Yu et al. Cell Regeneration (2023) 12:8
https://doi.org/10.1186/s13619-022-00143-6

Cell Regeneration CSCB

RESEARCH ARTICLE Open Access

Development of a 32-gene signature using machine learning for accurate prediction of inflammatory bowel disease

Shicheng Yu[1,2], Mengxian Zhang[3], Zhaofeng Ye[4], Yalong Wang[1,2], Xu Wang[2] and Ye-Guang Chen[2,3,5*]

Abstract

Inflammatory bowel disease (IBD) is a chronic inflammatory condition caused by multiple genetic and environmental factors. Numerous genes are implicated in the etiology of IBD, but the diagnosis of IBD is challenging. Here, XGBoost, a machine learning prediction model, has been used to distinguish IBD from healthy cases following elaborative feature selection. Using combined unsupervised clustering analysis and the XGBoost feature selection method, we successfully identified a 32-gene signature that can predict IBD occurrence in new cohorts with 0.8651 accuracy. The signature shows enrichment in neutrophil extracellular trap formation and cytokine signaling in the immune system. The probability threshold of the XGBoost-based classification model can be adjusted to fit personalized lifestyle and health status. Therefore, this study reveals potential IBD-related biomarkers that facilitate an effective personalized diagnosis of IBD.

Keywords: IBD, XGBoost, Signature genes, AI prediction, Biomarker

Background

Inflammatory bowel disease (IBD), including ulcerative colitis (UC) and Crohn's disease (CD), is an idiopathic inflammatory bowel disease that heavily interferes with people's quality of life (Rutgeerts et al. 2009). Ulcerative colitis unfolds the continuous inflammation of the colonic mucosa and submucosa, which usually involves the rectum first and gradually spreads to the whole colon, while Crohn's disease can involve the whole digestive tract, which is a discontinuous full-thickness inflammation, most often involving the terminal ileum, colon and perianal area (Feagins et al. 2011, Ordas et al. 2012). IBD patients are usually accompanied by disrupted stem cell dynamics and impaired epithelial regeneration capacity (Krishnan et al. 2011, Olafsson et al. 2020). Chronic inflammation can proceed to irreversible tissue destruction unless appropriate therapy is provided (Hosseinkhani et al. 2020). While the first step to treat IBD is to ease the pain, reduce inflammation, and facilitate tissue repair and regeneration (De Vry et al. 2007). Clinical data combined with mouse colitis models reveals that many important signaling pathways promote mucosal regeneration (Bergstrom et al. 2016, Han et al. 2020, He et al. 2018). For example, protease-activated receptor 2 (PAR2) signaling can mediate colonic mucosal regeneration through the stabilization of YAP (He et al. 2018). Clinical trials with consideration of mucosal regeneration and immune modulation may have promising results in treating IBD (Pak et al. 2018).

Genetic, microbial, environmental, and immunoregulatory factors have been suggested to contribute to IBD, but the exact cause of which is still unknown (Graham and Xavier 2020, Kiesslich et al. 2012). Thus, the identification of cellular and molecular mechanisms that contribute to different subtypes and developing phases of IBD is essential for developing targeted therapies (Eftychi et al. 2019, Matsukawa et al. 2016). Differentially expressed genes and pathways have been identified between inflamed and

*Correspondence: ygchen@tsinghua.edu.cn
[2] Guangzhou Laboratory, Guangzhou 510700, China
Full list of author information is available at the end of the article

healthy control tissues of IBD patients, such as NF-κB, TNF-α, immune response, proinflammatory cytokines, and chemokines (Allen et al. 2012, Gadaleta et al. 2011). However, till now there are no ideal biological markers for IBD due to the complex genetic background and environmental factors (Khaki-Khatibi et al. 2020).

Many high-throughput experimental IBD data have been analyzed with various machine learning methods (Gubatan et al. 2021). By integrating 30 gene features and training with 310 samples (269 IBD patients and 41 healthy controls), an Artificial Neural Network and molecular prognostic score system-based classification model was built to achieve an Area Under Curve (AUC) above 0.950 (Li et al. 2020). Isakov et al.'s model achieved an accuracy [(true positive + true negative) / (positive + negative)] of 0.808 using 16,390 genes on 229 IBD patients and 90 healthy controls with Random Forest Algorithm (Isakov et al. 2017). Using a random forests-based classification model, Han et al. introduced a novel pathway-based approach to distinguish UC and CD and received the best AUC of 0.764 on the validation sets (Han et al. 2018). However, most IBD-related datasets are limited by the small sample size, high dimensions, and severe category imbalance, which bring great challenges to the integration of the transcriptomic data of IBD cohorts (Lloyd-Price et al. 2019, Pittayanon et al. 2020).

Machine learning has facilitated the diagnosis and risk prediction of IBD, but there was considerable variability in the performance of the different algorithms across the various cohorts (Gubatan et al. 2021). A desirable model should perform similarly even with new cohorts. EXtreme Gradient Boosting (XGBoost), a wildly used tree-based machine learning ensemble algorithm, is a gradient boosting-based software library for supervised classification (Chen and Guestrin 2016). This algorithm shows a balance between prediction performance and explainability, which indicates the ability of machine learning algorithms to explain or justify the results in terms that are understandable by humans (Al'Aref et al. 2020, Chen and Guestrin 2016). XGBoost is a common choice for dealing with the classification problem of multiple diseases, such as Parkinson's disease (Gao et al. 2018), colon cancer (Koppad et al. 2022), and breast cancer (Thalor et al. 2022). Due to its simplicity, interpretability, and ability to handle imbalanced datasets, we chose the XGBoost algorithm to construct our IBD classifier (Shorthouse et al. 2018). This study presents a diagnostic model to analyze multiple IBD cohorts. With the help of the Uniform Manifold Approximation and Projection (UMAP) and the XGBoost algorithm to select features, a 32-gene IBD signature was identified, which showed enrichment in neutrophil extracellular trap formation and cytokine signaling in the immune system. In comparison with the 54-gene-based model, 30-gene-based model, 21-gene-based model, Path 1-2-3-based model, and the Top SHAP value gene-based model, our 32-gene-based model showed better performance in the new cohort/samples of IBD patients.

Results

Identification of a 32-gene signature associated with IBD

To search for new potential IBD biomarkers, we combined unsupervised clustering analysis and the XGBoost feature selection method to exploit the gene signature from multiple cohorts and reduce the effect of a single cohort for better detecting positive samples from multiple IBD cohorts. Specifically, we took advantage of the integration pipeline of Seurat, an R toolkit for single-cell genomics (Hao et al. 2021), for preliminary feature selection with unsupervised clustering analysis. Four IBD-associated datasets, including GSE112366, GSE3365, GSE75214, and the data from the Integrative Human Microbiome Project (iHMP) were integrated, and 41,307 features across 705 samples were chosen (Fig. 1). Finally, 9 clusters were obtained using unsupervised clustering, and most cohorts are distributed uniformly in distinct clusters (Fig. 2A-C). We conjectured distinct clusters that might represent specific disease states and the marker genes that can thus be used to distinguish IBD from healthy controls. One hundred sixty-nine marker genes were filtered out using Seurat's FindMarker function (Supplementary Table 1).

To single out our IBD signature, XGBoost was used for intensive feature selection. 251 samples of iHMP, 127 samples of GSE3365, and 167 samples of GSE112366 were used for model construction. From them, 22, 15, and 29 genes were identified with a SHapley Additive exPlanations (SHAP) value above 0.05 (Fig. 2D-F), respectively. Using intersection analysis, we screened out six genes (AQP9, CXCL1, MMP3, MUC1, APOL1, and MTATP6P1) (Supplementary Fig. 1) overlapping with the 169 marker genes (Supplementary Table 1). A higher SHAP value would imply higher feature importance. We added 6 and 3 selected genes with the SHAP value above 0.2 in the GSE3365 and GSE112366-based XGBoost feature selection. As iHMP is the most comprehensive public IBD transcriptomic data (Lloyd-Price et al. 2019), a higher weight was given to the selected genes obtained from iHMP. We added 22 genes with the SHAP value above 0.05 in the iHMP-based XGBoost feature selection. Finally, we obtained a 32-gene set, and these genes are referred to as the IBD signature (Supplementary Table 2). Furthermore, the patients were clustered into several clusters based on the 32 features, and the control, UC and CD patients could be separated in GSE3365 and GSE75214 (Supplementary Fig. 2 A, B).

