Journal of Inflammation Research

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

An Integrative Analysis of Transcriptome Combined with Machine Learning and Single-Cell RNA-Seq for the Common Biomarkers in Crohn's Disease and Kidney Stone Disease

Jiejie Zhu[1,*], Yishan Du[2,*], Luyao Gao[3], Jiajia Wang[3], Qiao Mei[1]

[1]Department of Gastroenterology, The First Affiliated Hospital of Anhui Medical University, Hefei City, Anhui Province, People's Republic of China; [2]Geriatric Department, The First Affiliated Hospital of Ningbo University, Ningbo City, Zhejiang Province, People's Republic of China; [3]Department of Pharmacology, School of Basic Medical Sciences, Anhui Medical University, Hefei City, Anhui Province, People's Republic of China

*These authors contributed equally to this work

Correspondence: Qiao Mei; Jiajia Wang, Email meiqiao@hotmail.com; wjj@ahmu.edu.cn

Background: The course of Crohn's disease (CD) is prolonged and many of them may develop kidney stone disease (KSD) with the need for surgical treatment. Therefore, finding biomarkers that can predict CD with KD become increasingly important.

Methods: We obtained three CD and one KSD dataset from GEO database. DEGs and module genes were identified utilizing Limma and WGCNA. We constructed a protein-protein interaction (PPI) network and employed machine learning algorithms to pinpoint potential hub genes (HGs) for diagnosing CD with KSD. We developed a nomogram and receiver operating characteristic (ROC) curve. Additionally, human intestinal cell and proximal tubular epithelial cell models were established to explore the HG levels. Next, we used Cytoscape to build the regulatory networks. Finally, single-cell analysis was performed to investigate specific cell types displaying these biomarkers in CD.

Results: We identified 36 common genes associated with CD and KSD. PYY, FOXA2, REG3A, REG1A, REG1B were identified as HGs utilizing the machine learning algorithm. The nomogram and all five potential HGs exhibited strong diagnostic capabilities. Cell experiments also verified that these genes were markedly expressed in cell models of CD and KSD. Meanwhile, we pinpointed four microRNAs and three transcriptional regulators intimately linked to five crucial genes. Finally, single-cell analysis indicated FOXA2, REG3A, REG1A and REG1B exhibited elevated expression in goblet cells, whereas PYY demonstrated high expression levels in coloncytes.

Conclusion: We determined five biomarkers, including PYY, FOXA2, REG3A, REG1A, REG1B. Our results offer useful perspectives for identifying CD with KSD.

Keywords: Crohn's disease, kidney stone disease, hub genes, bioinformatics analysis, machine learning, single-cell RNA-seq

Introduction

Inflammatory bowel disease (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), represents a chronic intestinal condition with a high recurrence rate that can lead to varying degrees of damage to the digestive system, dysfunction of organs and even disability.[1] CD is highly prevalent, and its incidence is increasing in adolescent and pediatric populations.[2] The CD is considered to be a heterogeneous disease, and an intricate interplay among genetic predisposition, specific environmental elements, and modified intestinal microbiota may lead to the dysregulation of immune responses in CD.[3–5]

Urolithiasis, also referred to as nephrolithiasis or kidney stone disease (KSD), ranks among the most prevalent urinary system disorders.[6] The prevalence and incidence of kidney stones have increased globally, partly due to the

Received: 7 November 2024
Accepted: 21 March 2025
Published: 10 April 2025

advancements in medical imaging.[7] Generally, urinary supersaturation and crystallization are influenced by a variety of factors, such as urine pH and concentrations of calcium, uric acids, and drugs, contributing to the development of kidney stones.[8] The calcium stone found in KSD primarily consists of calcium oxalate (CaOx), which may be present in a uniform composition or combined with calcium phosphate (CaP) and uric acid.[9] Research indicates that IBD is linked to an elevated incidence of kidney stone formation, occurring in 5 to 10%.[10] Moreover, a larger number of cases of KSD was more likely to be found in CD patients instead of UC, indicating that there was a strong correlation between KSD and CD.[11] The underlying aetiology of renal stones in CD is believed to be complex, involving intestinal dysfunction and enteric hyperoxaluria (EHO).[12] Up to 75% of patients with CD require intestinal resection and the association between kidney stone formation and bowel surgery in CD is widely recognized, particularly may occur in patients following ileostomy formation.[13,14] Intestinal inflammatory response and imbalances in the microbiota are also important factors leading to the occurrence of kidney stones in CD.[15]

KSD patients who suffer from CD are often asymptomatic, but oxalate deposition and recurrent urolithiasis can lead to severe chronic kidney disease in the long course of CD.[16] Therefore, early diagnosis and intervention have great clinical value in the subsequent development of the disease. Emerging research has shown genetic evidence supporting the causative relationship between genetically predicted CD and KSD, as well as suggesting that CD could raise the risk of KSD.[17] However, the underlying mechanisms of CD with KSD are far from elucidated and require further study. In our research, we employed diverse integrated bioinformatics approaches to uncover the hub genes (HGs) of CD with KSD by collecting three CD and one KSD dataset from the gene expression omnibus (GEO) repository. Such findings will likely lead to the identification of additional treatment targets.

Materials and Methods

Collection of Microarray Data

The research workflow is illustrated in Figure 1. The data used for analysis and validation came from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). Among them, genetic information from the CD datasets GSE95095, GSE36807 and GSE6731 and KSD-related dataset GSE73680 were mainly used for analysis, and the datasets GSE75214 and GSE36446 were used for verification of the results obtained from the analysis. It is important to point out that there are only a limited number of datasets on KSD. Among them, GSE117518 was excluded because of the small sample size (3 KSD tissues, 3 normal tissues). Hence, we chose the validation set for KSD (GSE36446) while it is a rat dataset. Meanwhile, the single cell dataset GSE214695 was also included in this investigation. Table 1 provides comprehensive details regarding the datasets.

Data Preprocessing

We downloaded the unprocessed and series matrix files of GSE95095, GSE36807 and GSE6731 from the GEO database. Regarding the unprocessed data, we extracted and standardized the probe expression matrices utilizing the R package "inSilicoMerging". Next, we used an robust method of empirical Bayes (EB) to adjust for batch effects in data.[18] Finally, the combination of multiple data sets was obtained after removing the batch effect.

Screening of Differentially Expressed Genes (DEGs)

The package "limma" was utilized in combined CD dataset as well as KSD dataset to screen the differentially expressed genes (DEGs), with meeting the threshold of Fold change > 1.5 and p <0.05. Meanwhile, the R packages "ggplot2" and "pheatmap" were utilized to create heat maps and volcano plots of DEGs between groups.

Weighted Gene Co-Expression Network Analysis (WGCNA)

To determine the crucial components associated with KSD, we conducted a WGCNA analysis via the "WGCNA" package in R software.[19] First, we computed the MAD (Median Absolute Deviation) for individual genes utilizing gene-expression arrays, and subsequently eliminated any abnormal samples. The association index between each gene pair was calculated to create a resemblance matrix. The adjacency was subsequently converted into a topological overlap matrix

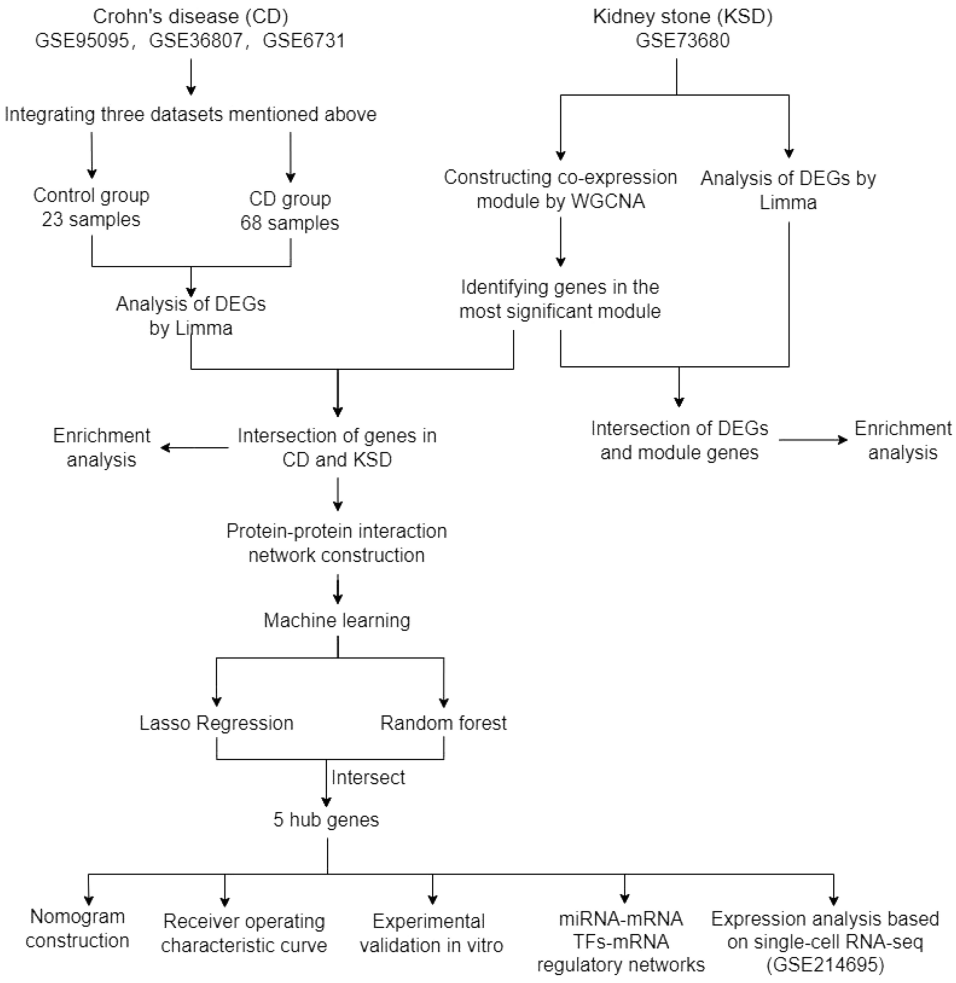

Figure 1 The workflow for this study.

