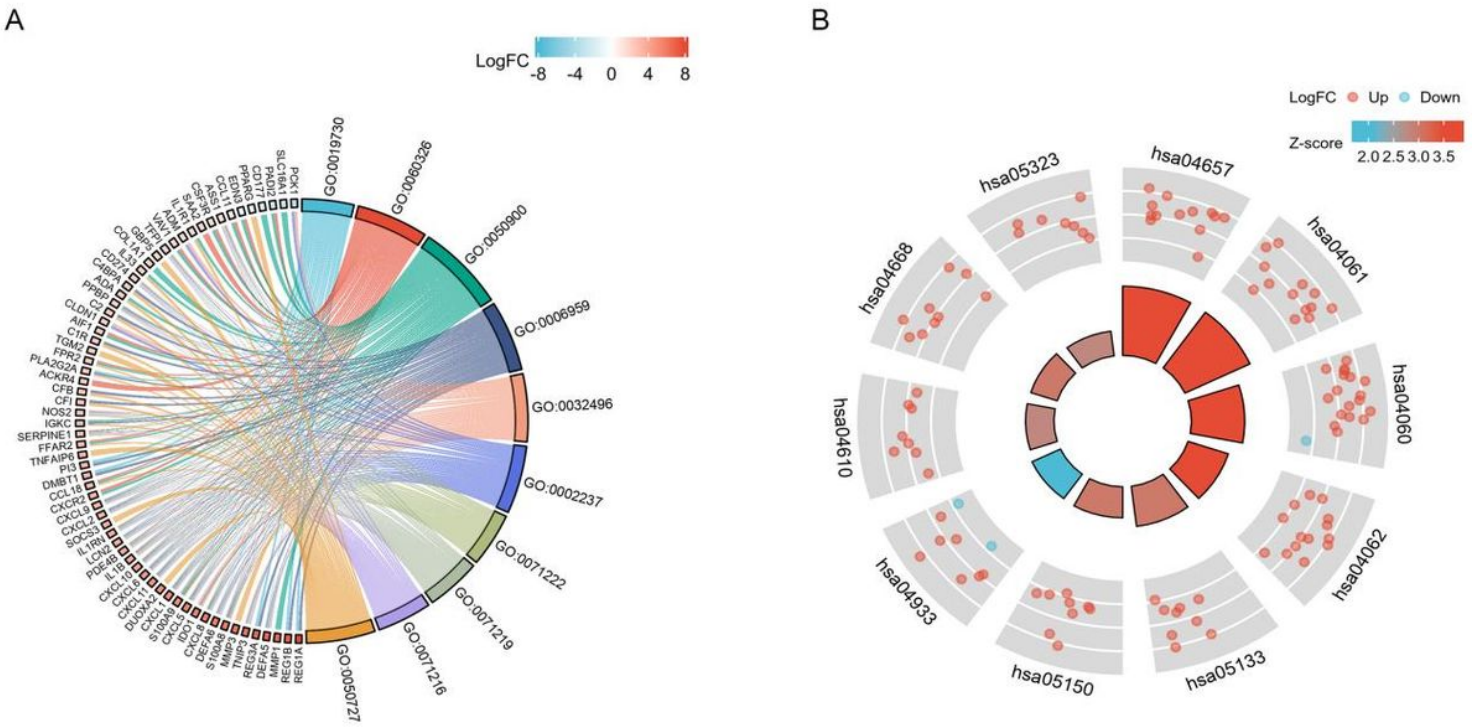

Figure 4

GO and KEGG pathway analyses of DEGs. (A) The chord plot showing the top 10 enriched biological processes of DEGs. GO:0006959, humoral immune response. GO:0032496, response to lipopolysaccharide. GO:0002237, response to molecule of bacterial origin. GO:0050727,regulation of inflammatory response. GO:0071222, cellular response to lipopolysaccharide. GO:0071219,cellular response to molecule of bacterial origin. GO:0071216, cellular response to biotic stimulus. GO:0019730, antimicrobial humoral response. GO:0060326, cell chemotaxis. GO:0050900, leukocyte migration. (B) The loop graph showing the top 10 enriched KEGG pathways of DEGs. hsa04060, Cytokine-cytokine receptor interaction. hsa04062, Chemokine signaling pathway. hsa04657, IL-17 signaling pathway. hsa04061, Viral protein interaction with cytokine and cytokine receptor. hsa04621, NOD-like receptor signaling pathway. hsa05133, Pertussis. hsa05150, Staphylococcus aureus infection. hsa04668, TNF signaling pathway. hsa04610, Complement and coagulation cascades. hsa05323, Rheumatoid arthritis.