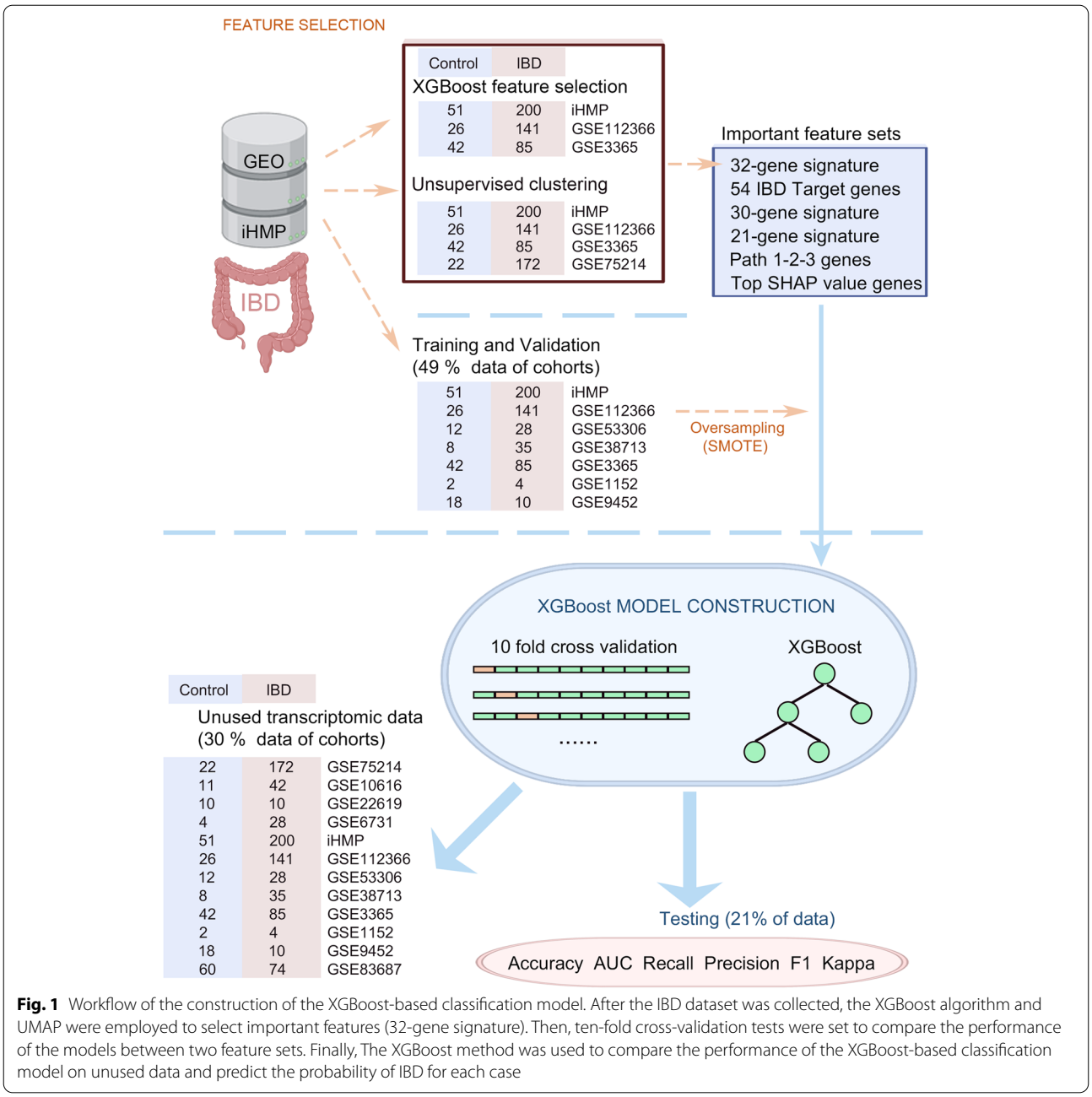

Fig. 1 Workflow of the construction of the XGBoost-based classification model. After the IBD dataset was collected, the XGBoost algorithm and UMAP were employed to select important features (32-gene signature). Then, ten-fold cross-validation tests were set to compare the performance of the models between two feature sets. Finally, The XGBoost method was used to compare the performance of the XGBoost-based classification model on unused data and predict the probability of IBD for each case

The 32-gene signature is mainly associated with immune response

To confirm the significance of the 32-gene signature in IBD prediction, a gene expression heatmap of these genes was performed in all four cohorts and seven additional IBD cohorts (GSE53306, GSE38713, GSE1152, GSE9452, GSE10616, GSE22619, and GSE6731) (Fig. 3A-K). In the iHMP cohort, the most abundant genes with a similar expression pattern were APOL1, AQP9, CCL24, COL4A1, CXCL1, CXCR1, FCGR3B, IFITM3, and MMP3. For example, the upregulation of AQP9 was identified in all the datasets. Further, we observed similar expression patterns in other cohorts. Across all data sets, we removed genes with missing expression levels and then normalized gene expression. Using Metascape, we analyzed the function of each gene and performed Gene Ontology (GO), Reactome Gene Sets, and KEGG Pathway enrichment analyses (Fig. 3L). As expected, the GO enrichment analysis revealed the enrichment of neutrophil extracellular trap formation (hsa04613). We also

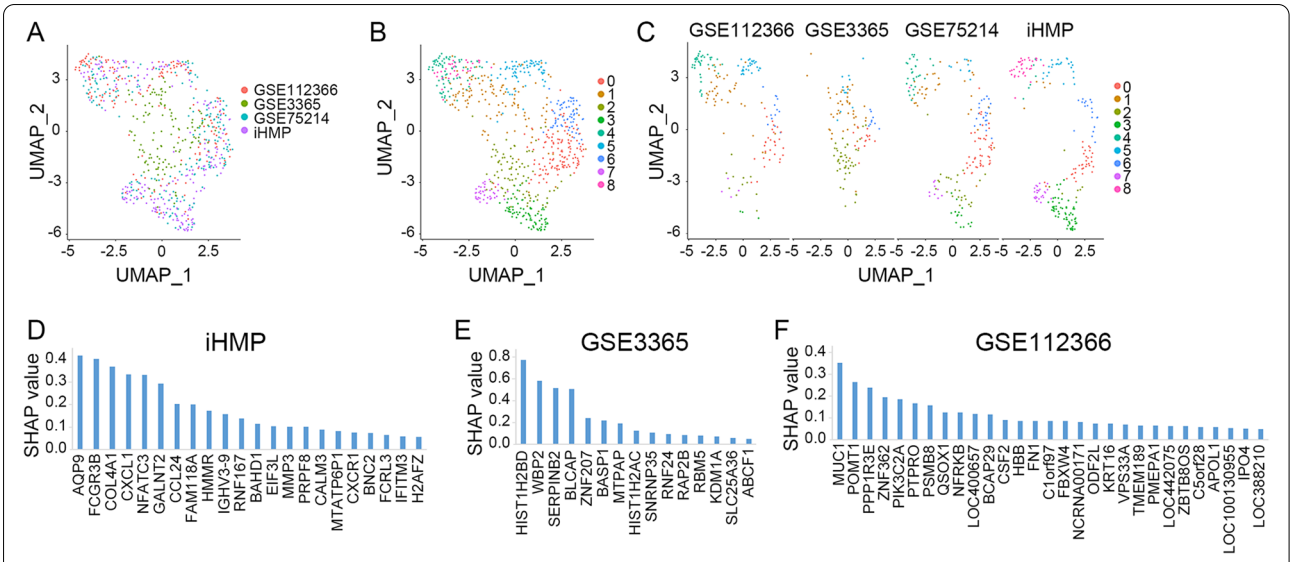

Fig. 2 Identification of the IBD-related 32-gene signature. **A** UMAP visualization of the integrated samples (GSE112366, GSE3365, GSE75214, and iHMP), color coded by samples from which cohort. **B** UMAP showing 9 clusters, color coded with cluster identities. **C** UMAP of GSE112366, GSE3365, GSE75214, and iHMP, color coded with cluster identities. SHAP values (feature importance scores) were calculated with XGBoost feature selection for iHMP (**D**), GSE3365 (**E**), and GSE112366 (**F**), respectively. X-axis represents gene names, Y-axis represents SHAP values

discovered the enrichment of cytokine signaling in the immune system (R-HSA-1280215) in the 32-gene set. For this analysis, AQP9, FCGR3B, H2AZ1, and H2BC5 were clustered into neutrophil extracellular trap formation, while CXCL1, MMP3, MUC1, SERPINB2, and IFITM3 were clustered into cytokine signaling in the immune system.

The 32-gene signature gives a more accurate IBD classification

Based on the 32-gene signature, an XGBoost-based classification model was built. The samples in the iHMP, GSE112366, GSE38713, GSE3365, GSE1152, and GSE9452 datasets were taken together for model training and testing. The dataset is split into training, validation, and testing sets in the ratio of 34.3%: 14.7%: 21%. Among the 961 samples included in the study, 462 samples were used for model training, validation, and testing, 288 samples (30% samples of all datasets) were used in the second testing step, and the other samples were not used in this study. Then, the feature importance score of each gene was calculated with SHAP on the 32-gene-based model (Fig. 4A).