(TOM), which was employed to determine the corresponding dissimilarity (1-TOM) and assess the network connectivity of a gene. The next phase involved identifying modules using hierarchical clustering and a dynamic tree cut algorithm. Grouping genes with similar expression patterns into gene modules was accomplished through average linkage hierarchical clustering, utilizing a TOM-based dissimilarity measure and setting a minimum threshold of 30 genes for the gene dendrogram. To further examine the modules, we selected a cut-off point for the module dendrogram, evaluated the dissimilarity of the characteristic genes within the modules, and merged several modules. In the end, eight co-expression

Table 1 Basic Information of GEO Datasets Utilized in This Investigation

GSE Series	Experiment Type	Sample Size		Platform	Species	Sample Source
		Control	CD			
GSE95095	Expression profiling by array	12	48	GPL14951	Homo sapiens	Intestinal biopsy
GSE36807	Expression profiling by array	7	13	GPL570	Homo sapiens	Intestinal biopsy
GSE6731	Expression profiling by array	4	7	GPL8300	Homo sapiens	Colonic mucosa
GSE75214	Expression profiling by array	22	75	GPL6244	Homo sapiens	Colonic mucosa
GSE214695	Expression profiling by high throughput sequencing	6	6	GPL18573	Homo sapiens	Colonic mucosa
		Control	KSD			
GSE73680	Expression profiling by array	6	29	GPL17077	Homo sapiens	Renal tissues
GSE36446	Expression profiling by array	6	6	GPL6101	Rattus norvegicus	Renal tissues

https://doi.org/10.2147/JIR.S502513

modules were identified. The modules having the most significant correlation with KSD were regarded as the key modules ($p < 0.05$).

Enrichment of Functions Analysis

The Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database are widely utilized to extract important functional pathways and processes of genes.[20] Analyses of GO and KEGG were executed utilizing the R package "clusterProfiler". The Sangerbox platform (http://vip.sangerbox.com/) provided us with visualization tools to present the findings for enrichment analysis.

The value of $p < 0.05$ is set to reach statistical significance for the standard.

Identification of Protein-Protein Interactions (PPIs)

In addition to the functional enrichment analysis, we also constructed the PPI map to explore the associations between the protein molecules we were interested in. The STRING repository serves as an online platform designed to build comprehensive protein association networks across organisms, encompassing both physical interactions and functional relationships.[21] PPI network on the basis of STRING (https://cn.string-db.org/, version 12.0), was constructed for important interacted gene identification. Homo sapiens was chosen as the target organism for further investigation. Interaction scores exceeding a confidence threshold of 0.400 were employed as the cut-off criteria for network visualization.

Screening HGs via Machine Learning

Following the acquisition of the candidate genes, a further screening process was conducted using two machine learning algorithms, encompassing LASSO regression and random forest (RF). Machine learning models have been spotted in cancer diagnosis and health care.[22,23] The above two kinds of algorithms were implemented utilizing the "glmnet" package and "randomForest" package in R software. The convergence of genes screened by LASSO and RF algorithms was considered to be the HGs for the diagnosis of CD with KSD.

Nomogram Plotting and Receiver Operating Characteristic (ROC) Analytic Method

In order to predict diagnosis in CD with KSD, nomogram is constructed based on logistic regression analysis with multiple factors by utilizing the R package "rms".[24]

Using the expression levels of individual genes, the score of each gene was calculated. The aggregate score signifies the total of all the aforementioned gene scores, facilitating prediction of the diagnosis of CD with KSD. Additionally, the ROC curve was developed to assess the predictive utility of the identified genes. After that, the 95% confidence interval (CI) and the area under the curve (AUC) were computed using the SPSS Statistics software. At last, the diagnostic model's efficacy was evaluated utilizing the validation datasets GSE75214 and GSE36446 to assess its diagnostic potential.

Prediction of miRNAs and Transcription Factors That Interact with HGs

Three online miRNA databases, including Targetscan(v8.0), miRWalk(v3.0), miRDB(v6.0), were employed to forecast the targeted miRNAs of five HGs. Drawing from the gene-miRNA targeting relationships, choose miRNAs that were included in at least two databases, then display and examine the co-expression network using CytoScape. Additional usage of the JASPAR profile database (http://jaspar.genereg.net) is needed to identify transcription factors (TFs) that interact with HGs.

Single-Cell RNA-Seq Analysis

CD single-cell sequencing dataset Sequencing data were procured from the GEO database (GSE214695) using 6 CD samples with 6 control samples for subsequent analysis.[25] Single-cell gene expression matrices were integrated using the Seurat R software package. Mitochondrial gene expression levels were calculated for each cell using the PercentageFeatureSet function, and expression complexity was calculated for each cell using Log10 (number of

genes)/Log10 (number of RNA transcripts). Substandard cells were eliminated using the subsequent criteria: (i) the quantity of identified genes was ≥200; (ii) the proportion of mitochondrial genes was <10%; and (iii) the count of erythrocytes was <3%. The data were integrated and standardized using the R software package "harmony", the standardized matrix was downscaled, and the single-cell data were mapped onto two-dimensional coordinates utilizing the Uniform manifold approximation and projection (UMAP). Data were mapped on a two-dimensional coordinate system, and UMAP plots were drawn according to their pathology-control profile. Cell clustering was performed on the single-cell data using the FindNeighbors and FindClusters functions in Seurat, respectively, and cell clusters were annotated using known marker genes.

Cell Culture and Treatment

The human colon adenocarcinoma cell line Caco-2 was provided by the National Collection of Authenticated Cell Cultures, Chinese Academy of Sciences. Human proximal tubular epithelial cell line HK-2 was procured from American Type Cell Culture (ATCC). Firstly, these two kinds of cells were maintained in a medium with Dulbecco's modified Eagle's medium (DMEM) comprising 10% fetal bovine serum (FBS) under a humidified environment of 5% CO_2 at 37°C. According to experimental needs, we used Lipopolysaccharide (LPS) (1μg/mL, Sigma-Aldrich) to stimulate Caco-2 cells for 24 hours while HK-2 cells were treated with CaOx crystals (2 mm; Sigma-Aldrich) for 24 h, as previously described.[26,27]

qRT-PCR

The extraction of total RNA from cells was executed utilizing the TRIzol reagent. The processes for RNA isolation and qRT-PCR were executed following the methodology delineated in an earlier investigation.[28] As an internal reference, GAPDH was employed. The following oligonucleotide primer sequences were utilized in this investigation: Forward: ACTTTGTCAAGCTCATTTCC, reverse: TGCAGCGAACTTTATTGATG for GAPDH. Forward: ACGGTCGCAATGCTGCTAAT, reverse: AAGGGGAGGTTCTCGCTGTC for PYY. Forward: TGGAGCAGCTACTATGCAG, reverse: CGTGTTCATGCCGTTCATC for FOXA2. Forward: GGCACCGAGCCCAATG, reverse: GGATTTCTCTCCCATGCAAAGT for REG3A. Forward: AGGAGAGTGGCACTGATGACTT, reverse: TAGGAGACCA GGGACCCACTG for REG1A. Forward: GGAGAGTAGCACTGATGACAGC, reverse: TCCCAGGACTTGTAGGAGACCA, for REG1B.

Statistics Analytical Method

R software version 4.3.3 and SPSS Version 26.0 were used to perform statistical analyses. We employed the independent Student's *t*-test to assess the statistical significance between variables with normal distribution in the two sets of continuous data. Conversely, the Mann–Whitney *U*-test was utilized to find differences between non-normal data. P-values below 0.05 indicated statistical significance for differences between groups.

Results

Identification of Differentially Expressed Genes

We chose three microarray datasets: GSE95095, GSE36807, and GSE6731, encompassing 68 CD-affected samples and 23 normal control samples. Examining the data prior to batch correction (Figure 2A–C) and following batch correction (Figure 2D–F) reveals that the batch effect in the combined data was eliminated, and the data underwent normalization. Initially, 745 DEGs were identified in the merged CD dataset utilizing the Limma approach (Fold change > 1.5 and *p*-value <0.05), with 474 exhibiting upregulation and 271 showing downregulation. The visualization of the previously mentioned data through heat maps and volcano plots is depicted in Figure 3A and B. Likewise, 5584 DEGs were screened out in the KSD dataset (GSE73680), comprising 4376 upregulated and 1208 down-regulated genes (Figure 4A and B).

