**ORIGINAL ARTICLE**
Iran J Allergy Asthma Immunol
In press.

# *APP*, *CXCL8*, *DMBT1*, *SAA1*, and *SLC40A1* as Immune-derived Biochemical Biomarkers for Crohn's Disease: Transcriptome-based Validation

**Xiaoxuan Xie, Jinyong Liu, Zhihan Liu, and Wei Yang**

*Department of Proctology, Shuguang Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China*

***Received: 9 May 2026; Received in revised form: 24 June 2026; Accepted: 3 July 2026***

## ABSTRACT

Intestinal immune dysregulation is a core biochemical feature of Crohn's disease (CD), yet reliable mucosa-derived biochemical biomarkers for the differential diagnosis of CD remain insufficiently defined.

Using 2 independent public transcriptome datasets (GSE20881 as the training cohort and GSE102133 for independent cross-dataset evaluation), we screened immune-related differentially expressed genes between CD intestinal mucosa and normal controls. After intersecting differential genes with ImmPort-derived immune gene sets, functional enrichment and 3 machine-learning algorithms (least absolute shrinkage and selection operator [LASSO] regression, random forest, and support vector machine-recursive feature elimination [SVM-RFE]) were sequentially applied to filter candidate biochemical markers.

In total, 42 immune-associated differentially expressed transcripts were identified, mainly enriched in complement activation, neutrophil degranulation, Toll-like receptor, and interleukin-mediated inflammatory cascades. Finally, 5 core immune biochemical biomarkers (*APP*, *CXCL8*, *DMBT1*, *SAA1*, and *SLC40A1*) were verified; the combined 5-gene panel yielded an area under the curve (AUC) of 0.835 in the training set and 0.990 in the validation cohort for discriminating CD from healthy mucosa.

Collectively, this immune gene signature provides novel candidate transcriptional biomarkers for exploratory auxiliary CD research, but further prospective biochemical and clinical specimen validation are required before any routine clinical application.

**Keywords:** *APP*; Biochemical diagnostic biomarker; Crohn's disease; *CXCL8*; Intestinal immunity; *SAA1*

## INTRODUCTION

Crohn's disease (CD) is a chronic inflammatory disorder of the intestine, usually marked by repeated remission and relapse. During the past several decades, inflammatory bowel disease (IBD) has increased in many regions, and the disease burden is no longer confined to Western countries.1 In daily practice, CD does not present in 1 fixed way.

Some patients mainly have abdominal pain and diarrhea, while others present with weight loss, fever, anemia, perianal lesions, or extraintestinal symptoms.

**Corresponding Author:** Wei Yang, MD;
Department of Proctology, Shuguang Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China.
Tel: (+86 21) 5382 7400, Email: 13091588927@163.com

The first and second authors contributed equally to this study

Copyright©, Iran J Allergy Asthma Immunol. All rights reserved.

Published by Tehran University of Medical Sciences (http://ijaai.tums.ac.ir)

IN PRESS

The bowel segment involved also varies. For this reason, diagnosis usually depends on a combination of symptoms, endoscopy, histology, imaging, and laboratory findings.[2,3]

Current diagnostic approaches are helpful, but they still have practical limits. Ileocolonoscopy and histological examination can show mucosal lesions directly, and cross-sectional imaging helps assess disease extent and complications. These examinations, however, are not always easy to repeat and may be affected by patient tolerance, cost, and differences in interpretation.[4] C-reactive protein (CRP), fecal calprotectin, and lactoferrin are commonly used to evaluate inflammatory activity, but their performance may differ according to disease location, inflammatory burden, and individual variation.[5] It is therefore reasonable to look at mucosal molecular information as a supplement to existing clinical tools, rather than as a replacement for them.

The biology of CD is layered. Genetic susceptibility, intestinal microbiota, epithelial barrier damage, and immune dysregulation have all been linked to disease onset and progression.[6] Large genetic studies have reported susceptibility loci related to host defense, autophagy, cytokine signaling, and immune cell function.[7,8] These findings make it clear that CD cannot be explained by 1 gene or 1 pathway. More often, the disease reflects an altered balance between epithelial cells, immune cells, microbes, and stromal components in the intestinal mucosa.[9]

Mucosal immunity is therefore a useful perspective for studying CD. Cytokines can maintain intestinal inflammation and influence tissue injury, repair, and complications.[10] The intestinal epithelium also takes part in antimicrobial defense, barrier maintenance, and communication with immune cells, rather than simply acting as a physical wall.[11] When mucosal inflammation develops, changes in epithelial defense, leukocyte recruitment, antimicrobial response, and cytokine-related signaling may be captured by transcriptome data. Previous studies of ileal mucosal expression have described CD-related patterns involving host-microbe interaction, oxidative stress, epithelial changes, and immune activation.[12,13] These studies support the value of mucosal expression data, although their results still need cautious interpretation.

Public transcriptome databases allow existing mucosal expression data to be reanalyzed with a focused question. The Gene Expression Omnibus (GEO) stores many functional genomics datasets, including expression profiles from IBD studies.[14] For an immune-related study, it is also necessary to use a defined immune gene source, so that the analysis is not based on subjective gene selection. ImmPort is one such resource for immunology-related data.[15] In the present work, the analysis was designed to first compare CD with normal intestinal mucosa, and then narrow the results to immune-related genes by intersecting them with the ImmPort gene set. This sequence maintains alignment with the study objective.

After immune-related differentially expressed genes are obtained, the next step is to reduce the list without relying on visual selection. Transcriptome data contain many variables, and a smaller candidate panel should be produced through a more structured process. Least absolute shrinkage and selection operator (LASSO) regression can shrink coefficients and retain fewer variables.[16] Random forest can rank variables according to their contribution to classification.[17] Support vector machine-recursive feature elimination (SVM-RFE) has also been used for feature selection in high-dimensional expression data.[18] These methods do not establish clinical utility, but they can help form a candidate set that can then be tested in another dataset.

Based on this background, this study analyzed differentially expressed immune genes between CD and normal intestinal mucosa using public transcriptome datasets. The analysis followed a planned route, including preprocessing, differential expression analysis, immune-gene intersection, functional enrichment, machine-learning screening, and independent cross-dataset evaluation. The aim was to describe mucosal immune-related transcriptional changes and to identify candidate biomarkers for further study, without claiming immediate clinical application

We selected ImmPort as the exclusive source of immune-related genes for 2 key reasons: (1) ImmPort manually curated immune genes covering adaptive immunity, innate immunity, chemokine signaling, and epithelial immune crosstalk, which fully matched the mucosal immune research focus of this study; (2) ImmPort provides standardized immune gene classification without redundant overlapping gene entries compared with alternative databases. MSigDB immune gene sets contain broader non-tissue-specific immune transcripts mixed with stromal and metabolic genes; InnateDB only focuses on innate immune genes and lacks adaptive immune modules relevant to CD

IN PRESS

pathogenesis. We acknowledged that cross-database comparison may enrich gene screening scope, and multi-database immune gene integration will be adopted in future biomarker screening work.

## MATERIALS AND METHODS

### Data Source

Transcriptome data were obtained from the GEO database. Datasets were considered if they contained intestinal mucosal samples from patients with CD and corresponding normal or non-IBD control samples. GSE20881 was used as the training dataset because it included both CD and control mucosal samples with a relatively large sample size. GSE102133 was used as an independent evaluation dataset.

GSE20881 was based on the GPL1708 platform and contained 99 CD samples and 73 normal control samples. GSE102133 was based on the GPL6244 platform and contained 65 CD samples and 12 normal control samples from ileal mucosa. All data were downloaded from public databases, and no additional patient samples were collected.