We compared Accuracy, AUC, Recall, Precision, F1, and Kappa between the XGBoost-based classification models constructed using the 32-gene signature, the 54 FDA-approved and failed target genes (Sahoo et al. 2021) (Supplementary Table 3), the 30 genes from Li et al. (Li et al. 2020), the 21 genes from Yuan et al. (Yuan et al. 2017), the Path 1-2-3 genes (Sahoo et al. 2021) and Top SHAP value genes with high SHAP value directly selected from XGBoost (top 16 genes of iHMP, top 8 genes of GSE112366, and top 8 genes of GSE3365). We found that the XGBoost-based classification model with the 32-gene signature obtained better performance than the XGBoost-based classification model with the 54 IBD target genes: the 32-gene signature yielded 0.8953 Accuracy, 0.9653 AUC, 0.9292 Recall, 0.8741 Precision, 0.8991 F1 and 0.7906 Kappa; the 54 IBD target genes gave 0.8804 Accuracy, 0.9521 AUC, 0.9246 Recall, 0.8551 Precision, 0.8863 F1 and 0.7607 Kappa; the 30-gene-based model produced 0.8657 Accuracy, 0.9390 AUC, 0.9243 Recall, 0.8292 Precision, 0.8734 F1 and 0.7315 Kappa; the 21-gene-based model generated 0.8533 Accuracy, 0.9294 AUC, 0.8654 Recall, 0.8555 Precision, 0.8557 F1 and 0.7065 Kappa; the Path 1-2-3-based model produced 0.8845 Accuracy, 0.9688 AUC, 0.9409 Recall, 0.8522 Precision, 0.8914 F1 and 0.7689 Kappa; and the Top SHAP value gene-based model generated 0.9038 Accuracy, 0.9636 AUC, 0.9333 Recall, 0.8858 Precision, 0.9073 F1 and 0.8075 Kappa (Fig. 4B). Therefore, our feature selection strategy could contribute to an improvement with considerable performance. However, the XGBoost algorithm achieved better performance on testing set than several other common algorithms, except for Random Forest algorithms (Supplementary Table 4).

The 32-gene-based model achieves a better prediction of IBD

The trained XGBoost-based classification model was applied to the unused transcriptomic data. First, we calculated the accuracy (ranging from 0.3500–0.9167) and the confusion matrix in the individual datasets separately using the 32-gene signature model (Supplementary Fig. 2C-N). We found that the accuracy of the GSE83687 was even lower for the control samples of this cohort that were harvested from normal noninflamed bowel from patients with colon cancer. For the rest of the analyses, the data from the unused part of the training cohorts (unused thirty percent data of iHMP, GSE112366, GSE38713, GSE3365, GSE1152, and GSE9452) and the remaining cohorts (GSE75214, GSE10616, GSE22619, and GSE6731) were combined for the following study. We compared the predictive accuracy between the XGBoost-based classification model constructed with the 32-gene signature, 54 IBD target genes, 30-gene signature, 21-gene signature, Path 1-2-3-gene signature, and Top SHAP value-gene signature, respectively. As shown in the confusion matrix, 32-gene-based model (0.8651 Accuracy) (Fig. 4C) obtained a higher performance than 54-gene-based model (0.6436 Accuracy) (Fig. 4D), 30-gene-based model (0.7958 Accuracy) (Fig. 4E), 21-gene-based model (0.7197 Accuracy) (Fig. 4F), Path 1-2-3-based model (0.8062 Accuracy) (Fig. 4G), and the Top SHAP value gene-based model (0.8235 Accuracy) (Fig. 4H). These results indicate that the 32-gene-based model performs better on the unused transcriptomic data than other models. The XGBoost-based classification model also performed better than the Random Forest-based classification model with a 32-gene signature on unused transcriptomic data (Fig. 4C, I). Taken together, the XGBoost algorithm works better than other common algorithms in IBD prediction.

Once trained, the XGBoost-based classification model could be applied to new data. We calculated the IBD scores for all 288 testing samples with the estimator.predict_proba function of PyCaret, and 8 IBD score pictures for each patient were shown in Fig. 5A. The abscissa represents the number of dots. In the plot, the predicted probability for each sample being positive for IBD was displayed with yellow dots. IBD scores (numbers of yellow dots) range from 0 to 100, with a higher score indicating a higher probability of developing IBD. A sample with more than 50 green dots was identified as a healthy control. A lower probability threshold was applied to higher-risk people. In the 32-gene-based model, when the probability threshold was reduced to 0.45, the correctly predicted IBD cases increased from 209 to 213. Subsequently, the accuracy of the 32-gene-based model slightly increased from 0.8651 to 0.8789 (Figs. 4C, 5B). We also observed that the accuracy (0.8512) of the 32-gene-based model was slightly decreased under the 0.55 probability threshold (Fig. 5C).

A 19-gene subtype classification model can distinguish between UC and CD

To distinguish UC from CD, we constructed an XGBoost subtype classification model. We obtained subtype signature genes from iHMP (74 UC vs 126 CD) and GSE3365 (26 UC vs 59 CD)-based XGBoost subtype feature selection. However, genes with SHAP values above 0.05 showed no overlap with the 169 marker genes. Then, six genes were selected with the GSE3365-based XGBoost subtype feature selection with SHAP values above 0.2. Even if only two cohorts were used for XGBoost subtype feature selection, a slightly higher weight was also given to the iHMP cohort, and 13 genes were achieved with the SHAP value above 0.1 in the iHMP-based XGBoost subtype feature selection. However, the heatmaps of four cohorts presenting the expression pattern of the 19 selected genes (ARRDC4, CCND2, CD4, CD59, ERI3, FKBP5, HLA-DQA2, HLA-H, IGHG1, IGKV2D-40, KDM8, KLF6, MT1M, PEMT, SH3YL1, SIGIRR, SLC37A2, SUPT4H1, and TSKU) exhibited no significant differences between UC and CD (Fig. 6A-D). Furthermore, the patients were clustered into several clusters based on the 19 features, and the control, UC and CD patients could be separated in GSE3365 and GSE75214 (Supplementary Fig. 3A, B). Due to the distinct expression pattern, the functional analysis of this gene set did not indicate to be clinically meaningful (Supplementary Table 5). Meanwhile, we found no intersections between the 19 gene set and the gene sets of 70 genes (Park et al. 2021) and 5 genes (Han et al. 2018) of previous subtype classification models. Still, 19 genes were used for the XGBoost subtype classification model building.

Based on the 19 selected genes, 54 IBD target genes, and 5 previously reported genes (Han et al. 2018), we constructed the subtype classification model with iHMP (74

(See figure on next page.)

Fig. 3 The 32-gene signature shows similar expression patterns in all cohorts and are mainly related to immune response. **A** iHMP, (**B**) GSE75214, (**C**) GSE3365, (**D**) GSE10616, (**E**) GSE6731 (**F**) GSE1152, (**G**) GSE112366, (**H**) GSE53306, (**I**) GSE38713, (**J**) GSE9452, and (**K**) GSE22619. Row Z-score gene expression heatmaps were generated using Log2(TPM + 1) values of the iHMP cohort and other cohorts' microarray expression profiles. **L** Gene-ontology analysis using metascape (http://metascape.org) of the 32-gene signature. The red dashed lines indicate upregulated genes