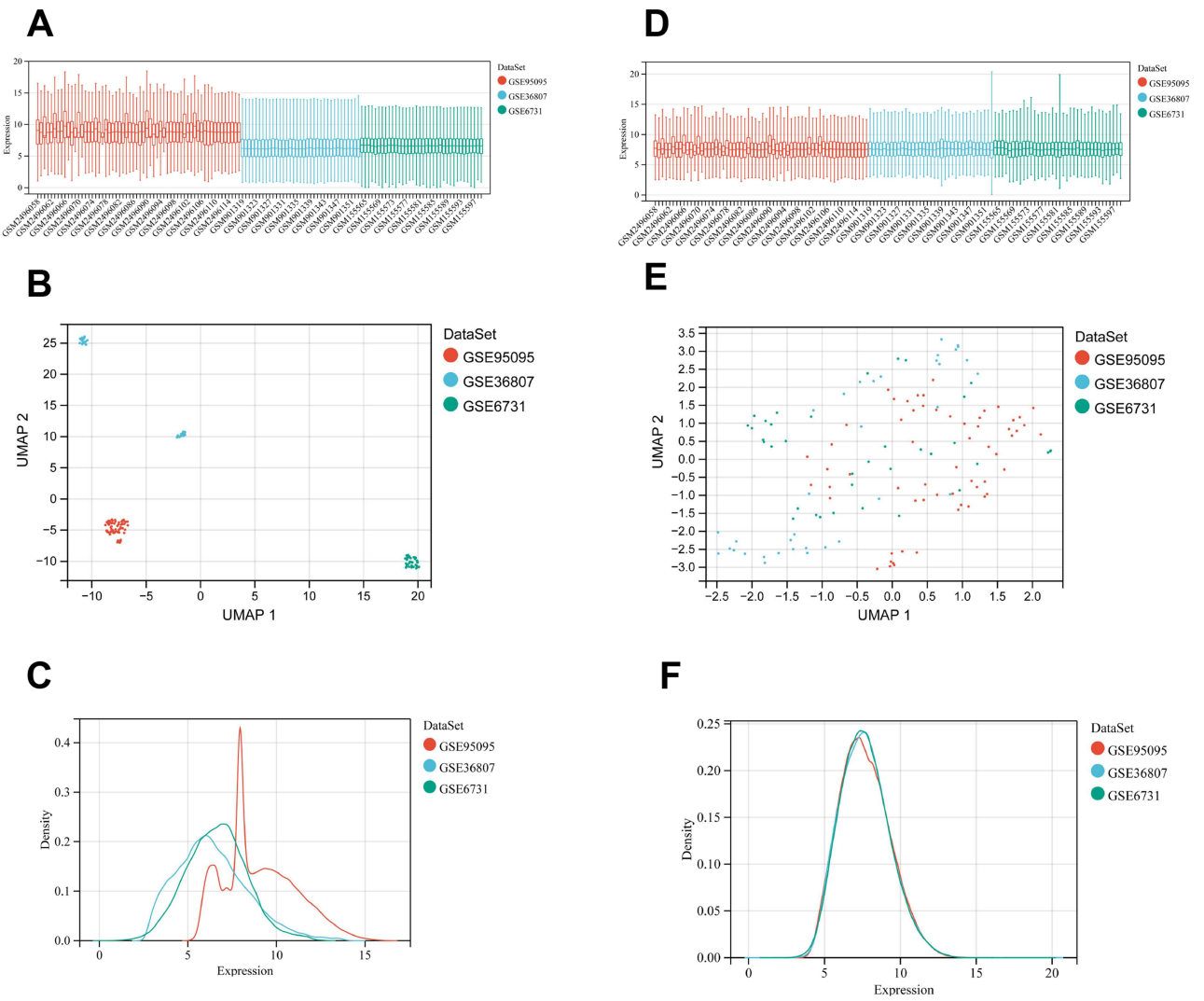

Figure 2 Data preprocessing. Box plot, principal component analyses and expression density plot were performed to remove batch correction of GSE95095, GSE36807 and GSE6731. (**A**–**C**) before batch correction and (**D**–**F**) after batch correction.

Weighted Gene Co-Expression Network Analysis and Identification of Key Modules

We employ WGCNA to pinpoint the module most closely linked to KSD. A "soft" threshold β of 28 (scale-free R^2 = 0.86) was chosen on the basis of the scale independence and mean connectivity (Figure 4C and D). Figure 4E illustrates the clustering tree graph of the KSD and control. Leveraging this analysis, eight colored gene co-expression modules were constructed, as shown in Figure 4F and G. Among them, the darkgreen module (Figure 4H) encompassing 2625 potential HGs, emerged as the central module due to its superior gene significance (r = 0.42, p = 0.01). Figure 4I presents the outcomes of a correlation examination between module membership and gene significance within the darkgreen module (r = 0.34). Hence, the darkgreen module was regarded as the key module for further analysis.

Functional Enrichment Analysis of Kidney Stone Disease

To assess if this dataset accurately represents the pathogenesis of KSD, we performed functional enrichment analysis utilizing the overlap between darkgreen module genes obtained from WGCNA and DEGs generated from Limma analysis of kidney stone dataset (GSE73680). Figure 5A shows the total number of shared genes that were found (1793). KEGG analysis showed that these genes were predominantly concentrated in the "Metabolic pathways" and "Antigen processing and presentation" (Figure 5B). GO investigation demonstrated that CGs were chiefly clustered in

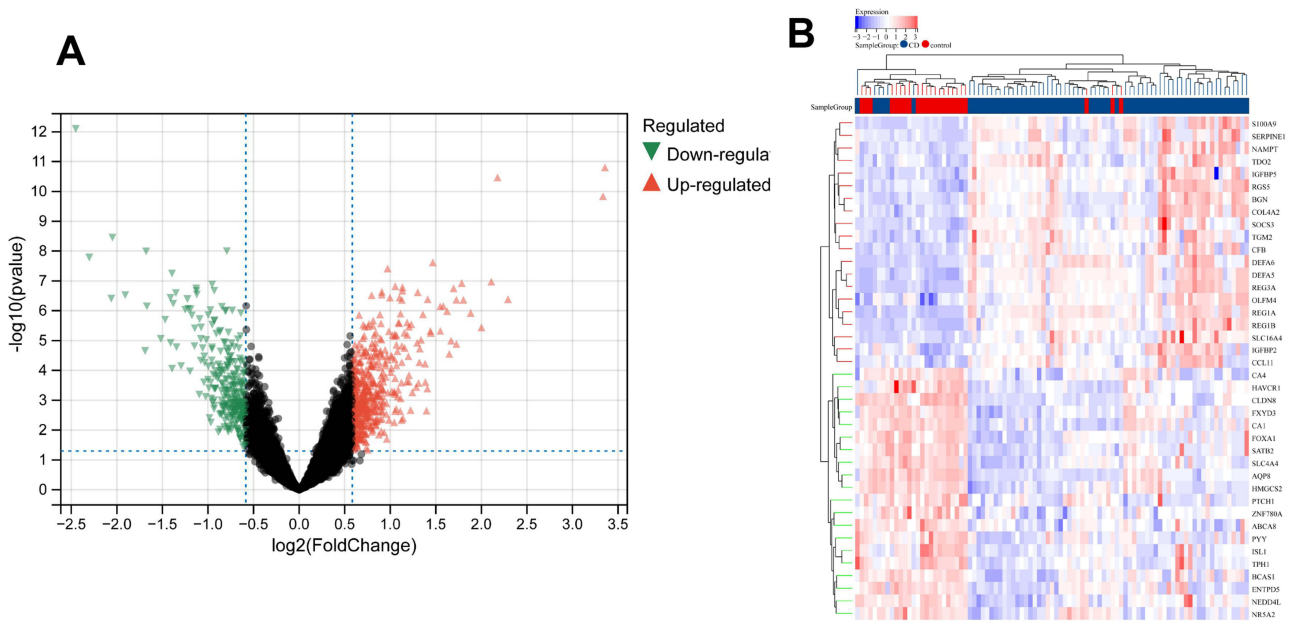

Figure 3 DEGs identified between CD and control groups from the integrated CD dataset. (**A**) Red and green in the volcano plot show the DEGs with significantly higher and lower gene expression level, respectively. (**B**) The heatmap displays the top upregulated and downregulated 20 genes in the CD and control groups.
Abbreviations: DEGs, differentially expressed genes; CD, Crohn's disease.

biological process (BP) terms, encompassing "chronic inflammatory response", "endocrine process" and "regulation of chronic inflammatory response" (Figure 5C). Concerning the cellular component (CC) ontology, the CGs were primarily situated in the "cell body membrane", "Golgi lumen" and "transporter complex" (Figure 5D). Molecular function (MF) assessment revealed that "transmembrane signaling receptor activity" was the most crucial item (Figure 5E). The enrichment study demonstrated that the immune and inflammatory responses were closely associated with KSD and dependable for subsequent CD examination.

Enrichment Analysis of CD with KSD and using the Network of Protein-Protein Interactions to Identify Node Genes

In order to investigate whether KSD-associated crucial genes could be connected to the pathogenesis of CD, 36 genes were found from the intersection of genes obtained from three CD datasets via Limma analysis and KSD module genes, as shown by the Venn diagram (Figure 6A). The KEGG analysis was executed utilizing these genes and the findings revealed that 36 genes were most often involved in pathways encompassing "Neuroactive ligand-receptor interaction" and "cAMP signaling pathway", as Figure 6B illustrates, were all intimately linked to the immune system. GO analysis discovered that these genes were associated with the "epithelium development", "regulation of signaling receptor activity" (BP); "extracellular region" and "extracellular space" (CC); and "signaling receptor binding", "receptor ligand activity" and "receptor regulator activity" (MF) (Figure 6C–E). Subsequently, we established an interactome network to pinpoint potential interacting HGs. The PPI network is depicted in Figure 6F, indicating potential interactions between 14 genes. These fourteen genes were sorted according to node numbers in Figure 6G.