Original sample metadata of GSE20881 and GSE102133 recorded only endoscopic mucosal inflammation status, without unified clinical disease activity indices (Crohn's Disease Activity Index [CDAI], Simple Endoscopic Score for Crohn's Disease [SES-CD]) for each CD sample, so subgroup analysis based on active/inactive disease could not be performed. All control mucosal samples in GSE20881 and GSE102133 were collected from healthy volunteers without IBD, infectious enteritis, or chronic gastrointestinal disorders, representing true healthy intestinal mucosa rather than non-IBD inflammatory controls.

### Data Preprocessing and Probe Annotation

All transcriptome analysis was completed using R software (v4.2.1). Key R packages included limma (v3.54.0), glmnet (v4.1-4), randomForest (v4.7-1.1), e1071 (v1.7-11), clusterProfiler (v4.6.2), and pheatmap (v1.0.12). The series matrix files and platform annotation files were downloaded for both datasets. Sample records were checked according to the original GEO information, and samples were divided into CD and normal control groups. Probe IDs in GSE20881 were annotated using the GPL1708 platform file. Probe IDs in GSE102133 were annotated using the GPL6244 annotation file.

Probes without valid gene symbols were removed. For multiple probes mapping to the same official gene symbol, we retained the probe with the median expression value across all samples as the representative expression level of the target gene; probes with low average expression (average signal < 20% quantile of all probes) were excluded in advance before probe collapsing. After preprocessing, GSE20881 was used for differential expression analysis and candidate gene screening, while GSE102133 was reserved for independent cross-dataset evaluation. Batch effect correction was not performed between GSE20881 (GPL1708) and GSE102133 (GPL6244), as the 2 datasets were independently processed as training and independent cross-dataset evaluation cohorts without merged expression matrix construction. Within each dataset, quantile normalization was implemented to reduce intra-platform variation among individual samples. Cross-platform expression differences were addressed by evaluating consistent differential expression trends rather than absolute expression values during biomarker screening.

### Differential Expression Analysis

Differential expression analysis was performed in GSE20881 by comparing CD samples with normal intestinal mucosal samples. Because the GSE20881 samples came from different intestinal sites (ileum, colon, rectum), biopsy anatomical location was incorporated as a fixed covariate in the linear regression model for differential expression analysis to reduce tissue location confounding effects. The full design matrix included binary group variable (1 = CD, 0 = healthy control) and multiple dummy variables corresponding to distinct biopsy sites; the model formula was: Gene expression ~ Group + Biopsy_Site. Genes with FDR < 0.05 and $|\log FC| \geq 0.2$ were regarded as differentially expressed genes. This relatively permissive $\log_2$-fold change threshold was selected for 2 reasons: (1) intestinal mucosal immune transcriptional changes in mild/moderate CD are often mild rather than sharply altered; strict $|\log FC| \geq 0.5$ would eliminate a large number of weakly regulated immune-related transcripts participating in innate immune and epithelial defense pathways; (2) consistent with previous bulk transcriptome studies of IBD mucosa that retained subtle immune gene alterations for subsequent machine learning screening, avoiding premature filtering of potential biomarker candidates.

IN PRESS

Principal component analysis (PCA) was used to view the overall distribution of samples. Volcano plots and heatmaps were drawn to show the distribution of differentially expressed genes and the expression pattern of representative genes. This was the first screening step of the study. No diagnostic marker gene was specified before this analysis.

**Identification of Differentially Expressed Immune-related Genes**

Immune-related genes were collected from the ImmPort database. The differentially expressed genes identified in GSE20881 were intersected with the ImmPort immune-related gene set. Genes present in both lists were defined as differentially expressed immune-related genes.

The expression pattern of these genes was shown by heatmap, and a volcano plot was used to display their distribution according to logFC and statistical significance. Upregulated and downregulated immune-related genes were counted separately. This step narrowed the analysis from all expression changes to immune-related transcriptional changes.

**Functional Enrichment Analysis**

Functional enrichment analysis was carried out for the differentially expressed immune-related genes. Gene Ontology (GO) biological process enrichment was used to describe related biological functions. Reactome pathway enrichment was used to examine known immune and inflammatory pathways involving these genes.

Enriched terms were ranked by statistical significance and displayed in enrichment plots. This analysis was used to describe the functional background of the immune-related gene set before marker screening, not to justify a preselected group of genes.

**Machine-learning Screening and Diagnostic Validation**

The differentially expressed immune-related genes were used as the input set for candidate diagnostic gene screening. Three machine-learning algorithms were applied: LASSO regression, random forest, and SVM-RFE, all implemented with 5-fold cross-validation repeated 10 times to reduce the risk of overfitting. LASSO regression (glmnet package, alpha = 1) was used to shrink redundant variables and retain genes with non-zero regression coefficients. Random forest (randomForest package, ntree = 1000, mtry = sqrt(total feature number)) ranked genes by mean decrease accuracy as feature importance score for group classification. SVM-RFE (e1071+caret package, linear kernel) provided an additional feature-selection result from the same gene set by recursively eliminating low-contribution immune genes.

Final core candidates were selected from the machine-learning candidate pool after considering expression-direction consistency between GSE20881 and GSE102133, evaluability in both datasets, and biological relevance to mucosal inflammation. Because GSE102133 contributed to this cross-dataset filtering step, it is described here as an independent evaluation cohort rather than a fully untouched external validation cohort.

Receiver operating characteristic (ROC) analysis was used to evaluate the diagnostic discrimination of the final candidate panel. A combined score based on the selected genes was calculated in GSE20881 and tested in GSE102133. Expression levels of the selected genes were also compared between CD and normal control samples in both datasets. The validation results were treated as evidence from public transcriptome data, not as clinical validation.

The five-gene diagnostic score was constructed using APP, CXCL8, DMBT1, SAA1, and SLC40A1. Gene expression values were standardized using the mean and standard deviation of the GSE20881 training cohort. The model included an intercept of −1.35, with coefficients of 0.42, 0.71, 0.38, 0.56, and −0.31 assigned to APP, CXCL8, DMBT1, SAA1, and SLC40A1, respectively. The weighted expression values and intercept were combined to generate a diagnostic score, which was converted into the predicted probability of Crohn's disease using a logistic transformation. In this analysis, the coefficients and training-derived standardization parameters were specified as fixed before evaluation in GSE102133, without further model refitting.

## RESULTS

**Dataset Overview and Analysis Framework**

Two public transcriptome datasets were included. GSE20881 served as the training dataset, and GSE102133 was used for independent cross-dataset evaluation. The analysis moved from data preprocessing and group annotation to differential expression analysis,

IN PRESS

immune-related gene screening, enrichment analysis, machine-learning gene selection, and validation of the candidate panel.

The workflow is shown in Figure 1.

Two public transcriptome datasets were included. GSE20881 served as the training dataset, and GSE102133 was used for external validation. The analysis moved from data preprocessing and group annotation to differential expression analysis, immune-related gene screening, enrichment analysis, machine-learning gene selection, and validation of the candidate panel.

The dataset information is listed in Table 1.