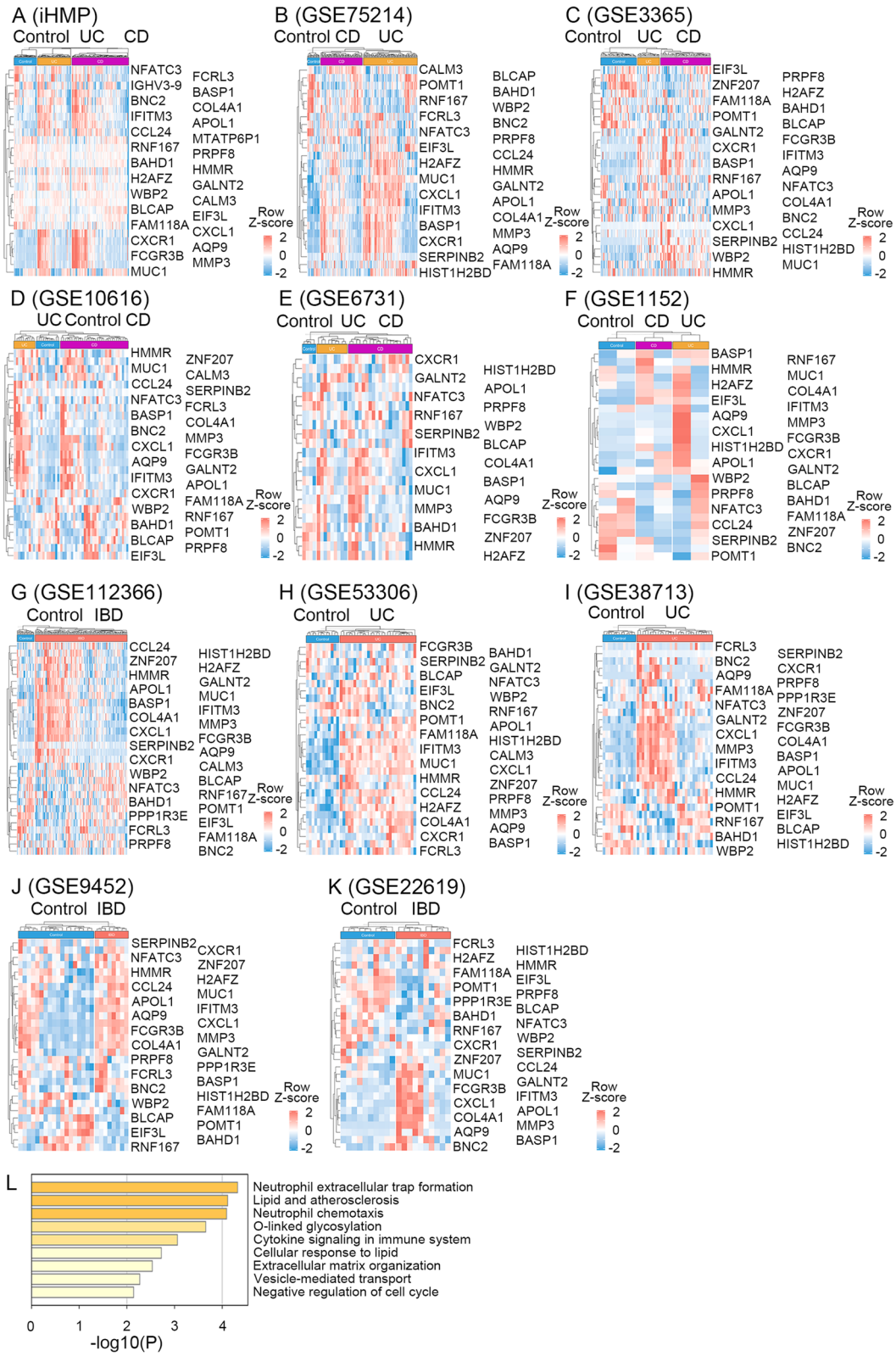

Fig. 3 (See legend on previous page.)

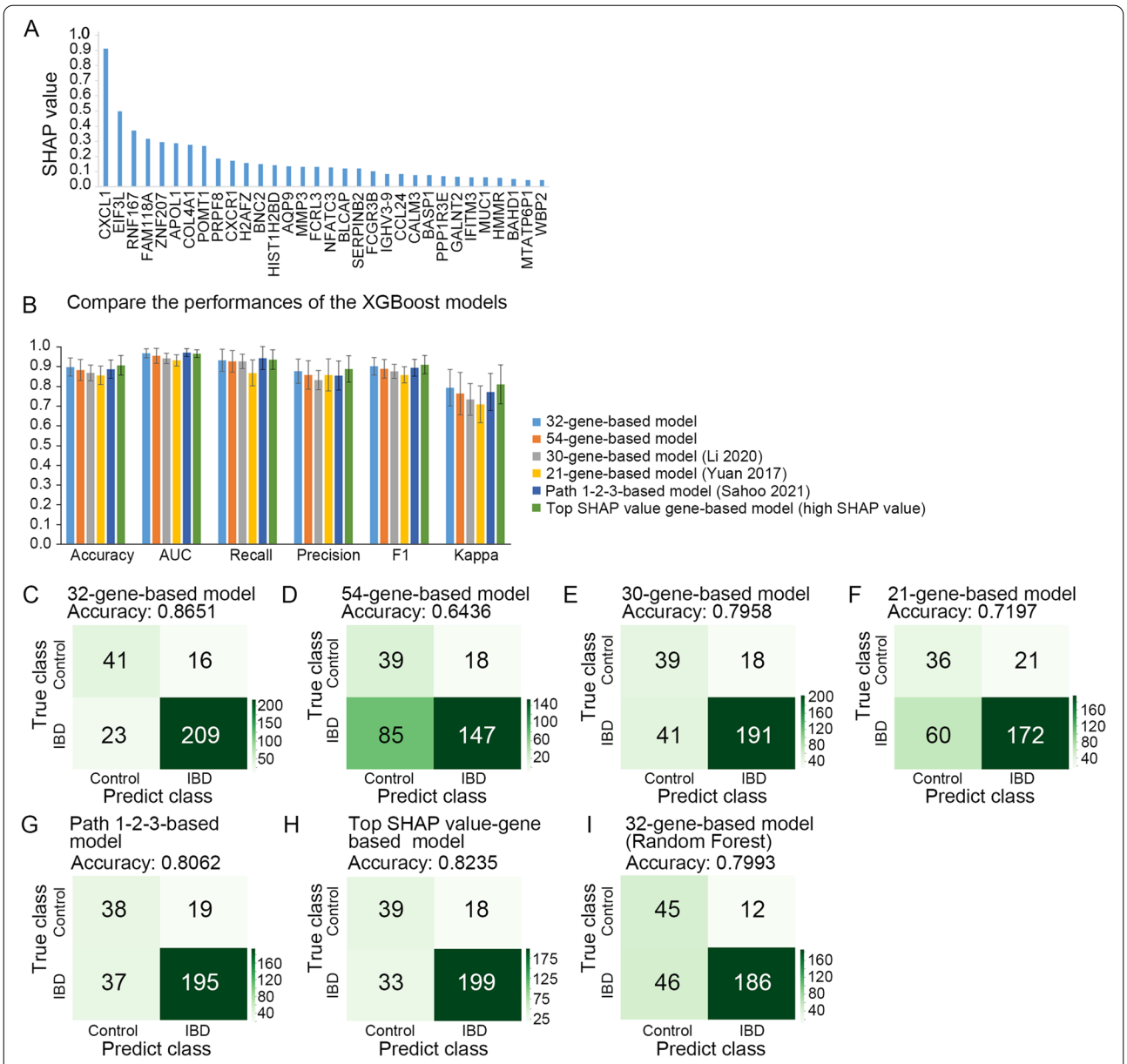

Fig. 4 The 32-gene-based model achieves a better performance than other models. **A** SHAP value (feature importance score) of the 32 genes. **B** The histogram comparison of the Accuracy, AUC, Recall, Precision, F1 and Kappa of two XGBoost-based classification models, those values range between 0 and 1 (0, poor performance; 1, good performance). **C**-**H** Confusion matrix of 32, 54, 30, 21, Path 1-2-3, and Top SHAP value gene-based XGBoost classification models with samples that were not used for training and validation. **I** Confusion matrix of the 32-gene-based Random Forest classification with samples that were not used for training and validation. Confusion matrix detailing the true positive (right lower), true negative (left upper), false positive (right upper), and false negative (left lower) predictions from XGBoost-based classification model. Accuracy = (true positive + true negative) / total

UC vs. 126 CD), GSE3365 (26 UC vs. 59 CD), GSE10616 (10UC vs. 32 CD), GSE6731(9 UC vs. 19 CD). In order to oversample the UC samples, the SMOTE algorithm was used. The 19-gene subtype classification model performed better than the other three models (the 32-gene, 54-gene, and 5-gene subtype classification model) on Accuracy, AUC, Recall, Precision, F1, and Kappa (Fig. 6E). Moreover, as shown in the confusion matrix, 19-gene subtype classification model (0.6395 Accuracy) gave a higher performance than the 32-gene subtype (0.5872 Accuracy), 54-gene subtype (0.5930 Accuracy), and 5-gene subtype classification model (0.4593 Accuracy) on unused transcriptomic data (Fig. 6F-I).