Machine Learning-Based Candidate HG Screening

Potential genes for nomogram development and diagnostic assessment were identified using LASSO regression and RF machine learning approaches. Figure 7A and B illustrated that the LASSO regression method pinpointed six prospective candidate markers, while Figure 7C and D demonstrated how the RF algorithm ranked genes based on their calculated significance (Figure 7C and D). Finally, the six potential candidate genes for LASSO were intersected with the top six

https://doi.org/10.2147/JIR.S502513

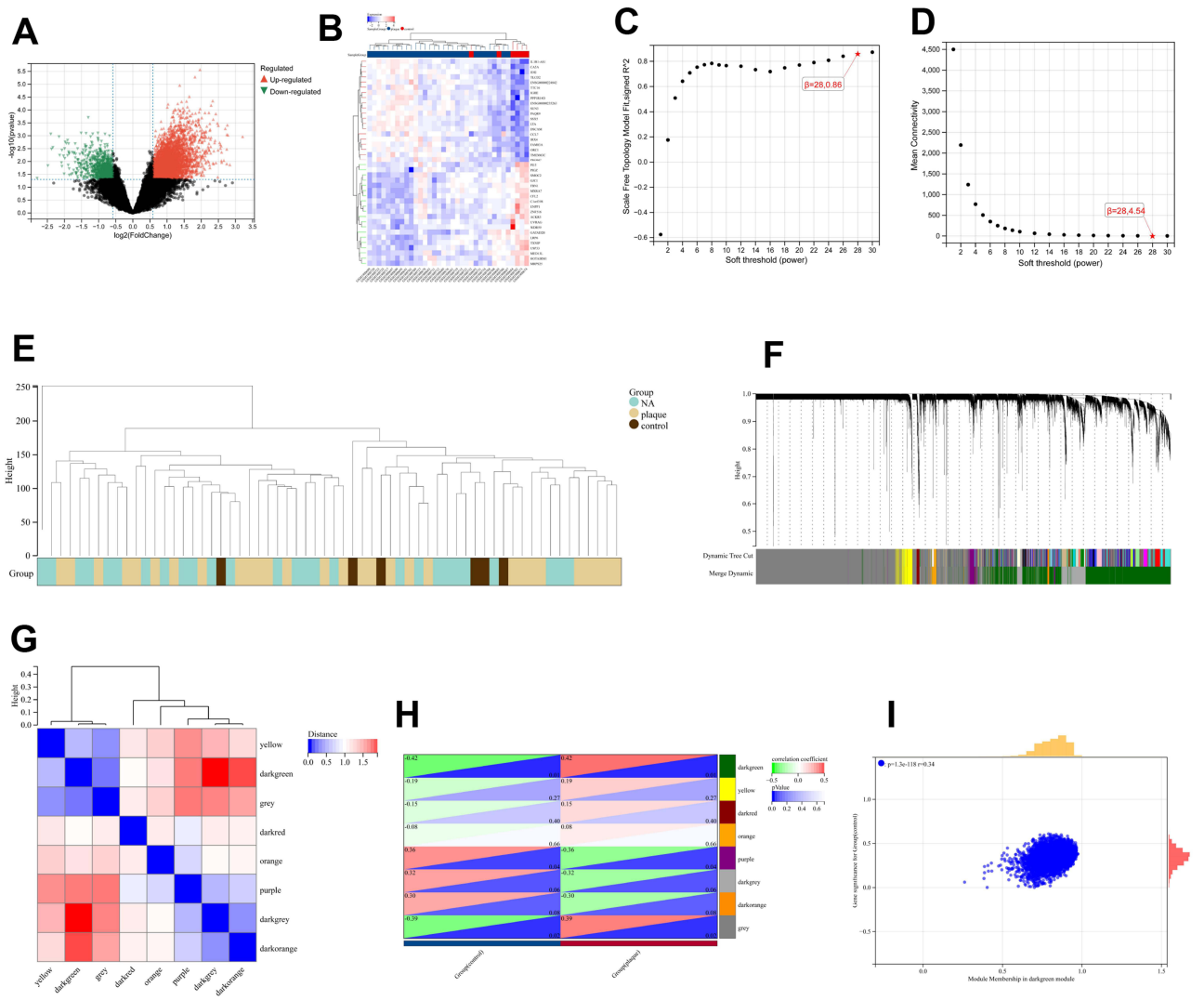

Figure 4 Identification of DEGs via Limma and module genes via weighted gene co-expression network analysis in KSD. (**A**) Volcano plot of DEGs in GSE73680. (**B**) Heatmap of DEGs in GSE73680. (**C** and **D**) Scale Independence and mean connectivity in GSE73680. (**E**) Cluster dendrogram of genes. (**F**) Gene co-expression modules represented by different colors under the gene tree. (**G**) Heatmap of eigengene adjacency. (**H**) Heatmap of the association between modules and KSD. (**I**) Correlation of module membership and gene significance in the darkgreen module.

Abbreviations: Limma, linear models for microarray data; DEGs, differentially expressed genes; KSD, Kidney stone disease.

most significant genes from RF algorithm and we obtained five HGs (PYY, FOXA2, REG3A, REG1A and REG1B) by intersecting the two machine learning results (Figure 7E).

Construction of a Nomogram and Diagnostic Value Assessment

To corroborate the the findings of our aforementioned bioinformatics investigation, we initially assessed the expression profiles of the five HGs in the combined CD dataset as well as KSD dataset GSE73680. The box plots illustrated that PYY, FOXA2, REG3A, REG1A, and REG1B exhibited notably elevated expression levels in individuals with CD or KSD relative to healthy subjects (Figure 8A and B). Next, we constructed the nomogram with five HGs (Figure 8C). As depicted in Figure 8C, the nomogram with five HGs was constructed. The ROC curves for the diagnostic model including PYY, FOXA2, REG3A, REG1A and REG1B were plotted, with AUCs of 0.958 and 0.862 respectively in the two cohorts (Figure 8D and E). Next, the diagnostic value of this model was confirmed using two external data set. In the GSE75214 validation set of CD, the AUC of this model was 0.972 (Figure 8F). Although the KSD gene set is limited and lack of the gene information of REG1B in the GSE36446 validation set of KSD, the diagnostic model of the

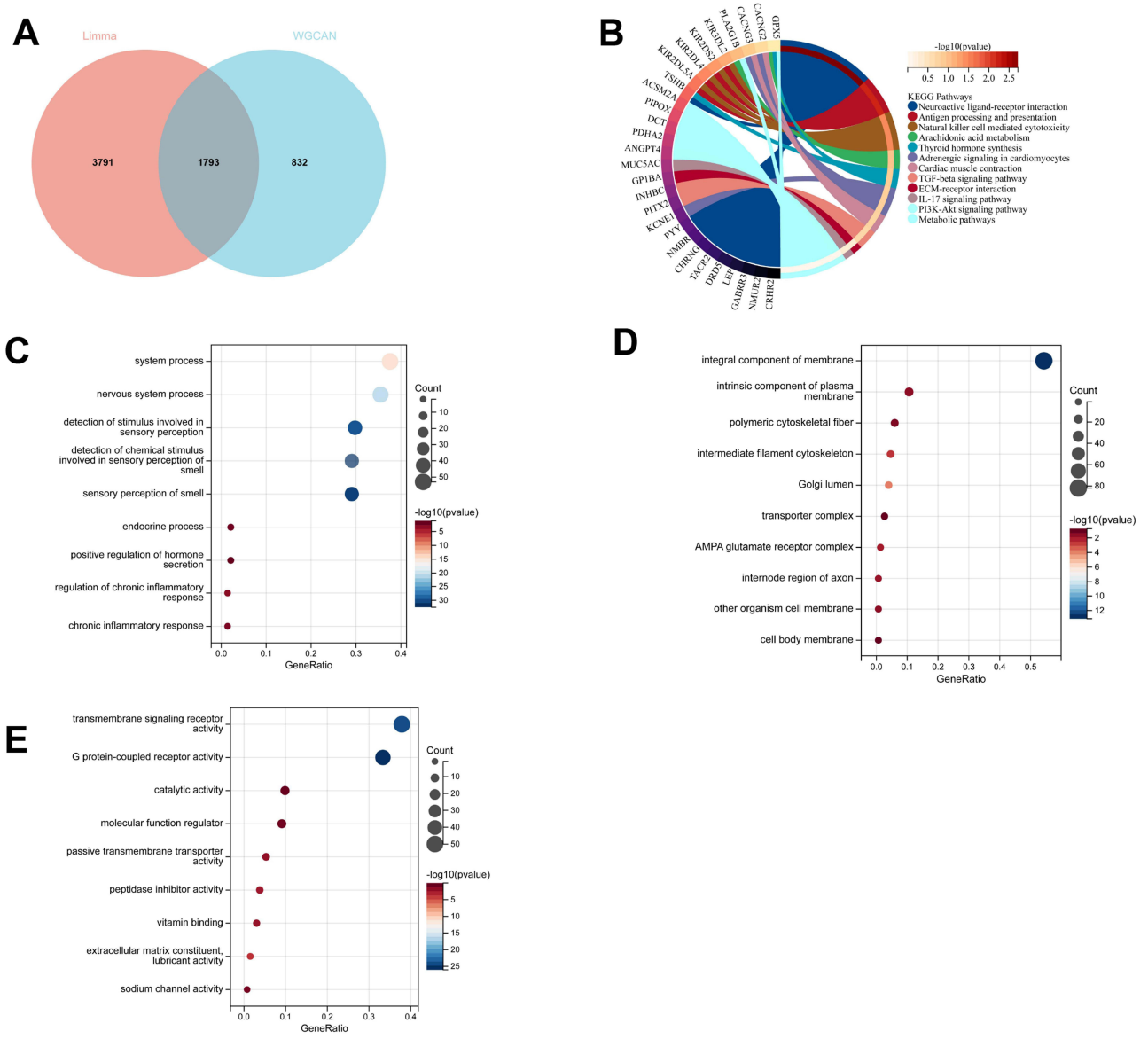

Figure 5 KEGG, GO analysis of the intersection of genes for KSD. (**A**) Venn diagram for overlapped genes between genes obtained from Limma analysis and WGCNA in KSD. (**B**) KEGG analysis of the overlapped genes. (**C**–**E**) Molecular function, cellular component, and biological process are all included in the GO analysis. GO terms are represented by the y-axis, while the gene ratio involved in related GO terms is represented by the x-axis. Gene numbers are represented by the size of the circles, and *p*-value is indicated by their color.