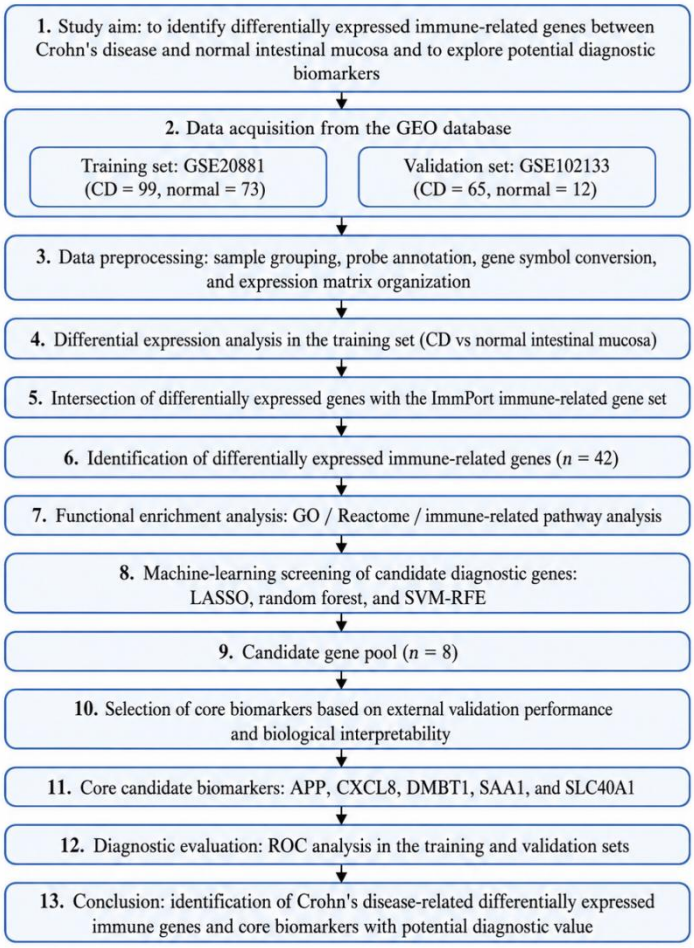


**Figure 1. Overall study workflow.**

**Table 1. Basic information of the included datasets.[a]**

| Dataset | Platform | Purpose | Sample source | Crohn's disease samples | Normal control samples | Main analytical use |
|---|---|---|---|---|---|---|
| GSE20881 | GPL1708 | Training set | Intestinal mucosal tissue | 99 | 73 | Differential expression analysis, identification of differentially expressed immune-related genes, and machine-learning screening |
| GSE102133 | GPL6244 | Validation set | Ileal mucosal tissue | 65 | 12 | External validation of candidate diagnostic biomarkers |

GSE20881 was used as the training set, and GSE102133 was used for independent cross-dataset evaluation.

GSE20881 included 172 intestinal mucosal samples, with 99 CD samples and 73 normal control samples. Because the samples came from different intestinal sites, biopsy site was considered in later analysis. After probe annotation and removal of invalid gene entries, 18 268 genes were retained. GSE102133 contained 77 ileal mucosal samples, including 65 CD samples and 12 normal control samples, and was used only for validation.

**Differential Expression Analysis in GSE20881**

The first analysis step compared transcriptome profiles between CD and normal intestinal mucosa in GSE20881. After adjustment for biopsy site, 75 genes met the criteria of FDR$<0.05$ and $|\text{logFC}| \geq 0.2$. Of these, 65 genes were upregulated and 10 genes were downregulated in CD samples.

The PCA plot after biopsy site adjustment showed partial separation between CD and normal mucosal samples, although overlap was still present (Figure 2A). According to intermediate analysis records, PCA without site correction displayed heavier intergroup overlap and weaker separation tendency, confirming that biopsy site confounding partially masked transcriptional differences between CD and normal mucosa. The volcano plot showed the distribution of differentially expressed genes after site adjustment (Figure 2B). The heatmap also showed different expression patterns between the 2 groups for the top differentially expressed genes (Figure 2C). Overall, the training dataset showed detectable mucosal transcriptional differences between CD and controls.

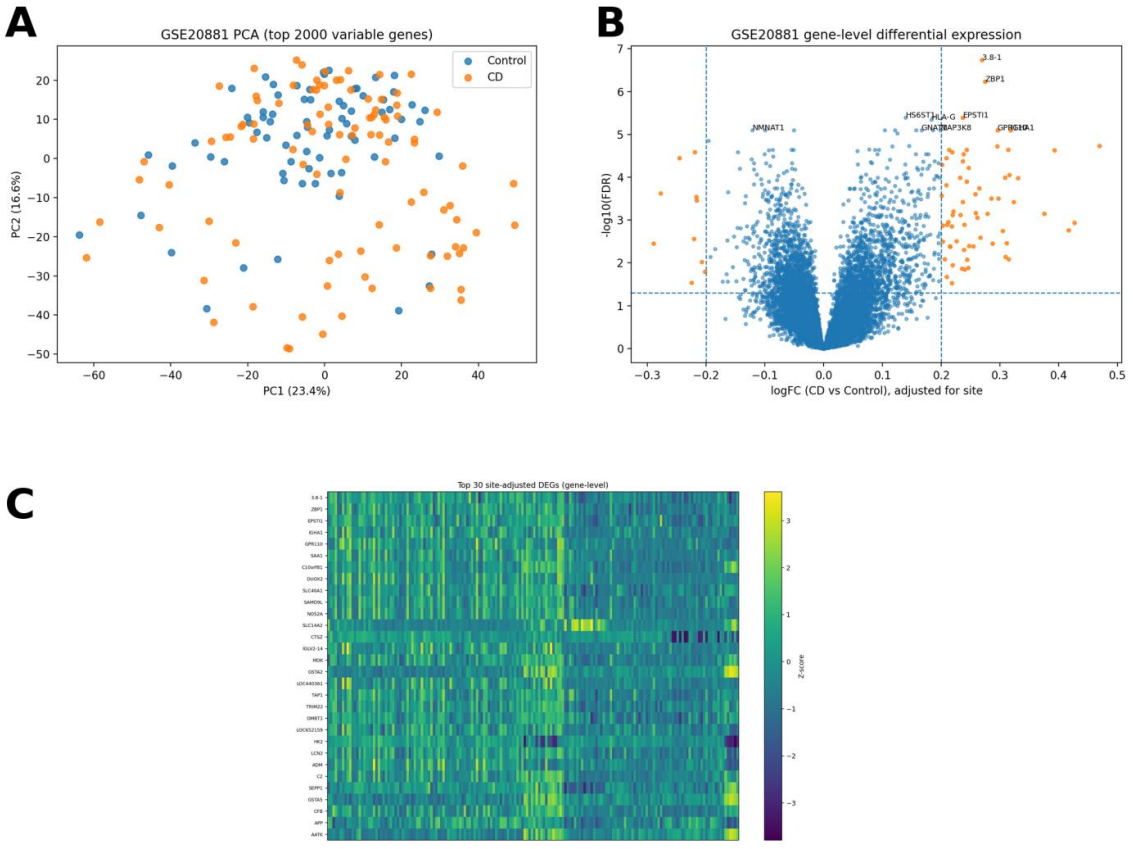


**Figure 2. Differential expression analysis between Crohn's disease and normal intestinal mucosa in GSE20881. A. Principal component analysis showing the overall distribution of Crohn's disease and normal intestinal mucosa samples. Each point represents 1 sample. B. Volcano plot showing the distribution of differentially expressed genes after adjustment for biopsy site. C. Heatmap showing the expression pattern of the top differentially expressed genes between Crohn's disease and normal intestinal mucosa.**

**Screening of Differentially Expressed Immune-Related Genes**

The 75 differentially expressed genes were then compared with the ImmPort immune-related gene set. This step produced 42 differentially expressed immune-related genes. Forty-one were upregulated in CD samples, and 1 was downregulated.