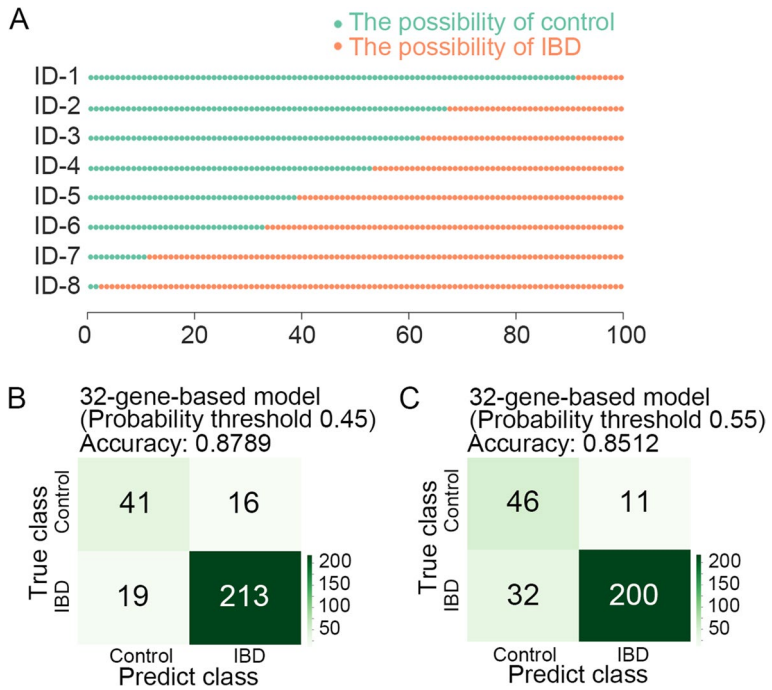

Fig. 5 The 32-gene-based model may predict well with a certain threshold. **A** The number of green dots indicates the possibility of health control, and the number of yellow dots indicates the possibility of IBD. The Y axis represents the ID of each sample, X axis represents the number of dots. **B**, **C** Confusion matrix for the 32-gene-based model with the probability threshold = 0.45 (**B**) and 0.55 (**C**)

Discussion

In this study, we successfully identified a 32-gene signature for IBD diagnosis using Seurat-based unsupervised clustering analysis and the XGBoost feature selection method. IBD is characterized by a dysregulated mucosal immune system (Punit et al. 2015), and the association between lipid mediators and cytokines has been extensively studied (Hamid and Tulic 2009). Among the 32 genes, the role of many of them in IBD has been previously demonstrated, such as AQP9 (Yu et al. 2021), BAHD1 (Zhu et al. 2015), BASP1 (Hong et al. 2018), BLCAP (Yuan et al. 2017), CALM3 (Park et al. 2015), CCL24 (Manousou et al. 2010), COL4A1 (Eshelman et al. 2020), CXCL1 (Cheng et al. 2019), CXCR1 (Ohtani et al. 2002), FAM118A (Khorasani et al. 2020), FCGR3B (Asano et al. 2013), FCRL3 (Martinez et al. 2007), GALNT2 (Nimmo et al. 2011), H2AFZ (alias: H2AZ1) (Chen et al. 2021), IFITM3 (Mo et al. 2013), IGHV3-9 (Yuan et al. 2021), MMP3 (Biancheri et al. 2015), MUC1 (Pothuraju et al. 2020), NFATC3 (Frigerio et al. 2021), SERPINB2 (Wei et al. 2015), and ZNF207 (Yuan et al. 2017). However, the role of APOL1, BNC2, EIF3L, HIST1H2BD (alias: H2BC5), HMMR, MTATP6P1, POMT1, PPP1R3E, PRPF8, RNF167, and WBP2 in IBD is unclear. Interestingly, PRPF8, a spliceosome component involved in the pre-mRNA splicing (Martinez-Gimeno et al. 2003), was correlated with neutrophil chemotaxis and cellular response to lipid (GO:0030593 and GO:0071396). WBP2, a transcriptional coactivator of estrogen receptor alpha and progesterone receptor (Lim et al. 2011), may play a role in neutrophil extracellular trap formation and cellular response to lipid (hsa04613 and GO:0071396). Those genes might be new potential markers for IBD. Our feature selection process can be used as a framework to identify potential biomarkers through comprehensive mining of public databases. However, biological experiments need to be performed to validate the function of these candidates in IBD.

We compared the performance of our XGBoost-based classification model among the 32-gene signature, the 54 IBD target genes, the 30-gene signature, and the 21-gene signature in terms of Accuracy, AUC, Recall, Precision, F1, and Kappa. The 32-gene-based model showed better performance than other models on test samples. An ideal model should have an accurate prediction performance in untrained cohorts, and our XGBoost-based classification model achieved a better accuracy (0.8651) in never trained cohorts/samples. Therefore, our model could achieve a robust prediction for the samples of multiple cohorts. However, our XGBoost subtype classification model with 19 selected genes still needs further improvement, although it gave a better prediction (0.6395 Accuracy) than other models.

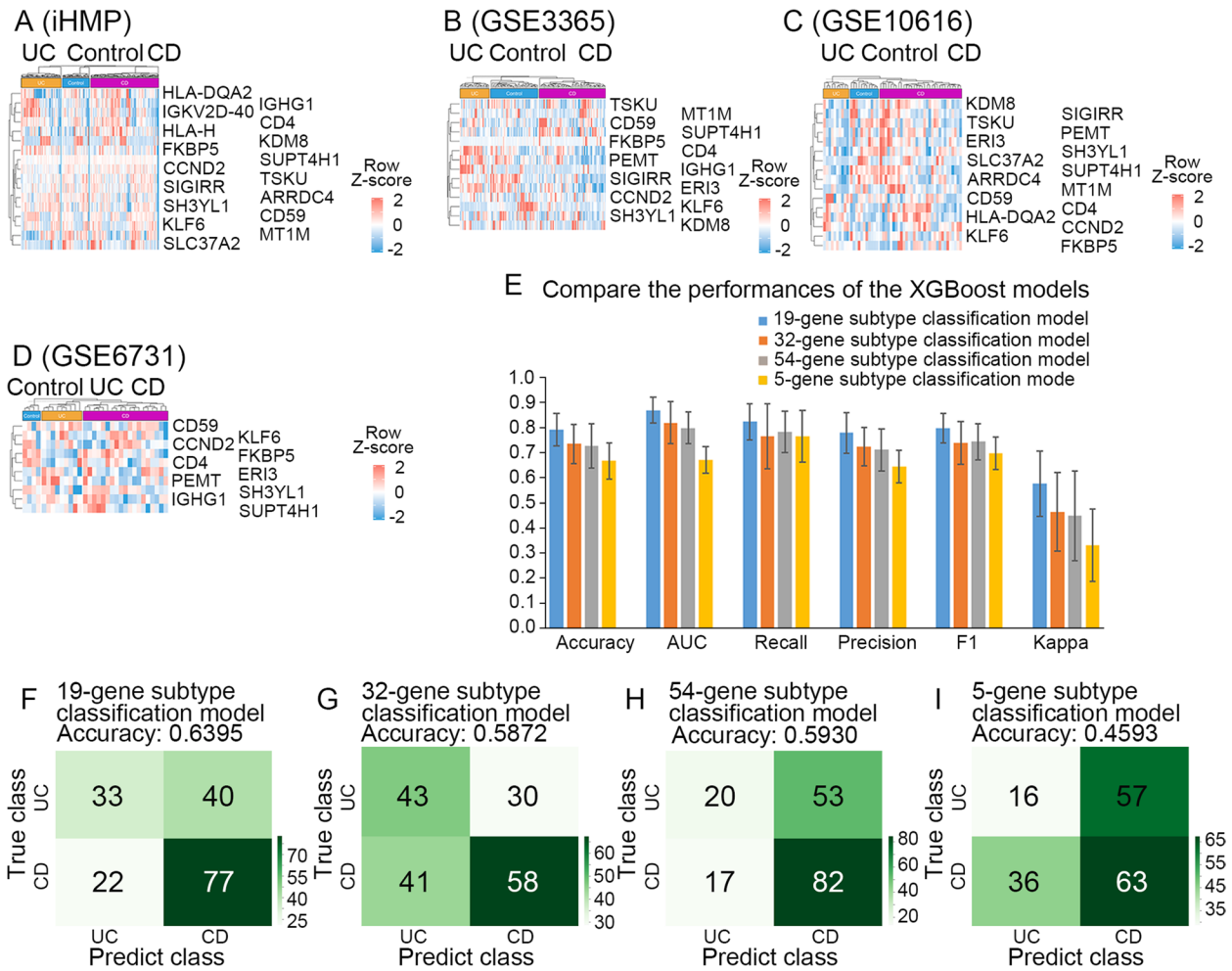

Fig. 6 The 19-gene subtype classification model could distinguish between UC and CD. **A-D** Heatmaps show the gene expression of the 19 selected genes in iHMP (**A**), GSE3365 (**B**), GSE10616 (**C**), and GSE6731 (**D**). **E** The histogram comparisons of the Accuracy, AUC, Recall, Precision, F1, and Kappa of two XGBoost-based classification models, those values range between 0 (poor performance) and 1 (good performance). **F-I** Confusion matrix for the 19-gene subtype classification model (**F**), the 32-gene subtype classification model (**G**), the 54-gene subtype classification model (**H**), the 5-gene subtype classification model (**I**)

Based on the XGBoost-based classification model, we calculate the IBD scores of each sample. These results can be used for personalized treatment for each patient. The high values of IBD scores suggest that more attention is needed for the individuals. The judgment may be inaccurate in patients with near 50 percent probability. It is also worth mentioning that the XGBoost-based classification model can be adjusted by changing the probability threshold. A lower probability threshold would increase the number of patients identified as IBD and decrease the number identified as health control. Individuals can adjust this probability threshold to fit their health status.