Abbreviations: KSD, Kidney stone disease; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; WGCNA, weighted gene co-expression network analysis; Limma, linear models for microarray data.

remaining four genes still had an AUC value of 0.944, showing excellent performance (Figure 8G). These results demonstrated a good predictive value for the diagnosis of CD with KSD. Additionally, we performed cell experiments to establish cell models for the study of CD and KSD. The qRT-PCR findings validated that the expression levels of PYY, REG3A, REG1A and REG1B were markedly elevated, while FOXA2 was decreased in Caco-2 cell model treated with LPS compared with the control samples (Figure 8H). HK-2 cells were treated with 2 mm CaOx to explore the expression of five HGs in CaOx stones and the results revealed a higher level of all five genes in the CaOx group relative to the control group (Figure 8I).

The Regulatory Network Analysis

Moreover, a co-expression network of mRNA and miRNA comprising 5 hGs, 109 nodes and 110 edges was visualized utilizing Cytoscape (Figure 9A). Among the miRNAs, miR-571, miR-1275, miR-4775, and miR-765 were the common

https://doi.org/10.2147/JIR.S502513

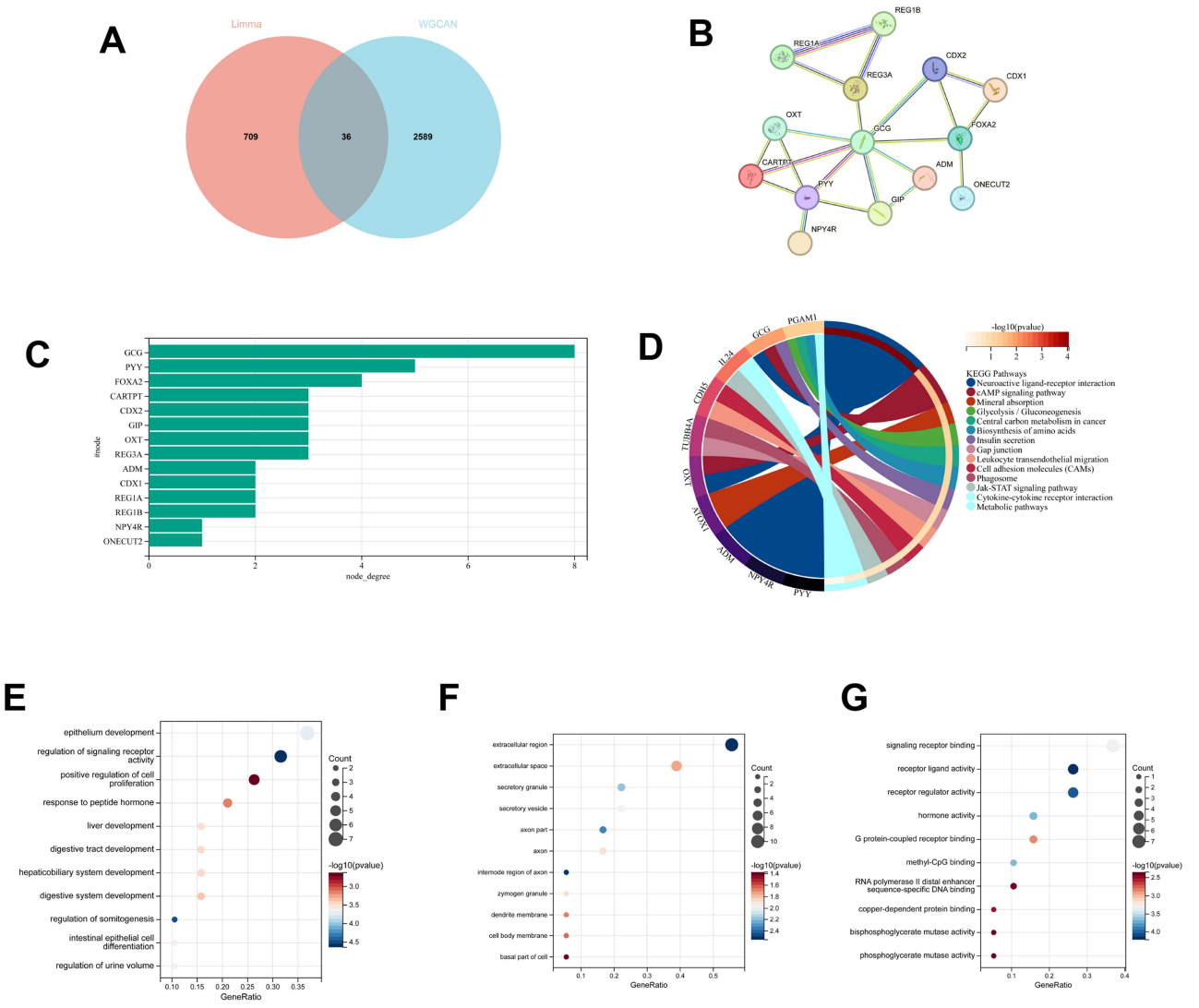

Figure 6 Enrichment analysis of genes related to CD and KSD and the discovery of node genes from PPI network. (**A**) Venn diagram shows that 36 common genes are identified from the intersection of genes in CD via Limma method and KSD with WGCNA. (**B**) KEGG analysis of the overlapped genes. (**C**–**E**) GO analysis (biological process, cellular component, and molecular function) of 36 genes. (**F**) 14 genes interact with each another, according to the PPI network. (**G**) The column displays the gene nodes of 14 genes in PPI network.

Abbreviations: CD, Crohn's disease; KSD, Kidney stone disease; Limma, linear models for microarray data; WGCNA, weighted gene co-expression network analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; PPI, protein-protein interaction.

miRNAs targeting two of the HGs. As depicted in Figure 9B, the regulatory network predicted TFs related to five HGs. The degree values of FOXC1, NR3C1, and GATA2 were equal to 3. Hence, they might be crucial regulators in the development of CD with KSD.

Single-Cell Dataset Analysis of HGs

The cellular distribution of PYY, FOXA2, REG3A, REG1A and REG1B, as well as the associated cell populations, was validated utilizing single-cell information from GSE214695. We obtained 11 cell groups by dimensionality reduction analysis, and the genes with elevated expression for each cluster were visualized on a heat map (Figure 10A–C). 10 kinds of cells, including T cells, coloncytes, plasma cells, macrophages, fibroblasts, B cells, goblet cells, mast cells, epithelial cells and glial cells were obtained by annotation (Figure 10D). Obviously, the infiltration levels of immune cells encompassing B cells, macrophages, and T cells were relatively high in CD patients (Figure 10E and F). Finally, the expression levels of PYY, FOXA2, REG3A, REG1A and REG1B in the different types of cells are shown in Figure 10G and H. The findings