The Venn diagram shows the overlap between the differential expression results and the ImmPort gene set (Figure 3A). The heatmap showed that these 42 immune-related genes separated the 2 sample groups to a visible degree (Figure 3B). Their distribution is shown in the volcano plot (Figure 3C), and the direction of expression change is summarized in Figure 3D. Details of these genes, including logFC, FDR, direction, and representative immune-related terms, are given in Table 2.

**Functional Enrichment Analysis**

The 42 differentially expressed immune-related genes were used for functional enrichment analysis. In the GO biological process results, the main terms included antimicrobial humoral response, leukocyte chemotaxis, myeloid leukocyte migration, complement activation, and leukocyte migration involved in inflammatory response (Figure 4A). These terms mainly point to immune cell movement and mucosal inflammatory response.

Reactome enrichment analysis identified several immune and inflammatory pathways, including interleukin-10 signaling, interleukin-4 and interleukin-13 signaling, neutrophil degranulation, Toll-like receptor cascades, antimicrobial peptides, interferon gamma signaling, and interleukin-1 family signaling (Figure 4B). The enrichment pattern was consistent with the immune-related gene screening step and supported the next stage of candidate gene selection.

**Machine-learning Screening of Candidate Diagnostic Genes**

The 42 differentially expressed immune-related genes were then entered into 3 feature-selection methods. LASSO regression retained 19 genes with non-zero coefficients (Figure 5A). Random forest ranked the genes by their contribution to sample classification (Figure 5B). SVM-RFE was used as a third method to narrow the candidate range (Figure 5C).

After these analyses were integrated, 8 genes were kept as the candidate pool: *APP*, *CTSZ*, *CXCL8*, *DMBT1*, *GSTA2*, *S100A9*, *SAA1*, and *SLC40A1* (Figure 5D). This pool was not considered final. The genes were further reviewed according to expression consistency across datasets, validation performance, and biological relevance to mucosal inflammation. Consistent with the 3 objective screening criteria defined in the Methods section, *APP*, *CXCL8*, *DMBT1*, *SAA1*, and *SLC40A1* were retained as core candidate biomarkers. External AUC values of the 8 candidate genes are shown in Figure 5E.

**Diagnostic Validation of the Core Candidate Biomarkers**

The 5 selected genes were evaluated in the training and evaluation datasets. In GSE20881, the 5-gene model had an AUC of 0.835 (95% CI, 0.772–0.886) for distinguishing CD samples from normal intestinal mucosa samples (Figure 6A). In GSE102133, the same model showed an AUC of 0.990 (95% CI, 0.941–1.000) (Figure 6B).

*APP*, *CXCL8*, *DMBT1*, *SAA1*, and *SLC40A1* were then compared between groups at the expression level. In GSE20881, the 5 genes generally showed higher expression in CD samples than in controls (Figure 6C). A similar tendency was seen in GSE102133 (Figure 6D). These results support their further evaluation as candidate biomarkers in public transcriptome data, while further clinical and experimental validation is still required

IN PRESS

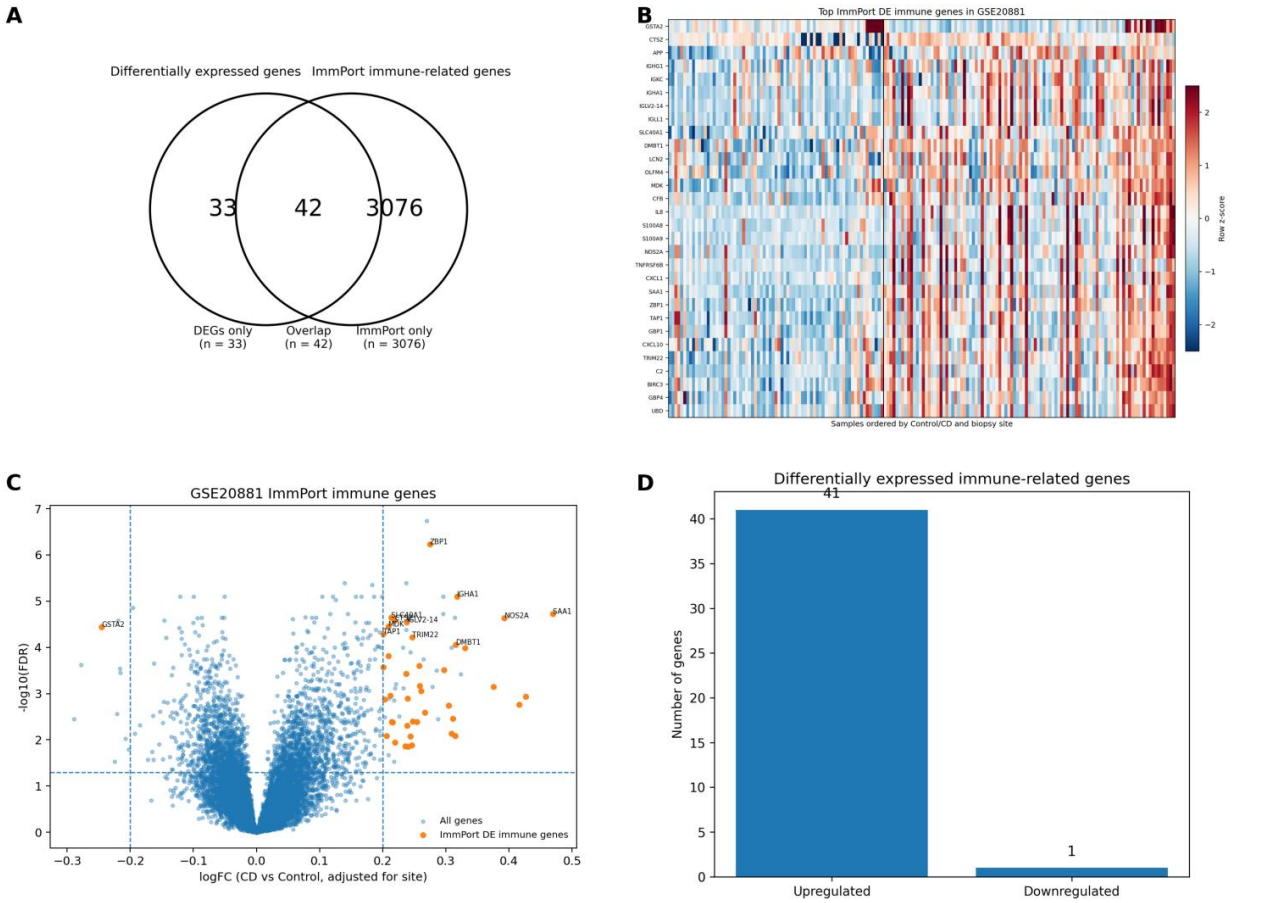


**Figure 3. Identification of differentially expressed immune-related genes in GSE20881. A. Venn diagram showing the overlap between differentially expressed genes identified in GSE20881 and immune-related genes obtained from the ImmPort database. B. Heatmap showing the expression pattern of the differentially expressed immune-related genes in Crohn's disease and normal intestinal mucosa samples. C. Volcano plot of the differentially expressed immune-related genes. D. Bar plot summarizing the numbers of upregulated and downregulated immune-related genes.**