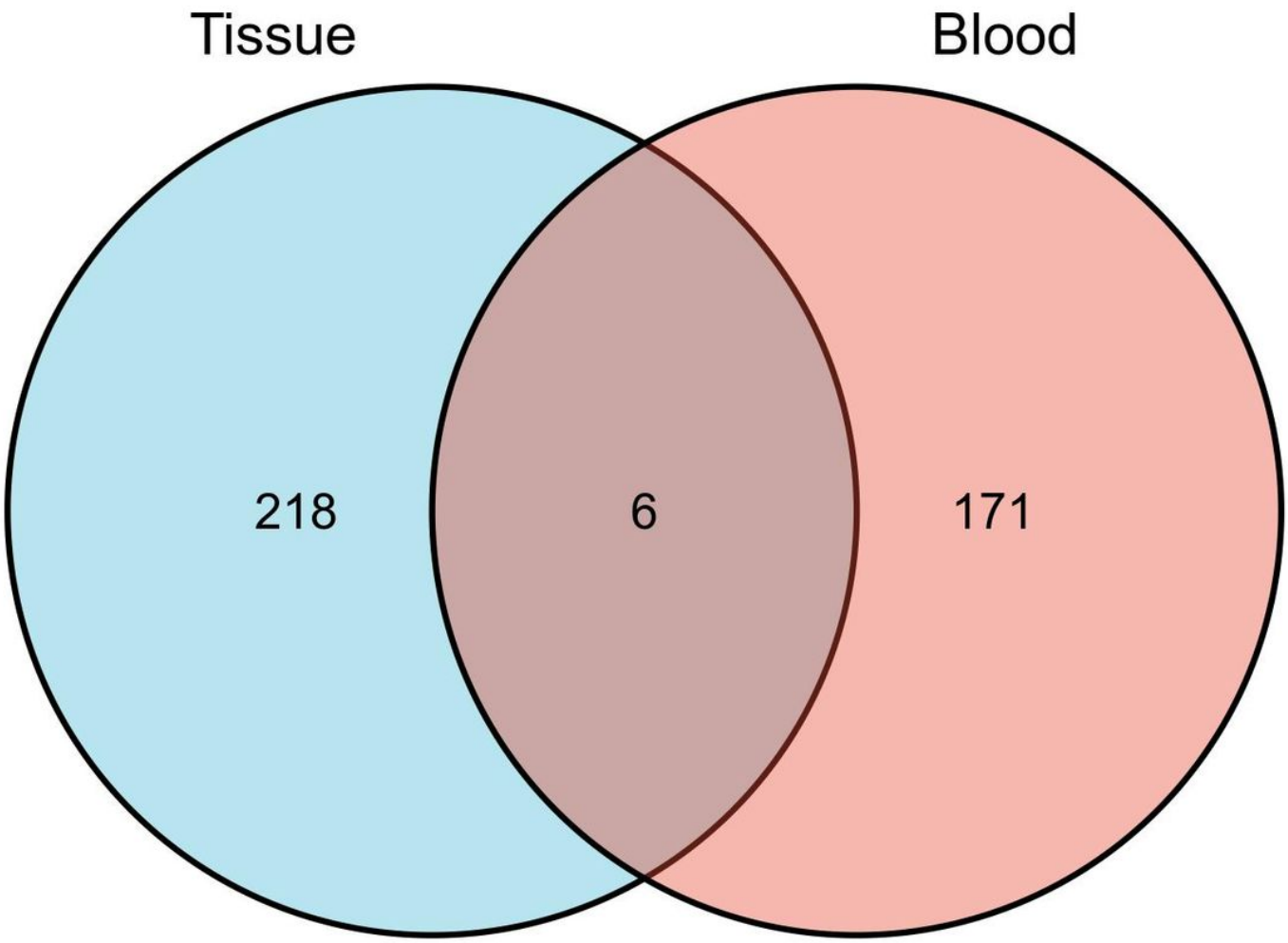

Figure 5

Co-expression of the differentially expressed genes of tissue and blood groups. DEGs were identified in mucosal biopsies from CD and N samples of the GSE16879, GSE95095 and GSE126124 datasets. 177 DEGs in blood between CD and N samples were identified