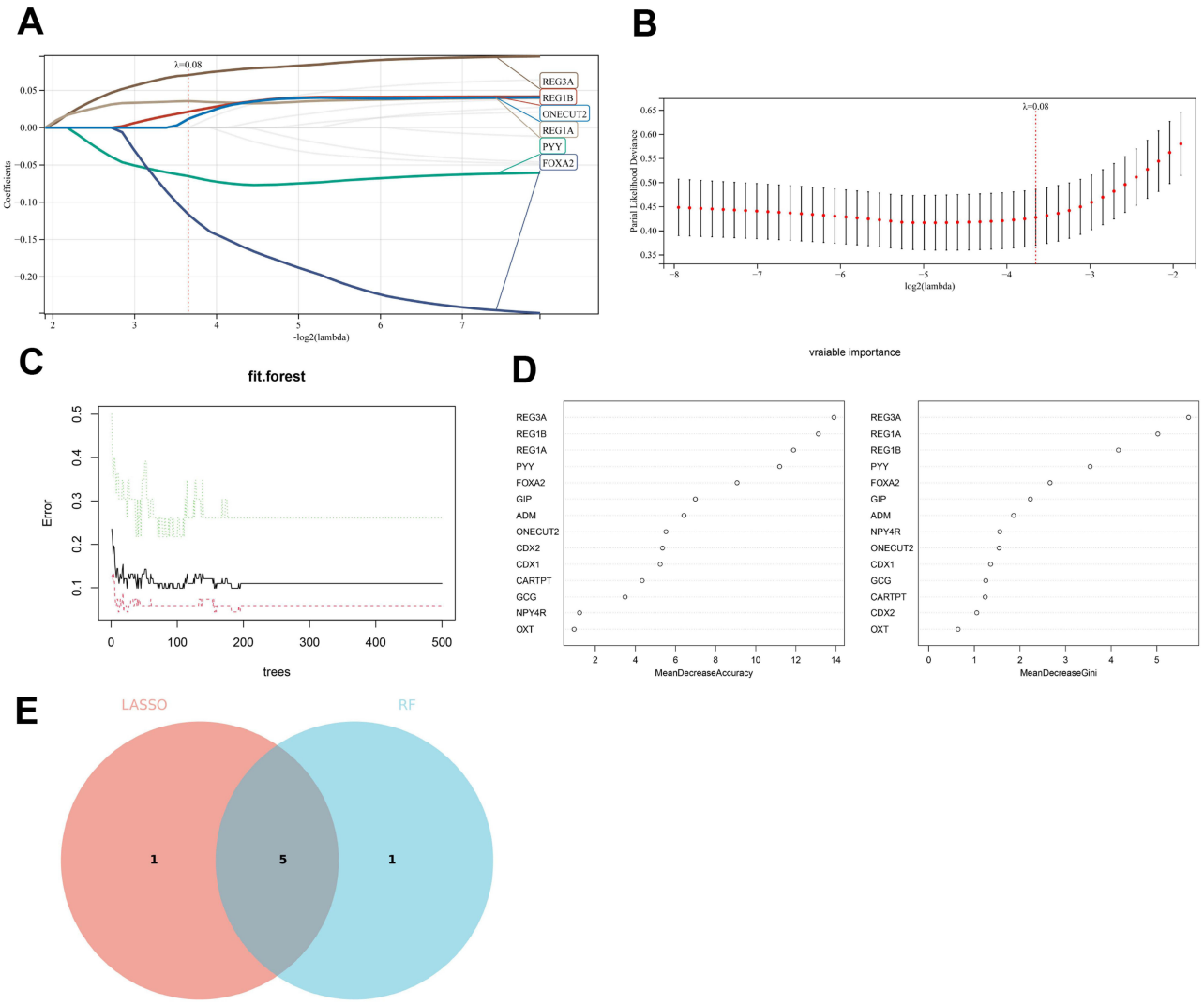

Figure 7 Machine learning in screening candidate HGs for CD with KSD. (**A** and **B**) The Lasso model for biomarker screening. The most appropriate number of genes (n = 6) for the diagnosis of CD with KSD is the one that corresponds to a minimum in the curve the curve. (**C**) Random Forest approach for selecting genes. (**D**) 14 genes are ordered by the importance score through the random forest algorithm. (**E**) Venn diagram demonstrates that the results of two aforementioned methods have an overlap in genes.
Abbreviations: CD, Crohn's disease; KSD, Kidney stone disease.

indicated that FOXA2, REG3A, REG1A, and REG1B exhibited elevated expression in goblet cells, while PYY was highly expressed in coloncytes.

Discussion

The formation of kidney stones is one of the main urological complications of CD, which might complicate existing disorder.[29] Inflammatory condition and metabolic abnormalities in CD may contribute to the development of kidney stones.[11] Although KSD usually occur many years after CD has been diagnosed, the symptoms of KSD may precede or overshadow gastrointestinal symptoms, and the possibility of underlying CD needs to be considered.[30,31] However, comparatively few studies combined these two diseases and there are no specific markers for CD with KSD. Early identification and intervention of CD and KSD have important clinical significance and may improve the prognosis of the disease. Consequently, this research offers a comprehensive examination of CD and KSD employing bioinformatics techniques and machine learning algorithms.

https://doi.org/10.2147/JIR.S502513

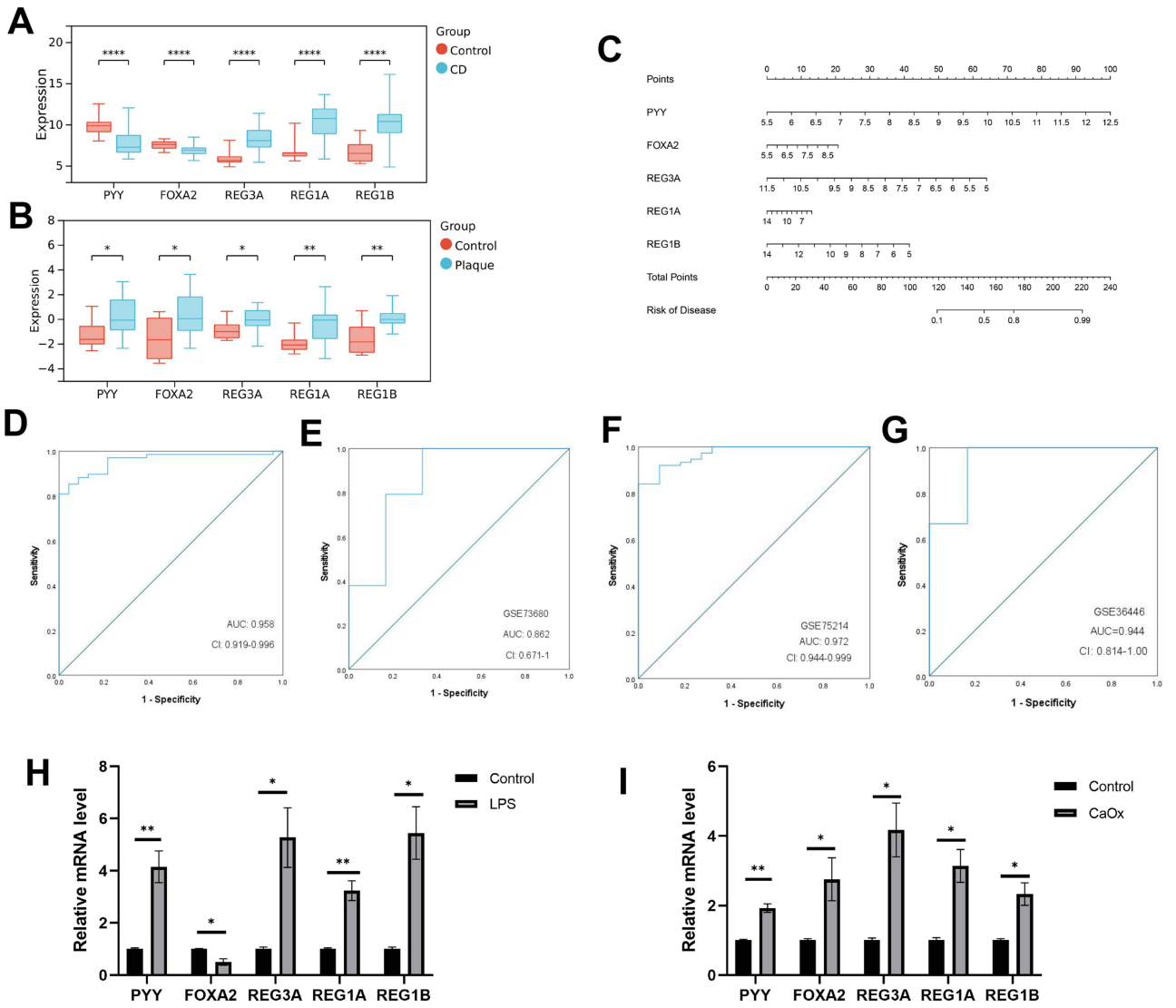

Figure 8 The nomogram construction and diagnostic value validation. (**A**) The expression of the HGs in the combined CD dataset. (**B**) The expression of the HGs in the KSD dataset (GSE73680). (**C**) The nomogram was constructed based on the five genes. (**D**) The ROC curves of the model in the integrated CD dataset. (**E**) The ROC curves of the model in the KSD dataset (GSE73680). (**F** and **G**) ROC curve analysis of the model in the CD (GSE75214) and KSD (GSE36446) validation set. AUC, area under the curve. (**H**) RT-qPCR results of 5 key genes in Caco-2 cell samples. (**I**) qRT-PCR results of 5 key genes in HK-2 cell samples. *p< 0.05, **p < 0.01, ****p < 0.0001.
Abbreviations: CD, Crohn's disease; KSD, Kidney stone disease; ROC, receiver operating characteristic.

Firstly, this investigation employed the LIMMA approach to pinpoint KSD-related key genes based on GSE73680 dataset and found 5584 DEGs through differential analysis. WGCNA is increasingly being employed to elucidate the association patterns between genes across microarray specimens and have been successfully applied in biology.[19] The co-expression network was established using WGCNA, and clusters exhibiting strong correlation with KSD were detected, and the darkgreen module was selected as the key module. Following obtaining the intersection of DEGs and genes in darkgreen module, the function and pathway enrichment analysis was performed and found metabolic pathways and chronic inflammatory response were enriched, consisting with the pathogenesis of KSD. Next, three microarray data sets from CD were merged and normalized for LIMMA analysis. After the intersection of the DEGs obtained by LIMMA analysis in CD and the darkgreen module genes in KSD, a set of 36 genes emerged, which were mainly enriched in the pathways of neuroactive ligand-receptor interaction and cAMP signaling. More importantly, we discovered five common HGs of CD and KSD on the basis of machine learning, including PYY, FOXA2, REG3A, REG1A and REG1B. In order to assess the value of the diagnostic model, we also built a nomogram as well as ROC curves. We then tested the model