**Table 2. Differentially expressed immune-related genes identified in GSE20881.**

| Gene symbol | logFC (CD vs control)[a] | FDR | Direction |
|---|---|---|---|
| *ZBP1* | 0.275 | $5.83 \times 10^{-7}$ | Up in CD |
| *IGHA1* | 0.318 | $7.97 \times 10^{-6}$ | Up in CD |
| *SAA1* | 0.470 | $1.86 \times 10^{-5}$ | Up in CD |
| *SLC40A1* | 0.214 | $2.28 \times 10^{-5}$ | Up in CD |
| *NOS2A* | 0.393 | $2.33 \times 10^{-5}$ | Up in CD |
| *CTSZ* | 0.219 | $2.60 \times 10^{-5}$ | Up in CD |
| *IGLV2-14* | 0.238 | $2.89 \times 10^{-5}$ | Up in CD |
| *MDK* | 0.209 | $3.53 \times 10^{-5}$ | Up in CD |
| *GSTA2* | −0.246 | $3.58 \times 10^{-5}$ | Down in CD |
| *TAP1* | 0.201 | $5.12 \times 10^{-5}$ | Up in CD |
| *TRIM22* | 0.247 | $6.03 \times 10^{-5}$ | Up in CD |
| *DMBT1* | 0.316 | $8.84 \times 10^{-5}$ | Up in CD |
| *LCN2* | 0.331 | $1.03 \times 10^{-4}$ | Up in CD |
| *C2* | 0.209 | $1.52 \times 10^{-4}$ | Up in CD |
| *CFB* | 0.258 | $2.49 \times 10^{-4}$ | Up in CD |
| *APP* | 0.201 | $2.68 \times 10^{-4}$ | Up in CD |
| *IGLL1* | 0.297 | $3.11 \times 10^{-4}$ | Up in CD |
| *GBP1* | 0.237 | $3.75 \times 10^{-4}$ | Up in CD |
| *GBP4* | 0.259 | $6.79 \times 10^{-4}$ | Up in CD |
| *CXCL8* | 0.376 | $7.17 \times 10^{-4}$ | Up in CD |
| *IGKC* | 0.261 | $8.73 \times 10^{-4}$ | Up in CD |
| *TNFRSF6B* | 0.212 | $1.11 \times 10^{-3}$ | Up in CD |
| *S100A8* | 0.427 | $1.15 \times 10^{-3}$ | Up in CD |
| *CXCL10* | 0.240 | $1.29 \times 10^{-3}$ | Up in CD |
| *BIRC3* | 0.203 | $1.33 \times 10^{-3}$ | Up in CD |
| *CXCL1* | 0.417 | $1.73 \times 10^{-3}$ | Up in CD |
| *OLFM4* | 0.305 | $1.82 \times 10^{-3}$ | Up in CD |
| *IGHG1* | 0.267 | $2.58 \times 10^{-3}$ | Up in CD |
| *UBD* | 0.312 | $3.44 \times 10^{-3}$ | Up in CD |
| *S100A9* | 0.248 | $4.04 \times 10^{-3}$ | Up in CD |
| *CCL20* | 0.215 | $4.07 \times 10^{-3}$ | Up in CD |
| *NCF2* | 0.255 | $4.12 \times 10^{-3}$ | Up in CD |
| *CXCL11* | 0.216 | $4.15 \times 10^{-3}$ | Up in CD |
| *CXCL2* | 0.239 | $4.91 \times 10^{-3}$ | Up in CD |
| *INDO* | 0.309 | $7.26 \times 10^{-3}$ | Up in CD |
| *IL1B* | 0.206 | $8.14 \times 10^{-3}$ | Up in CD |
| *MMP3* | 0.315 | $8.26 \times 10^{-3}$ | Up in CD |
| *LYZ* | 0.244 | $8.35 \times 10^{-3}$ | Up in CD |
| *TNIP3* | 0.219 | $1.13 \times 10^{-2}$ | Up in CD |
| *S100P* | 0.246 | $1.31 \times 10^{-2}$ | Up in CD |
| *CXCL3* | 0.236 | $1.38 \times 10^{-2}$ | Up in CD |
| *ITGB2* | 0.240 | $1.41 \times 10^{-2}$ | Up in CD |

[a]logFC indicates expression change in Crohn's disease relative to normal intestinal mucosa.

[b]CD: Crohn's disease; FDR: false discovery rate; logFC: log fold change.

IN PRESS



Published by Tehran University of Medical Sciences (http://ijaai.tums.ac.ir)

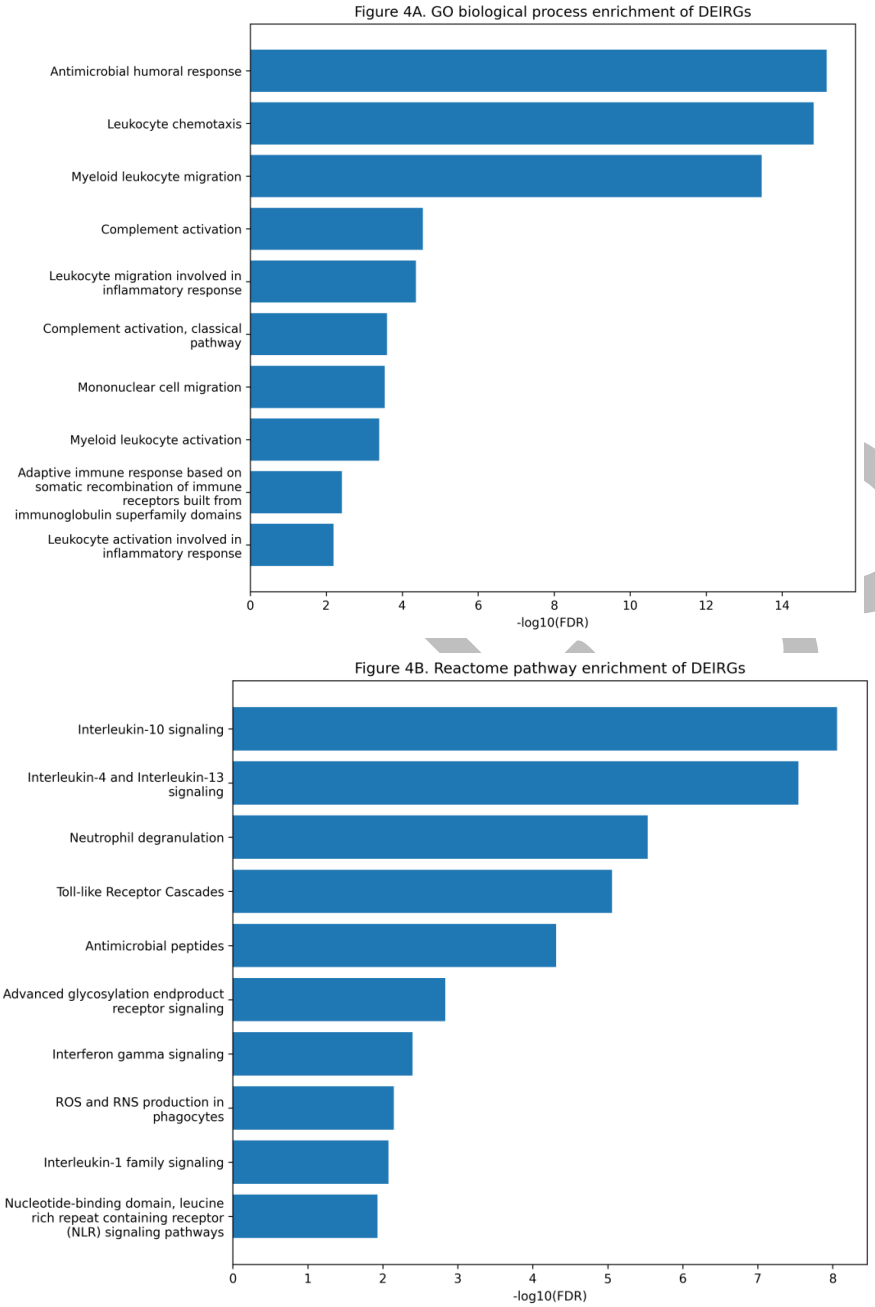


**Figure 4. Functional enrichment analysis of differentially expressed immune-related genes. A. GO biological process enrichment analysis of the differentially expressed immune-related genes identified in GSE20881. B. Reactome pathway enrichment analysis of the differentially expressed immune-related genes.**