from the GSE119600 datasets. DEGs, differentially expressed genes. CD, Crohn's Disease. N, control samples.

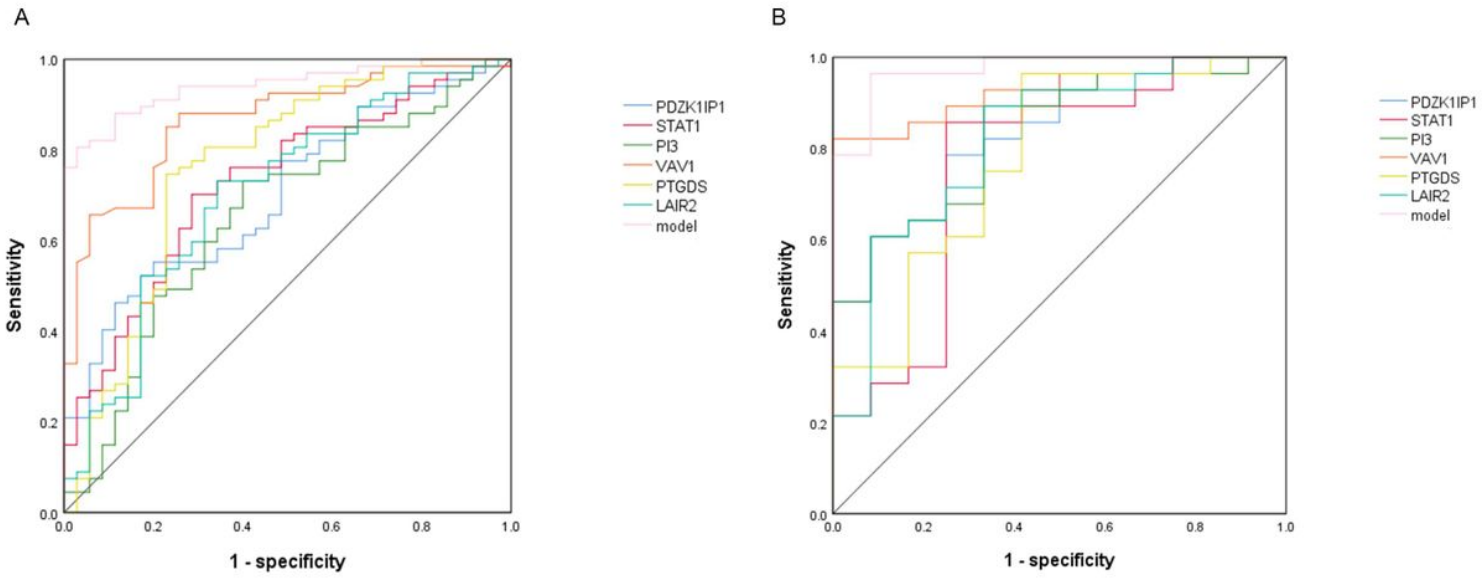

Figure 6

ROC curve of the 6 specifically expressed hub genes and model for detection CD. (A) ROC curve in training set. (B) ROC curve in validation set.