Conclusions

In this study, we use machine learning to develop a 32-gene signature for accurate prediction of IBD. We demonstrate a better performance of the 32-gene-based XGBoost model on transcriptomic data with multiple cohorts. Among the 32 genes, some have been reported to be associated with IBD development, but the others are new potential IBD biomarkers, such as APOL1, BNC2, EIF3L, HIST1H2BD, HMMR, MTATP6P1, POMT1, PPP1R3E, PRPF8, RNF167, and WBP2. We further show that adjusting the probability threshold can facilitate an effective personalized diagnosis of IBD.

Methods

Data sources and organization

The transcriptomic data were downloaded from the Gene Expression Omnibus (GEO) database (GEO: GSE112366, GSE53306, GSE38713, GSE3365, GSE1152, GSE9452, GSE75214, GSE10616, GSE22619, GSE6731, and GSE83687) and the iHMP (Lloyd-Price et al. 2019). GEO databases were downloaded from GEO with the limma R package and GEO query (Barrett et al. 2013). ComBat function of sva package was used to remove the batch effect for each cohort (Leek et al. 2012). A total of 846 patients (182 controls and 664 IBD) were included in the study. Among the cohorts, 70% of the data of iHMP, GSE112366, GSE38713, GSE3365, GSE1152, and GSE9452 served as the training & validation data to construct the machine learning model, and the remaining data of those cohorts and the data of GSE75214, GSE10616, GSE22619, and GSE6731 cohorts used as the test data to analyze the model's accuracy. In this training-validation set, 70% were in the training set, and 30% were in the validation set. iHMP, GSE3365, and GSE112366 were used for feature selection.

Unsupervised clustering with UMAP clustering

The different sources and the data sampling impact may affect the identification of significantly differential genes. To minimize the impacts of cohort differences on the classification model, we combined the bulk data and selected characteristic genes that were not affected by sampling. Specifically, the FindIntegrationAnchors package was obtained to integrate GSE112366, GSE3365, GSE75214, and iHMP. Due to the small sample size, we set twenty dims. UMAP was run with the R package Seurat (version 4.0) (Becht et al. 2018). Patients were clustered into several clusters. Finally, the FindAllMarkers function of Seurat (version 4.0) was used to identify the marker genes for each cluster used for AI model building (Butler et al. 2018). Among the clusters, marker genes with p_val_adj < 0.000000000000001 and avg_logFC > 0.5 were selected as significant genes.

Feature selection with XGBoost

Before the feature selection and the model construction, each input data of the patient was normalized using MinMax to yield values between 0 and 1. This study calculated feature importance scores and performed feature selection with XGBoost Extreme Gradient Boosting (XGBoost) on iHMP, GSE3365, and GSE112366, respectively (Chen and Guestrin 2016, Ogunleye and Wang 2020). We fed all detected genes of GSE3365 and GSE112366 to construct the XGBoost-based classification model, respectively. On the other hand, 5000 top variable genes of iHMP cohort were screened with var function of the R package and these genes were used to construct the XGBoost-based classification model to reduce the input dimension. Genes with an absolute SHAP value above 0.05 were selected in 3 cohorts. Then, genes with an absolute SHAP value above 0.2 of iHMP, genes with an absolute SHAP value above 0.1 of GSE3365, and genes with an absolute SHAP value above 0.1 of GSE112366 were selected. At last, an intersection was taken between these selected genes, and marker genes were identified in the Principal component analysis.

Feature importance, gene expression, and Gene Ontology analysis

In order to analyze feature importance, we used SHAP, a method for estimating instance-wise Shapley values that represent true estimates of the effects of each feature on a prediction (Lundberg et al. 2020). Log2(TPM + 1) transformation was performed to normalize TPM values of each cohort. The microarray expression profiles of other cohorts were obtained from the Gene Expression Omnibus (GEO) public microarray database. The R statistical package (version 4.0.3) was used to handle missing values, scale normalization, and median centering. The heatmaps were created using the ComplexHeatmap R package (https://github.com/jokergoo/ComplexHeatmap). The gene function annotation was conducted using Metascape software (https://metascape.org/gp/index.html#/main/step1) using default settings (Zhou et al. 2019).

XGBoost-based classification model construction and Evaluation of the classification model

Based on the likelihood of FDA approval and failure (Sahoo et al. 2021), we collected 54 target genes of IBD. We also collected gene sets that are important for diagnosing IBD in Li et al. study (Li et al. 2020) and Yuan, et al. study (Yuan et al. 2017). Then, 32-gene signature, 54 IBD target genes, 30-gene signature (Li et al. 2020), 21-gene signature (Yuan et al. 2017), Path 1-2-3-gene signature (Sahoo et al. 2021), and Top SHAP value-gene signature were fed into the Extreme Gradient Boosting algorithm XGBoost ('xgboost' package in python), respectively. We fed our AI model with transcriptomic data and tested the constructed AI model with unused transcriptomic data. In order to avoid data imbalance, SMOTE function of imblearn package was used. In cases of uneven distribution of classes, tenfold cross-validation was carried out to determine Accuracy, AUC, Recall, Precision, F1, and Kappa. We also calculated the accuracy in the individual datasets separately for the 32-gene signature model.

The probability threshold and the IBD possibility for each patient

By manually adjusting the probability threshold of the predict_model function of PyCaret (https://pycaret.org/) (Ali 2020), the prediction result was changed. Based on the prediction score obtained with the predict_model function of PyCaret, each sample was indicated for IBD possibility.

Statistical analysis

R (https://www.r-project.org/) and python (https://www.python.org/) were performed for statistical analysis of sequencing data. In XGBoost, "SHAP values" for each gene were calculated based on the SHAP package. The comparison was done using the student's t-test or Wilcoxon ranks test. The p_val_adj was calculated using the Bonferroni correction compared with all genes in the dataset (https://satijalab.org/seurat/reference/findmarkers). Seurat (version 4.0) was obtained for quantification and Statistical analysis (https://satijalab.org/seurat/).

Abbreviations

AUC: Area under the curve; GEO: Gene Expression Omnibus; IBD: Inflammatory bowel disease; iHMP: The Integrative Human Microbiome Project; SHAP: Shapley Additive exPlanations; UMAP: Uniform Manifold Approximation and Projection; XGBoost: XGBoost Extreme Gradient Boosting.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13619-022-00143-6.

Additional file 1: Supplementary Figure 1. Venn diagram: overlapped genes between the gene features selected by unsupervised clustering analysis and by the XGBoost. **Supplementary Figure 2. A** UMAP visualization of the integrated samples (GSE112366, GSE3365, GSE75214, and iHMP) shows the clustering based on the 32-gene signature, color coded by healthy controls, UC and CD. **B** UMAP showing 9 clusters based on the 32-gene signature, color coded with cluster identities. **C**-**N** Confusion matrix of 32-gene-based XGBoost classification models with 30 percent samples of GSE53306 (**C**), iHMP (**D**), GSE3365 (**E**), GSE112366 (**F**), GSE75214 (**G**), GSE6731 (**H**), GSE10616 (**I**), GSE38713 (**J**), GSE22619 (**K**), GSE9452 (**L**), and GSE1152 (**M**), and GSE83687 (N) that were not used for training and validation. Confusion matrix detailing the true positive (right lower), true negative (left upper), false positive (right upper), and false negative (left lower) predictions from XGBoostbased classification model. Accuracy = (true positive + true negative) / total. **Supplementary Figure 3. A** UMAP visualization of the integrated samples (GSE112366, GSE3365, GSE75214, and iHMP) shows the clustering based on the 19-gene signature, color coded by healthy controls, UC and CD. **B** UMAP showing 9 clusters based on the 19-gene signature, color coded with cluster identities.

Additional file 2: Supplementary Table 1. Marker genes filtered out with FindMarker function of Seurat.

Additional file 3: Supplementary Table 2. Annotation of 32-gene signature.

Additional file 4: Supplementary Table 3. List of 54 FDA-approved and failed target genes.

Additional file 5: Supplementary Table 4. 32-gene-based XGBoost-based classification model achieved better performance than most common models.

Additional file 6: Supplementary Table 5. Annotation of 19-gene signature.

Acknowledgements

Not applicable.