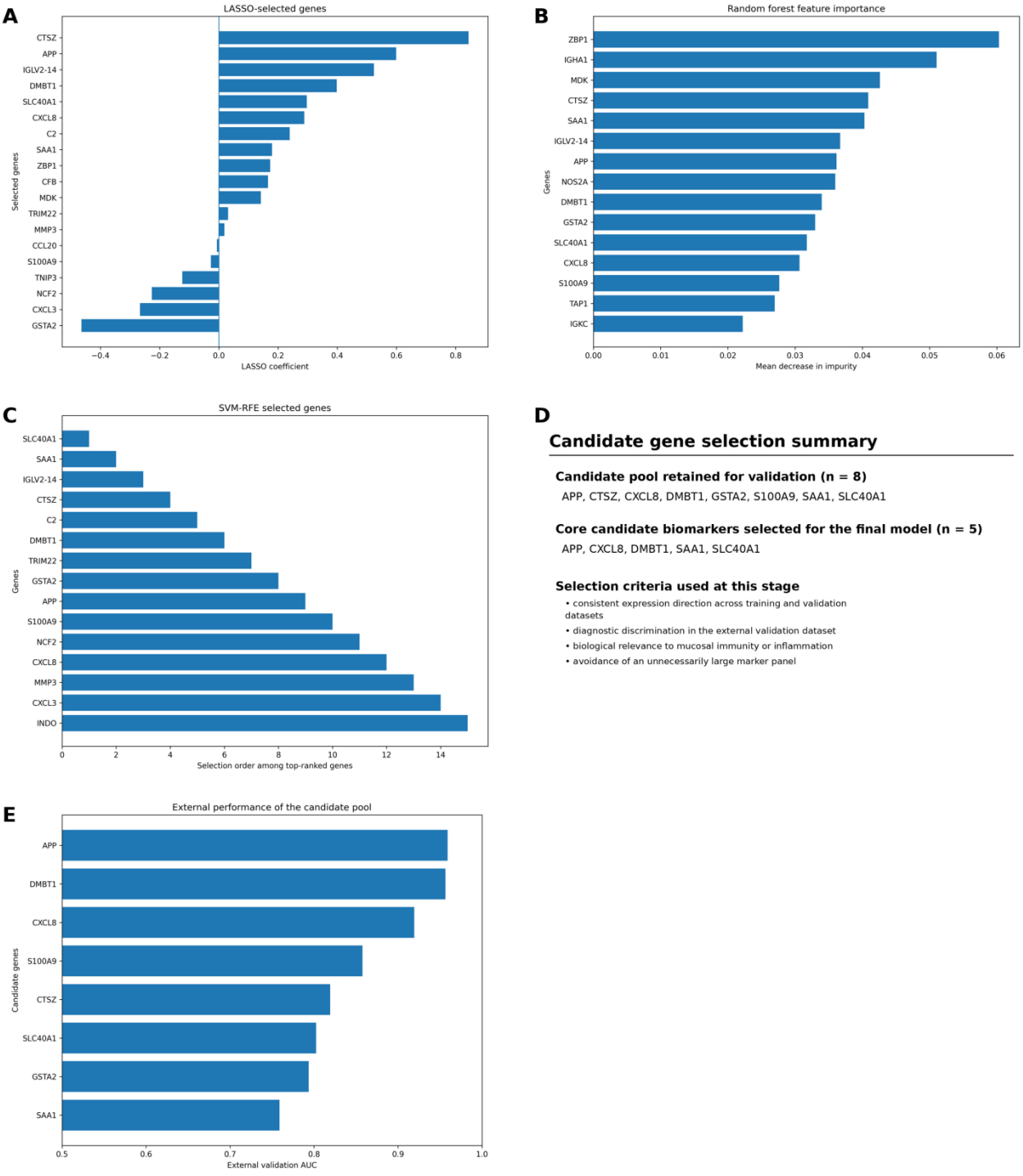


**Figure 5. Machine-learning screening of candidate diagnostic genes. A. Genes selected by the LASSO model. B. Genes ranked by random forest feature importance. C. Genes retained by SVM-RFE. D. Summary of the candidate gene pool and the final core candidate biomarkers. E. External validation AUC values of the 8 candidate genes.**

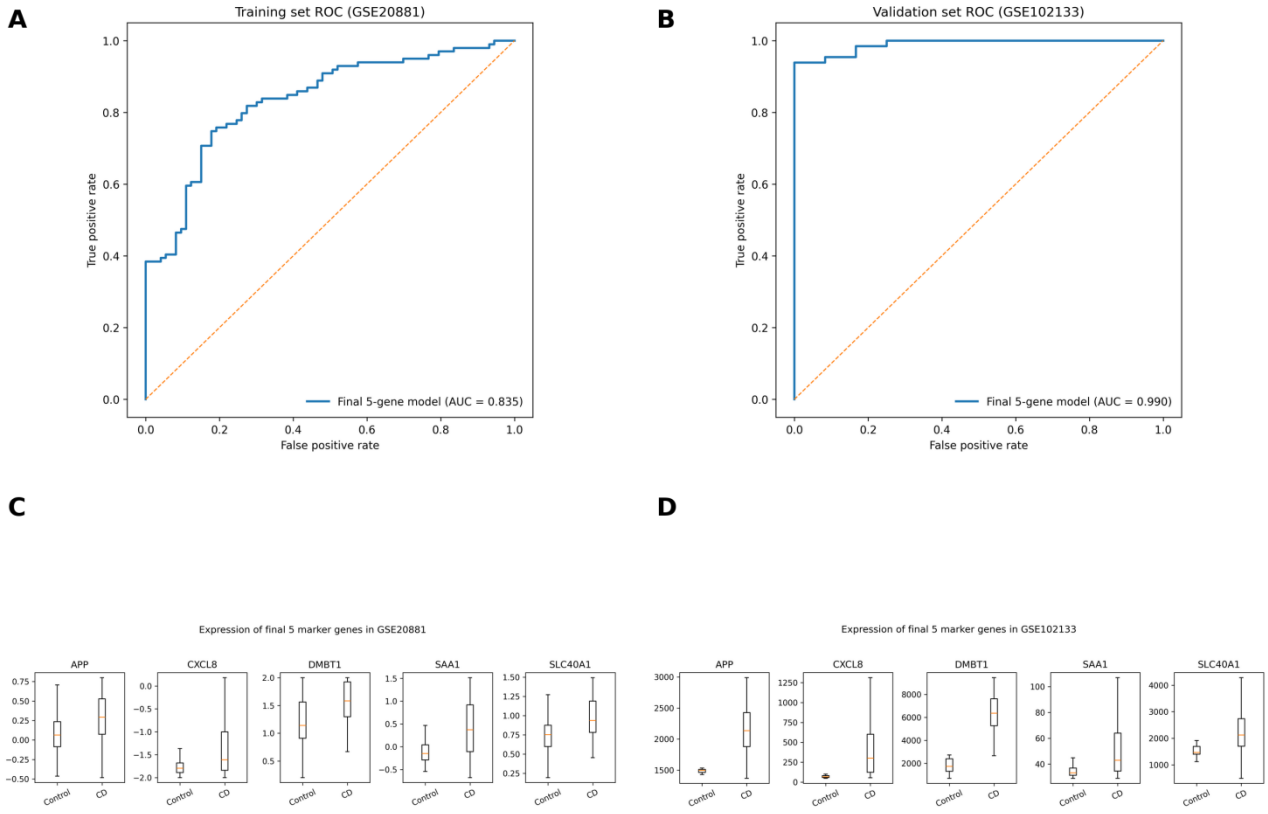


**Figure 6. Diagnostic validation of the 5 core candidate biomarkers. A. ROC curve of the 5-gene model in the GSE20881 training set. B. ROC curve of the 5-gene model in the GSE102133 validation set. C. Expression levels of *APP*, *CXCL8*, *DMBT1*, *SAA1*, and *SLC40A1* in GSE20881. D. Expression levels of the same 5 genes in GSE102133.**