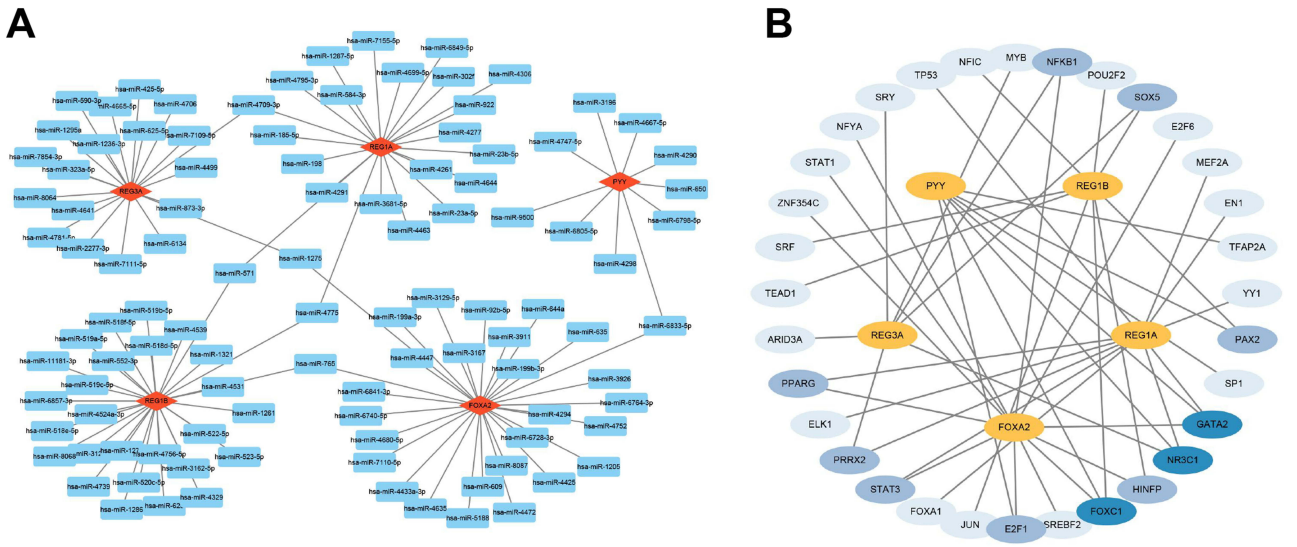

Figure 9 Prediction of potential miRNAs and TF-mRNA network of 5 hGs. (**A**) An Interaction network of five HGs and potential miRNAs-targeted where red diamonds represent genes, the blue rectangles represent predicted miRNAs. (**B**) Diagram of TF-mRNA regulatory network where the blue ellipses represent TF, the yellow ellipses represent genes.
Abbreviation: TF, transcription factor.

using two external datasets, which further confirmed its diagnostic value. In cell models of CD and KSD, these five genes also exert significant expression.

Peptide YY (PYY) is a 36 amino acid hormone and widely distributed in the digestive tract, involved in the regulation of energy homeostasis and metabolism.[32] According to research findings, patients with impaired kidney function had elevated PYY levels, which raises the possibility of a regulatory role in KSD.[33,34] FOXA2 is a member of human forkhead box (FOX) gene family.[35] The lineage-specifying transcription factor FOXA2 has been shown to be related with the development of congenital anomalies of the kidneys and urinary tract as well as bladder cancer.[36,37] Oxidative stress, mitochondrial dysfunction and cell death play pivotal roles in the pathogenesis of kidney stone formation.[38] Recent findings have shed light onto the possible role of FOXA2 in the mediation of oxidative stress and apoptosis in renal tubular cells.[39] REG3A, REG1A and REG1B are three members from the regenerating gene (REG) family, highly expressed in inflamed mucosa during IBD-related inflammation.[40] Accumulating evidence has clarified the potential roles of the REG genes in the development of IBD and gastrointestinal cancer, which display multifunctional biological activities, such as antiapoptotic, anti-inflammatory, antimicrobial effects.[41,42] However, the pathophysiological significance of the REGs in KSD is still elusive and remains to be explored.

MiRNAs are a large group of small non-protein coding RNAs that modulate gene expression and maintain general homeostasis via interacting with a wide range of target genes involved in various biological processes.[43] Studies have shown that alteration in miRNA expression may affect the progression of numerous disorders.[44–46] The increased level of microRNA 31 (MIR31) was found in colon tissues from patients with CD compared with controls, resulting in the reduction of the inflammatory response in colon epithelium of mice by suppressing expression of cytokine related genes.[47] In a rat model of hyperoxaluria, eight downregulated miRNAs and twenty upregulated miRNAs were observed to be differentially expressed in the renal tissues, and these miRNAs were predicted to serve important roles in insulin and mitogen-activated protein kinase (MAPK) signaling pathways as indicated by the pathway analysis.[48] Various digital tools and algorithms have been engineered to forecast miRNA binding locations on messenger RNAs.[49–51] In the present study, we developed the miRNA-mRNA regulatory network and found 4 crucial nodes of the network (miR-571, miR-1275, miR-4775, miR-765). Transcription factors (TFs) could specifically bind to the DNA sequences of a variety of genes and control the transcription of genes.[52] TF-mRNA network was built and it is found that FOXC1, NR3C1, and GATA2 are closely related to the HGs. Nevertheless, research on these miRNAs and TFs in CD and KSD that we have identified is limited and it still needs to be further investigated.

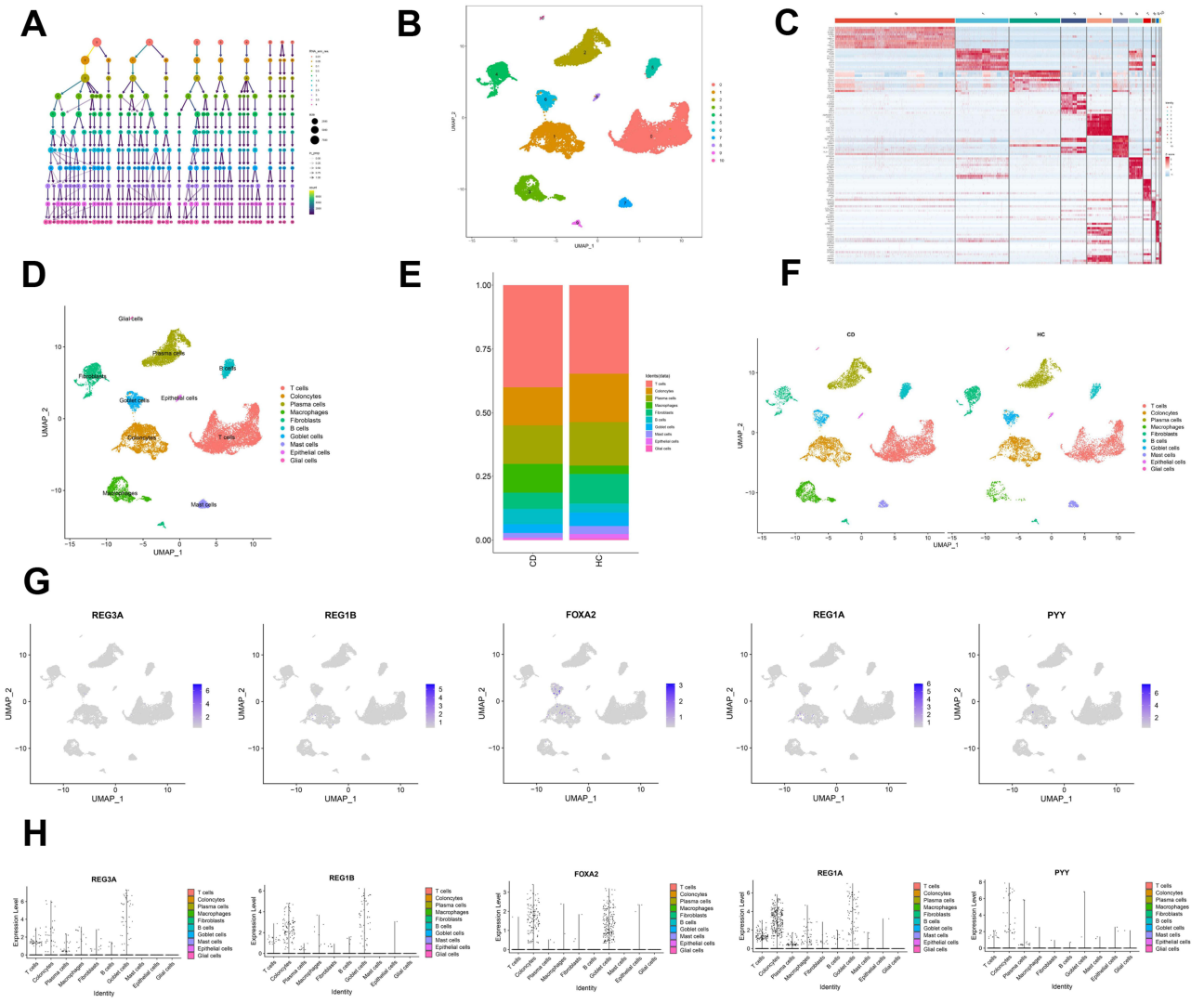

Figure 10 Single-cell transcriptome analysis depicting the cellular distribution and cell type mapping of HGs in the CD single-cell sequencing dataset (GSE214695). (**A**) Dendrogram of single-cell clustering. (**B**) UMAP downscaling plot showing 10 cell clusters obtained from single-cell sequencing analysis (resolution=0.1). (**C**) Heatmap of marker gene expression. (**D**) UMAP plots colored according to the major cell types. (**E**) Bar graph of the number and proportion of different cell types in the samples. (**F**) Split-plane presentation of the UMAP plots of the CD samples versus the control samples. (**G**) Coordinate mapping plot of core genes in different cells. (**H**) Expression violin plot of core genes in different cells.

Abbreviations: CD, Crohn's disease; UMAP, uniform manifold approximation and projection.

Finally, a single-cell sequencing data set from CD samples was retrieved in order to perform single cell annotation analysis. The single-cell RNA sequencing technologies have made pivotal advancements in single cell research, such as the physiological heterogeneity or lineage information in individual cells, which paves the way to new understandings in cellular processes and molecular mechanisms.[53] Our results indicated that the five HGs were expressed to varying degrees in the ten distinct types of annotated cells. The HGs were substantially expressed in goblet cells except PYY. In the gut, goblet cells have various functions, including mucus and mucin secretion, influencing the immune system, and interactions with intestinal microbiota, which is closely related to CD etiology.[54] This study offers a reliable direction for the in-depth investigation of the goblet cells not merely focused on various intestinal epithelial cells and may be helpful in understanding the mechanism of CD and KSD.