Authors' contributions

SCY and YGC conceived the study and wrote the manuscript. SCY analyzed the data. MXZ, YLW, and XW helped in organizing this manuscript. ZY helped bioinformatic analysis. All authors read and approved the final manuscript.

Funding

This study was supported by grants from Guangdong Postdoctoral Research Foundation (CN) (O0390302 to SCY), National Natural Science Foundation of China (31988101 and 31730056 to YGC).

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files. Requests for materials should be addressed to the corresponding author.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

YGC is the Editor-in-Chief of Cell Regeneration. He was not involved in the review or decision related to this manuscript. All the authors delare no competing interests.

Author details

[1]Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, 190 Kaiyuan Avenue, Guangzhou Science Park, Luogang District, Guangzhou 510530, China. [2]Guangzhou Laboratory, Guangzhou 510700, China. [3]The State Key Laboratory of Membrane Biology, Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China. [4]School of Medicine, Tsinghua University, Beijing 100084, China. [5]School of Basic Medicine, Nanchang University, Nanchang 330031, China.

Received: 1 August 2022 Accepted: 9 October 2022
Published online: 05 January 2023

References

Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. Eur Heart J. 2020;41(3):359–67. https://doi.org/10.1093/eurheartj/ehz565.

Ali M. PyCaret: an Open Source, Low-Code Machine Learning Library in Python. PyCaret version 1.0.0. https://pycaret.org/.

Allen IC, Wilson JE, Schneider M, Lich JD, Roberts RA, Arthur JC, et al. NLRP12 suppresses colon inflammation and tumorigenesis through the negative regulation of noncanonical NF-kappaB signaling. Immunity. 2012;36(5):742–54. https://doi.org/10.1016/j.immuni.2012.03.012.

Asano K, Matsumoto T, Umeno J, Hirano A, Esaki M, Hosono N, et al. Impact of allele copy number of polymorphisms in FCGR3A and FCGR3B genes on susceptibility to ulcerative colitis. Inflamm Bowel Dis. 2013;19(10):2061–8. https://doi.org/10.1097/MIB.0b013e318298118e.

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):D991-5. https://doi.org/10.1093/nar/gks1193.

Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018. https://doi.org/10.1038/nbt.4314.

Bergstrom K, Liu X, Zhao Y, Gao N, Wu Q, Song K, et al. Defective Intestinal Mucin-Type O-Glycosylation Causes Spontaneous Colitis-Associated Cancer in Mice. Gastroenterology. 2016;151(1):152-64 e11. https://doi.org/10.1053/j.gastro.2016.03.039.

Biancheri P, Brezski RJ, Di Sabatino A, Greenplate AR, Soring KL, Corazza GR, et al. Proteolytic cleavage and loss of function of biologic agents that neutralize tumor necrosis factor in the mucosa of patients with inflammatory bowel disease. Gastroenterology. 2015;149(6):1564-74 e3. https://doi.org/10.1053/j.gastro.2015.07.002.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20. https://doi.org/10.1038/nbt.4096.

Chen Y, Lei J, He S. m(6)A Modification Mediates Mucosal Immune Microenvironment and Therapeutic Response in Inflammatory Bowel Disease. Front Cell Dev Biol. 2021;9:692160. https://doi.org/10.3389/fcell.2021.692160.

Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–94. https://doi.org/10.1145/2939672.2939785

Cheng Y, Ma XL, Wei YQ, Wei XW. Potential roles and targeted therapy of the CXCLs/CXCR2 axis in cancer and inflammatory diseases. Biochim Biophys Acta Rev Cancer. 2019;1871(2):289–312. https://doi.org/10.1016/j.bbcan.2019.01.005.

De Vry CG, Prasad S, Komuves L, Lorenzana C, Parham C, Le T, et al. Non-viral delivery of nuclear factor-kappaB decoy ameliorates murine inflammatory bowel disease and restores tissue homeostasis. Gut. 2007;56(4):524–33. https://doi.org/10.1136/gut.2006.096487.

Eftychi C, Schwarzer R, Vlantis K, Wachsmuth L, Basic M, Wagle P, et al. Temporally Distinct Functions of the Cytokines IL-12 and IL-23 Drive Chronic Colon Inflammation in Response to Intestinal Barrier Impairment. Immunity. 2019;51(2):367-80 e4. https://doi.org/10.1016/j.immuni.2019.06.008.

Eshelman MA, Harris L, Deiling S, Koltun WA, Jeganathan NA, Yochum GS. Transcriptomic analysis of ileal tissue from Crohn's disease patients identifies extracellular matrix genes that distinguish individuals by age at diagnosis. Physiol Genomics. 2020;52(10):478–84. https://doi.org/10.1152/physiolgenomics.00062.2020.

Feagins LA, Holubar SD, Kane SV, Spechler SJ. Current strategies in the management of intra-abdominal abscesses in Crohn's disease. Clin Gastroenterol Hepatol. 2011;9(10):842–50. https://doi.org/10.1016/j.cgh.2011.04.023.

Frigerio S, Lartey DA, D'Haens GR, Grootjans J. The Role of the Immune System in IBD-Associated Colorectal Cancer: From Pro to Anti-Tumorigenic Mechanisms. Int J Mol Sci. 2021;22(23). https://doi.org/10.3390/ijms222312739.

Gadaleta RM, van Erpecum KJ, Oldenburg B, Willemsen EC, Renooij W, Murzilli S, et al. Farnesoid X receptor activation inhibits inflammation and preserves the intestinal barrier in inflammatory bowel disease. Gut. 2011;60(4):463–72. https://doi.org/10.1136/gut.2010.212159.

Gao C, Sun H, Wang T, Tang M, Bohnen NI, Muller M, et al. Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. Sci Rep. 2018;8(1):7129. https://doi.org/10.1038/s41598-018-24783-4.

Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature. 2020;578(7796):527–39. https://doi.org/10.1038/s41586-020-2025-2.

Gubatan J, Levitte S, Patel A, Balabanis T, Wei MT, Sinha SR. Artificial intelligence applications in inflammatory bowel disease: Emerging technologies and future directions. World J Gastroenterol. 2021;27(17):1920–35. https://doi.org/10.3748/wjg.v27.i17.1920.

Hamid Q, Tulic M. Immunobiology of asthma. Annu Rev Physiol. 2009;71:489–507. https://doi.org/10.1146/annurev.physiol.010908.163200.

Han L, Maciejewski M, Brockel C, Gordon W, Snapper SB, Korzenik JR, et al. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. Bioinformatics. 2018;34(6):985–93. https://doi.org/10.1093/bioinformatics/btx651.

Han T, Goswami S, Hu Y, Tang F, Zafra MP, Murphy C, et al. Lineage Reversion Drives WNT Independence in Intestinal Cancer. Cancer Discov. 2020;10(10):1590–609. https://doi.org/10.1158/2159-8290.CD-19-1536.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87 e29. https://doi.org/10.1016/j.cell.2021.04.048.

He L, Ma Y, Li W, Han W, Zhao X, Wang H. Protease-activated receptor 2 signaling modulates susceptibility of colonic epithelium to injury through stabilization of YAP in vivo. Cell Death Dis. 2018;9(10):949. https://doi.org/10.1038/s41419-018-0995-x.

Hong M, Ye BD, Yang SK, Jung S, Lee HS, Kim BM, et al. Immunochip Meta-Analysis of Inflammatory Bowel Disease Identifies Three Novel Loci and Four Novel Associations in Previously Reported Loci. J Crohns Colitis. 2018;12(6):730–41. https://doi.org/10.1093/ecco-jcc/jjy002.

Hosseinkhani B, van den Akker NMS, Molin DGM, Michiels L. (Sub)populations of extracellular vesicles released by TNF-alpha -triggered human endothelial cells promote vascular inflammation and monocyte migration. J Extracell Vesicles. 2020;9(1):1801153. https://doi.org/10.1080/20013078.2020.1801153.

Isakov O, Dotan I, Ben-Shachar S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. Inflamm Bowel Dis. 2017;23(9):1516–23. https://doi.org/10.1097/MIB.0000000000001222.

Khaki-Khatibi F, Qujeq D, Kashifard M, Moein S, Maniati M, Vaghari-Tabari M. Calprotectin in inflammatory bowel disease. Clin Chim Acta. 2020;510:556–65. https://doi.org/10.1016/j.cca.2020.08.025.

Khorasani HM, Usefi H, Pena-Castillo L. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. Sci Rep. 2020;10(1):13744. https://doi.org/10.1038/s41598-020-70583-0.

Kiesslich R, Duckworth CA, Moussata D, Gloeckner A, Lim LG, Goetz M, et al. Local barrier dysfunction identified by confocal laser endomicroscopy predicts relapse in inflammatory bowel disease. Gut. 2012;61(8):1146–53. https://doi.org/10.1136/gutjnl-2011-300695.