## DISCUSSION

This study used public intestinal mucosal transcriptome datasets to examine immune-related expression differences between CD and normal mucosa. In the training dataset, 75 differentially expressed genes were identified, and 42 of them overlapped with the ImmPort immune-related gene set. Enrichment analysis linked these genes mainly to antimicrobial humoral response, leukocyte chemotaxis, myeloid leukocyte migration, complement activation, neutrophil degranulation, Toll-like receptor cascades, and interleukin-related pathways. After feature selection and cross-dataset evaluation, APP, CXCL8, DMBT1, SAA1, and SLC40A1 were retained as core candidate biomarkers. The 5-gene panel showed discrimination in both the training and evaluation datasets.

The stepwise selection process is important for interpreting the results. The 5 genes were obtained only after the differential-expression, immune-gene intersection, and feature-selection steps. It began with a fixed question: whether immune-related transcriptional differences could be identified between CD and normal intestinal mucosa. The later enrichment results were also consistent with this analytical framework. Leukocyte migration, neutrophil-related response, antimicrobial defense, and cytokine signaling are all plausible processes in inflamed intestinal mucosa. These findings support the biological relevance of the screened gene set, but they do not prove that the genes drive disease development.

Collectively, the 5-gene signature reflects a coordinated immune-related transcriptional pattern involving intestinal epithelial defense, innate immune cell recruitment, acute inflammatory response, and inflammatory iron homeostasis in CD mucosa: DMBT1 maintains epithelial antimicrobial barrier as the frontline mucosal protective factor; CXCL8 correlates with neutrophil chemotaxis and local innate immune cell infiltration after epithelial injury; SAA1 amplifies acute-phase inflammatory signaling downstream of leukocyte activation; APP modulates epithelial cytokine secretion

and epithelial cell phenotype to sustain inflammatory microenvironment; SLC40A1 balances macrophage iron export disturbed by persistent mucosal inflammation, forming a feedback loop between iron metabolism and inflammatory cascade. The coordinated upregulation of all 5 genes jointly reflects the comprehensive activation of epithelial-immune crosstalk in inflamed CD intestinal tissue, rather than isolated single-gene dysfunction.

Among the final 5 genes, *CXCL8* has the most direct connection with intestinal inflammation. *CXCL8*, also called IL-8, is a chemokine involved in neutrophil recruitment. Earlier work reported increased IL-8 in inflamed mucosa from patients with IBD, and the level was related to neutrophil infiltration and local inflammation.[19] In the present analysis, *CXCL8* was upregulated in CD mucosa and was retained in the final panel. This result is reasonable, but not specific to CD. *CXCL8* should be interpreted as part of an inflammatory expression pattern, rather than as a disease-specific marker on its own.

*DMBT1* also has a reasonable mucosal background. Previous work suggested that *DMBT1* is expressed in intestinal epithelial surface cells and Paneth cells, and that its expression increases in inflamed IBD tissues. A deletion variant of *DMBT1* was associated with CD risk, and Dmbt1-deficient mice were more susceptible to experimental colitis.[20] Its inclusion in the 5-gene panel may reflect epithelial defense and mucosal protection. This point is useful because CD mucosal injury involves not only immune cells, but also the epithelial response at the intestinal surface.

*SAA1* should also be interpreted with restraint. Serum amyloid A is an acute-phase protein and has been studied as a marker of inflammatory activity in CD. One study reported that SAA was related to mucosal and histologic inflammation and might help in some patients with active disease but without elevated CRP.[21] In our analysis, *SAA1* was upregulated in CD mucosa and remained after screening. This may reflect inflammatory activity rather than a CD-specific mechanism. It is therefore more appropriate to regard *SAA1* as 1 component of a combined inflammatory and immune-related signature.

*APP* was retained in the model, although its link to CD inflammation is less direct than that of *CXCL8*, *DMBT1*, or *SAA1*. Beyond its well-known neural functions, *APP* is expressed in human colonic epithelial cells; in vitro evidence indicates that it modulates epithelial phenotype by promoting pro-inflammatory cytokine secretion and altering intestinal lipid metabolism.[22] Following mucosal injury, upregulated epithelial *APP* may contribute to phenotypic transition and amplify local inflammatory signaling. Although animal model validation for CD-specific inflammation is lacking, *APP* serves as a unique epithelial-derived biomarker, complementing the immune-cell-originated signals of *CXCL8* and *SAA1* with additional epithelial transcriptional information.

*SLC40A1* encodes ferroportin, an iron export protein. Iron metabolism is often disturbed in IBD, and inflammation can affect iron handling. A recent study reported that butyrate may promote macrophage iron export through ferroportin and reduce the anemia-inflammation circuit in IBD contexts.[23] In the present analysis, *SLC40A1* showed a consistent tendency across datasets and was retained in the final panel. This may point to a connection between mucosal inflammation and iron metabolism, although the current data cannot tell whether *SLC40A1* is a cause, a response, or a marker of cellular composition changes.

The 5-gene model showed an AUC of 0.835 in GSE20881 and 0.990 in GSE102133. These values indicate that the selected genes separated CD mucosa from normal mucosa in the public datasets used here. The high AUC in the evaluation dataset should be viewed carefully. GSE102133 contained fewer controls than CD samples and consisted of ileal mucosal tissues, while GSE20881 included several intestinal sites. Platform differences, cohort background, and disease status may also affect model performance. The validation result therefore supports the stability of the expression pattern, but it is not enough to claim clinical diagnostic applicability.

Several limitations should be noted. The work was based on public datasets, and no newly collected clinical samples were used. Although an external dataset was included, the findings were not verified by qPCR, immunohistochemistry, or prospective clinical testing. The training dataset contained samples from different intestinal sites; site was considered in the analysis, but residual tissue-site effects may remain. Public datasets also differ in patient background, medication exposure, disease activity, and sample processing. In addition, the data came from bulk transcriptome profiles, so expression changes may reflect both gene regulation and changes in the proportions of epithelial cells, neutrophils, macrophages, lymphocytes, and other cell types.

IN PRESS

The extremely high AUC = 0.990 observed in the GSE102133 validation cohort requires cautious interpretation and potential overfitting assessment. Three factorsmay have contributed to this high discrimination value: first, severe sample imbalance (65 CD vs 12 healthy controls) in the evaluation dataset may inflate classification performance metrics; second, all validation samples were restricted to ileal mucosa, while the training set contained multi-segment intestinal tissues, reducing transcriptional heterogeneity within validation group; third, the 5-gene signature mainly reflects robust pan-inflammatory transcriptional signals strongly enriched in inflamed ileal lesions. We adopted repeated 5-fold cross-validation during machine learning screening to mitigate overfitting, but independent cohorts with more balanced case-control numbers are still required to verify stable diagnostic efficacy.

In addition, only binary classification labels (CD vs healthy control) were available in public GEO datasets, without continuous clinical inflammatory outcome indicators, so model calibration analysis could not be performed alongside ROC evaluation; calibration curves will be incorporated in future prospective clinical cohort validation.

We attempted to identify an additional independent GEO mucosal transcriptome cohort for secondary independent cross-dataset evaluation, but no public dataset simultaneously met all inclusion criteria (paired CD and healthy intestinal mucosal samples, complete biopsy location metadata, compatible microarray detection platform). Merging multiple heterogeneous datasets would introduce uncorrectable batch effects across different chip platforms, so we did not integrate extra cohorts for validation.

All data in this study originated from bulk intestinal mucosal transcriptome profiling, which cannot distinguish 2 distinct sources of differential expression: intrinsic transcriptional upregulation within specific cell types, and shifts in tissue cellular composition (e.g., increased proportions of neutrophils, macrophages, and lymphocytes in inflamed CD lesions). The elevated expression of the 5 core biomarkers may partially result from immune cell infiltration rather than endogenous gene activation in epithelial cells. Due to the lack of public paired single-cell and bulk mucosal datasets for CD patients, we could not perform single-cell validation in the current analysis. Single-cell RNA sequencing would be required in future studies to assign the expression signals of *APP*, *CXCL8*, *DMBT1*, *SAA1*, and *SLC40A1* to their specific cell-of-origin, thereby eliminating confounding from tissue cellular composition shifts.

Lack of individual disease activity scores for all CD samples restricted subgroup analysis of biomarker expression across active and remission phases; the 5-gene panel's expression fluctuation with disease activity remains uncharacterized in current public datasets.

Detailed medication history (corticosteroids, anti-tumor necrosis factor [TNF] biologic agents, immunomodulators) was not uniformly recorded for all subjects in the 2 GEO datasets. Pharmacological intervention can significantly reshape intestinal mucosal immune transcriptome and suppress the expression of inflammatory genes including *CXCL8* and *SAA1*, which may introduce unmeasured confounding bias to the current signature performance.

Overall, the study identified differentially expressed immune-related genes between CD and normal intestinal mucosa and narrowed them to a 5-gene candidate panel. The results point to immune cell migration, antimicrobial response, epithelial defense, cytokine signaling, and iron-related inflammatory changes in the mucosal expression pattern of CD. These findings provide a basis for further work, but they should remain exploratory. Future studies should test the 5-gene panel in independent clinical cohorts, examine its relationship with disease activity and intestinal location, and determine whether it adds information beyond existing clinical and laboratory indicators.

## CONCLUSIONS

This study used public transcriptome datasets to analyze immune-related expression differences between CD and normal intestinal mucosa. Differential expression analysis followed by intersection with the ImmPort gene set identified 42 differentially expressed immune-related genes. These genes were mainly enriched in immune and inflammatory processes, including leukocyte chemotaxis, antimicrobial response, complement activation, neutrophil degranulation, and cytokine-related pathways.

Through a stepwise screening strategy, APP, CXCL8, DMBT1, SAA1, and SLC40A1 were selected as core candidate biomarkers. The 5-gene model showed discrimination in both the training dataset and the independent cross-dataset evaluation dataset. These results suggest that the selected genes may reflect part of

IN PRESS

the mucosal immune changes in CD and may be useful for future biomarker research. The findings are still based on public transcriptome data, so independent clinical samples and experimental validation are needed before clinical application can be considered.

## STATEMENT OF ETHICS

This study was a secondary analysis of deidentified, publicly available transcriptome datasets (GSE20881 and GSE102133) and involved no new participant recruitment, intervention, or specimen collection.

## FUNDING

Not applicable.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest.

## ACKNOWLEDGMENTS

Not applicable.

## DATA AVAILABILITY

The datasets analyzed in this study are publicly available from the NCBI Gene Expression Omnibus under accession numbers GSE20881 and GSE102133. The analysis workflow and derived candidate-gene results are described in the article.

## AI ASSISTANCE DISCLOSURE

No artificial intelligence (AI) tools were used in the preparation, writing, editing, or analysis of this manuscript.

## REFERENCES

1. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, 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.
2. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. Lancet. 2017;389(10080):1741–55.
3. Roda G, Chien Ng S, Kotze PG, Argollo M, Panaccione R, Spinelli A, et al. Crohn's disease. Nat Rev Dis Primers. 2020;6(1):22.
4. Maaser C, Sturm A, Vavricka SR, Kucharzik T, Fiorino G, Annese V, et al. ECCO-ESGAR guideline for diagnostic assessment in IBD Part 1: initial diagnosis, monitoring of known IBD, detection of complications. J Crohns Colitis. 2019;13(2):144–64.
5. Mosli MH, Zou G, Garg SK, Feagan SG, MacDonald JK, Chande N, et al. C-reactive protein, fecal calprotectin, and stool lactoferrin for detection of endoscopic activity in symptomatic inflammatory bowel disease patients: a systematic review and meta-analysis. Am J Gastroenterol. 2015;110(6):802–19.
6. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature. 2011;474(7351):307–17.
7. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–24.
8. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49(2):256–61.
9. Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature. 2020;578(7796):527–539.
10. Neurath MF. Cytokines in inflammatory bowel disease. Nat Rev Immunol. 2014;14(5):329–42.
11. Ramanan D, Cadwell K. Intrinsic defense mechanisms of the intestinal epithelium. Cell Host Microbe. 2016;19(4):434–41.
12. Haberman Y, Tickle TL, Dexheimer PJ, Kim MO, Tang D, Karns R, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest. 2014;124(8):3617–33.
13. Kugathasan S, Denson LA, Walters TD, Kim MO, Marigorta UM, Schirmer M, et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. Lancet. 2017;389(10080):1710–8.
14. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41(Database issue).
15. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: disseminating data to the public



IN PRESS

for the future of immunology. Immunol Res. 2014;58(2-3):234–9.
16. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
17. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7:3.
18. Ding Y, Wilkins D. Improving the performance of SVM-RFE to select genes in microarray data. BMC Bioinformatics. 2006;7 Suppl 2:S12.
19. Mitsuyama K, Toyonaga A, Sasaki E, Watanabe K, Tateishi H, Nishiyama T, et al. IL-8 as an important chemoattractant for neutrophils in ulcerative colitis and Crohn's disease. Clin Exp Immunol. 1994;96(3):432–6.
20. Renner M, Bergmann G, Krebs I, End C, Lyer S, Hilberg F, et al. *DMBT1* confers mucosal protection in vivo and a deletion variant is associated with Crohn's disease. Gastroenterology. 2007;133(5):1499–509.
21. Yarur AJ, Quintero MA, Jain A, Czul F, Barkin JS, Abreu MT. Serum amyloid A as a surrogate marker for mucosal and histologic inflammation in patients with Crohn's disease. Inflamm Bowel Dis. 2017;23(1):158–64.
22. Puig KL, Manocha GD, Combs CK. Amyloid precursor protein mediated changes in intestinal epithelial phenotype in vitro. PLoS One. 2015;10(3): e0119534.
23. Xiao P, Cai X, Zhang Z, Li Y, Chen Y, Luo Y, et al. Butyrate prevents the pathogenic anemia-inflammation circuit by facilitating macrophage iron export. Adv Sci (Weinh). 2024;11(12):2306571.

IN PRESS