Koppad S, Basava A, Nash K, Gkoutos GV, Acharjee A. Machine Learning-Based Identification of Colon Cancer Candidate Diagnostics Genes. Biology (Basel). 2022;11(3). https://doi.org/10.3390/biology11030365.

Krishnan K, Arnone B, Buchman A. Intestinal growth factors: potential use in the treatment of inflammatory bowel disease and their role in mucosal healing. Inflamm Bowel Dis. 2011;17(1):410–22. https://doi.org/10.1002/ibd.21316.

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3. https://doi.org/10.1093/bioinformatics/bts034.

Li H, Lai L, Shen J. Development of a susceptibility gene based novel predictive model for the diagnosis of ulcerative colitis using random forest and artificial neural network. Aging. 2020;12(20):20471–82. https://doi.org/10.18632/aging.103861.

Lim SK, Orhant-Prioux M, Toy W, Tan KY, Lim YP. Tyrosine phosphorylation of transcriptional coactivator WW-domain binding protein 2 regulates estrogen receptor alpha function in breast cancer via the Wnt pathway. FASEB J. 2011;25(9):3004–18. https://doi.org/10.1096/fj.10-169136.

Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655–62. https://doi.org/10.1038/s41586-019-1237-9.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.

Manousou P, Kolios G, Valatas V, Drygiannakis I, Bourikas L, Pyrovolaki K, et al. Increased expression of chemokine receptor CCR3 and its ligands in ulcerative colitis: the role of colonic epithelial cells in in vitro studies. Clin Exp Immunol. 2010;162(2):337–47. https://doi.org/10.1111/j.1365-2249.2010.04248.x.

Martinez A, Nunez C, Martin MC, Mendoza JL, Taxonera C, Diaz-Rubio M, et al. Epistatic interaction between FCRL3 and MHC in Spanish patients with IBD. Tissue Antigens. 2007;69(4):313–7. https://doi.org/10.1111/j.1399-0039.2007.00816.x.

Martinez-Gimeno M, Gamundi MJ, Hernan I, Maseras M, Milla E, Ayuso C, et al. Mutations in the pre-mRNA splicing-factor genes PRPF3, PRPF8, and PRPF31 in Spanish families with autosomal dominant retinitis

pigmentosa. Invest Ophthalmol vis Sci. 2003;44(5):2171–7. https://doi.org/10.1167/iovs.02-0871.
Matsukawa T, Izawa K, Isobe M, Takahashi M, Maehara A, Yamanishi Y, et al. Ceramide-CD300f binding suppresses experimental colitis by inhibiting ATP-mediated mast cell activation. Gut. 2016;65(5):777–87. https://doi.org/10.1136/gutjnl-2014-308900.
Mo JS, Na KS, Yu JI, Chae SC. Identification of the polymorphisms in IFITM1 gene and their association in a Korean population with ulcerative colitis. Immunol Lett. 2013;156(1–2):118–22. https://doi.org/10.1016/j.imlet.2013.09.026.
Nimmo ER, Stevens C, Phillips AM, Smith A, Drummond HE, Noble CL, et al. TLE1 modifies the effects of NOD2 in the pathogenesis of Crohn's disease. Gastroenterology. 2011;141(3):972-81 e1-2. https://doi.org/10.1053/j.gastro.2011.05.043.
Ogunleye A, Wang QG. XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM Trans Comput Biol Bioinform. 2020;17(6):2131–40. https://doi.org/10.1109/TCBB.2019.2911071.
Ohtani N, Ohtani H, Oki M, Naganuma H, Nagura H. CXC chemokine receptor 1 (CXCR1) is expressed mainly by neutrophils in inflamed gut and stomach tissues. Tohoku J Exp Med. 2002;196(3):179–84. https://doi.org/10.1620/tjem.196.179.
Olafsson S, McIntyre RE, Coorens T, Butler T, Jung H, Robinson PS, et al. Somatic Evolution in Non-neoplastic IBD-Affected Colon. Cell. 2020;182(3):672-84 e11. https://doi.org/10.1016/j.cell.2020.06.036.
Ordas I, Eckmann L, Talamini M, Baumgart DC, Sandborn WJ. Ulcerative colitis. Lancet. 2012;380(9853):1606–19. https://doi.org/10.1016/S0140-6736(12)60150-0.
Pak S, Hwang SW, Shim IK, Bae SM, Ryu YM, Kim HB, et al. Endoscopic Transplantation of Mesenchymal Stem Cell Sheets in Experimental Colitis in Rats. Sci Rep. 2018;8(1):11314. https://doi.org/10.1038/s41598-018-29617-x.
Park YS, Chung SH, Lee SK, Kim JH, Kim JB, Kim TK, et al. Melatonin improves experimental colitis with sleep deprivation. Int J Mol Med. 2015;35(4):979–86. https://doi.org/10.3892/ijmm.2015.2080.
Park SK, Kim S, Lee GY, Kim SY, Kim W, Lee CW, et al. Development of a Machine Learning Model to Distinguish between Ulcerative Colitis and Crohn's Disease Using RNA Sequencing Data. Diagnostics (Basel). 2021;11(12). https://doi.org/10.3390/diagnostics11122365.
Pittayanon R, Lau JT, Leontiadis GI, Tse F, Yuan Y, Surette M, et al. Differences in Gut Microbiota in Patients With vs Without Inflammatory Bowel Diseases: A Systematic Review. Gastroenterology. 2020;158(4):930-46 e1. https://doi.org/10.1053/j.gastro.2019.11.294.
Pothuraju R, Krishn SR, Gautam SK, Pai P, Ganguly K, Chaudhary S, et al. Mechanistic and Functional Shades of Mucins and Associated Glycans in Colon Cancer. Cancers (Basel). 2020;12(3). https://doi.org/10.3390/cancers12030649.
Punit S, Dube PE, Liu CY, Girish N, Washington MK, Polk DB. Tumor Necrosis Factor Receptor 2 Restricts the Pathogenicity of CD8(+) T Cells in Mice With Colitis. Gastroenterology. 2015;149(4):993-1005 e2. https://doi.org/10.1053/j.gastro.2015.06.004.
Rutgeerts P, Vermeire S, Van Assche G. Biological therapies for inflammatory bowel diseases. Gastroenterology. 2009;136(4):1182–97. https://doi.org/10.1053/j.gastro.2009.02.001.
Sahoo D, Swanson L, Sayed IM, Katkar GD, Ibeawuchi SR, Mittal Y, et al. Artificial intelligence guided discovery of a barrier-protective therapy in inflammatory bowel disease. Nat Commun. 2021;12(1):4246. https://doi.org/10.1038/s41467-021-24470-5.
Shorthouse D, Riedel A, Kerr E, Pedro L, Bihary D, Samarajiwa S, et al. Exploring the role of stromal osmoregulation in cancer and disease using executable modelling. Nat Commun. 2018;9(1):3011. https://doi.org/10.1038/s41467-018-05414-y.
Thalor A, Kumar Joon H, Singh G, Roy S, Gupta D. Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. Comput Struct Biotechnol J. 2022;20:1618–31. https://doi.org/10.1016/j.csbj.2022.03.019.
Wei K, Zhang D, Hong J, Zhang C, Feng X, Huang Y, et al. Herb-Partitioned Moxibustion and the miRNAs Related to Crohn's Disease: A Study Based on Rat Models. Evid Based Complement Alternat Med. 2015;2015:265238. https://doi.org/10.1155/2015/265238.
Yu B, Yin YX, Tang YP, Wei KL, Pan ZG, Li KZ, et al. Diagnostic and Predictive Value of Immune-Related Genes in Crohn's Disease. Front Immunol. 2021;12:643036. https://doi.org/10.3389/fimmu.2021.643036.
Yuan F, Zhang YH, Kong XY, Cai YD. Identification of Candidate Genes Related to Inflammatory Bowel Disease Using Minimum Redundancy Maximum Relevance, Incremental Feature Selection, and the Shortest-Path Approach. Biomed Res Int. 2017;2017:5741948. https://doi.org/10.1155/2017/5741948.
Yuan X, Chen B, Duan Z, Xia Z, Ding Y, Chen T, et al. Depression and anxiety in patients with active ulcerative colitis: crosstalk of gut microbiota, metabolomics and proteomics. Gut Microbes. 2021;13(1):1987779. https://doi.org/10.1080/19490976.2021.1987779.
Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523. https://doi.org/10.1038/s41467-019-09234-6.
Zhu H, Wan X, Li J, Han L, Bo X, Chen W, et al. Computational Prediction and Validation of BAHD1 as a Novel Molecule for Ulcerative Colitis. Sci Rep. 2015;5:12227. https://doi.org/10.1038/srep12227.