In an effort to elucidate the molecular mechanisms underlying CD and KSD, our work discovers and identifies common DEGs and HGs for the first time, as well as examines viable regulatory factors. Nevertheless, it is important to recognize a few limitations. First, the available microarray data are limited, especially data from KSD samples, and may

not fully represent patients with KSD. Furthermore, little is known about the precise molecular mechanisms in which TFs, miRNAs, and HGs influence these illnesses. In order to verify our results and investigate possible processes, more experiments in vitro or in vivo are required.

Conclusion

Based on bioinformatics analysis and machine learning, we pinpointed five HGs (PYY, FOXA2, REG3A, REG1A, and REG1B). The miRNA-mRNA and TF-mRNA networks were also established and the results revealed central miRNAs and TFs. Subsequent single-cell sequencing analyse provide deeper insight at the cellular level. The screened genes and regulatory molecules in this work may have the potential to serve as diagnostic and therapeutic targets for CD and KSD and facilitate the exploration of molecular mechanisms in the future.

Data Sharing Statement

The data that support the findings of this study are available on request.

Ethics Approval and Consent to Participate

The human sample data involved in this study followed the national and institutional ethical guidelines and this study was approved by the Ethics Committee of the First Affiliated Hospital of Anhui Medical University (PJ2022-07-27).

Acknowledgments

We express our gratitude to the team at the GEO for allowing us to utilize the data.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research was conducted without receiving any dedicated funding from either public, commercial, or not-for-profit funding agencies.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Zhang YZ, Li YY. Inflammatory bowel disease: pathogenesis. *World J Gastroenterol*. 2014;20(1):91–99.
2. Weidner J, Glauche I, Manuwald U, et al. Correlation of socioeconomic and environmental factors with incidence of Crohn disease in children and adolescents: systematic review and meta-regression. *JMIR Public Health Surveill*. 2024;10:e48682.
3. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. *Lancet*. 2017;389(10080):1741–1755. doi:10.1016/S0140-6736(16)31711-1
4. Veauthier B, Hornecker JR. Crohn's disease: diagnosis and management. *Am Fam Physician*. 2018;98(11):661–669.
5. Gajendran M, Loganathan P, Catinella AP, Hashash JG. A comprehensive review and update on Crohn's disease. *Dis Mon*. 2018;64(2):20–57. doi:10.1016/j.disamonth.2017.07.001
6. Peerapen P, Thongboonkerd V. Kidney Stone Prevention. *Adv Nutr*. 2023;14(3):555–569. doi:10.1016/j.advnut.2023.03.002
7. Thongprayoon C, Krambeck AE, Rule AD. Determining the true burden of kidney stone disease. *Nat Rev Nephrol*. 2020;16(12):736–746. doi:10.1038/s41581-020-0320-7
8. Wang Z, Zhang Y, Zhang J, Deng Q, Liang H. Recent advances on the mechanisms of kidney stone formation. *Int J mol Med*. 2021;48(2):149. doi:10.3892/ijmm.2021.4982
9. Coe FL, Worcester EM, Evan AP. Idiopathic hypercalciuria and formation of calcium renal stones. *Nat Rev Nephrol*. 2016;12(9):519–533. doi:10.1038/nrneph.2016.101
10. Herbert J, Teeter E, Burstiner LS, et al. Urinary manifestations in African American and Caucasian inflammatory bowel disease patients: a retrospective cohort study. *BMC Urol*. 2022;22(1):1. doi:10.1186/s12894-021-00951-z

11. McConnell N, Campbell S, Gillanders I, et al. Risk factors for developing renal stones in inflammatory bowel disease. *BJU Int*. 2002;89(9):835–841. doi:10.1046/j.1464-410X.2002.02739.x
12. Oikonomou K, Kapsoritakis A, Eleftheriadis T, Stefanidis I, Potamianos S. Renal manifestations and complications of inflammatory bowel disease. *Inflamm Bowel Dis*. 2011;17(4):1034–1045. doi:10.1002/ibd.21468
13. van Loo ES, Dijkstra G, Ploeg RJ, Nieuwenhuijs VB. Prevention of postoperative recurrence of Crohn's disease. *J Crohns Colitis*. 2012;6(6):637–646. doi:10.1016/j.crohns.2011.12.006
14. Nightingale J. Small Bowel and Nutrition Committee of the British Society of Gastroenterology. Guidelines for management of patients with a short bowel. *Gut*. 2006;55(Suppl 4):iv1–iv12. doi:10.1136/gut.2006.091108
15. Knauf F, Brewer JR, Flavell RA. Immunity, microbiota and kidney disease. *Nat Rev Nephrol*. 2019;15(5):263–274. doi:10.1038/s41581-019-0118-7
16. Yang Y, Ludvigsson JF, Olén O, Sjölander A, Carrero JJ. Absolute and relative risks of kidney and urological complications in patients with inflammatory bowel disease. *Am J Gastroenterol*. 2024;119(1):138–146. doi:10.14309/ajg.0000000000002473
17. Zhang H, Huang Y, Zhang J, Su H, Ge C. Causal effects of inflammatory bowel diseases on the risk of kidney stone disease: a two-sample bidirectional Mendelian randomization. *BMC Urol*. 2023;23(1):162. doi:10.1186/s12894-023-01332-4
18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–127. doi:10.1093/biostatistics/kxj037
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559. doi:10.1186/1471-2105-9-559
20. Chen L, Zhang YH, Lu G, Huang T, Cai YD. Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. *Artif Intell Med*. 2017;76:27–36. doi:10.1016/j.artmed.2017.02.001
21. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605–D612. doi:10.1093/nar/gkaa1074
22. Lee YW, Choi JW, Shin EH. Machine learning model for predicting malaria using clinical information. *Comput Biol Med*. 2021;129:104151. doi:10.1016/j.compbiomed.2020.104151
23. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery [published correction appears in Lancet Oncol. 2019 Jun;20(6):293. doi: 10.1016/S1470-2045(19)30294-3]. *Lancet Oncol*. 2019;20(5):e262–e273. doi:10.1016/S1470-2045(19)30149-4
24. Pan X, Jin X, Wang J, Hu Q, Dai B. Placenta inflammation is closely associated with gestational diabetes mellitus. *Am J Transl Res*. 2021;13(5):4068–4079.
25. Garrido-Trigo A, Corraliza AM, Veny M, et al. Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. *Nat Commun*. 2023;14(1):4506. doi:10.1038/s41467-023-40156-6
26. Liu Q, Wang D, Yang X, et al. The mechanosensitive ion channel PIEZO1 in intestinal epithelial cells mediates inflammation through the NOD-like receptor 3 pathway in Crohn's disease. *Inflamm Bowel Dis*. 2023;29(1):103–115. doi:10.1093/ibd/izac152
27. Xie J, Ye Z, Li L, et al. Ferrostatin-1 alleviates oxalate-induced renal tubular epithelial cell injury, fibrosis and calcium oxalate stone formation by inhibiting ferroptosis. *mol Med Rep*. 2022;26(2):256. doi:10.3892/mmr.2022.12772
28. Zhu J, Wu Y, Ge X, Chen X, Mei Q. Discovery and validation of ferroptosis-associated genes of ulcerative colitis. *J Inflamm Res*. 2024;17:4467–4482. doi:10.2147/JIR.S463042
29. Ben-Ami H, Ginesin Y, Behar DM, Fischer D, Edoute Y, Lavy A. Diagnosis and treatment of urinary tract complications in Crohn's disease: an experience over 15 years. *Can J Gastroenterol*. 2002;16(4):225–229. doi:10.1155/2002/204614
30. Joy HM, Fairhurst JJ, Beattie RM. Renal calculus at presentation in a child with Crohn's disease. *Pediatr Radiol*. 2003;33(4):250–252. doi:10.1007/s00247-002-0819-z
31. Manganiotis AN, Banner MP, Malkowicz SB. Urologic complications of Crohn's disease. *Surg Clin North Am*. 2001;81(1):197–x. doi:10.1016/S0039-6109(05)70281-4
32. Tseng WW, Liu CD. Peptide YY and cancer: current findings and potential clinical applications. *Peptides*. 2002;23(2):389–395. doi:10.1016/S0196-9781(01)00616-7
33. Pérez-Fontán M, Cordido F, Rodríguez-Carmona A, et al. Short-term regulation of peptide YY secretion by a mixed meal or peritoneal glucose-based dialysate in patients with chronic renal failure. *Nephrol Dial Transplant*. 2008;23(11):3696–3703. doi:10.1093/ndt/gfn297
34. Haj-Yehia E, Mertens RW, Kahles F, et al. Peptide YY (PYY) is associated with cardiovascular risk in patients with acute myocardial infarction. *J Clin Med*. 2020;9(12):3952. doi:10.3390/jcm9123952
35. Katoh M, Katoh M. Human FOX gene family (Review). *Int J Oncol*. 2004;25(5):1495–1500.
36. Zheng B, Seltzsam S, Wang C, et al. Whole-exome sequencing identifies FOXL2, FOXA2 and FOXA3 as candidate genes for monogenic congenital anomalies of the kidneys and urinary tract. *Nephrol Dial Transplant*. 2022;37(10):1833–1843. doi:10.1093/ndt/gfab253
37. Yamashita H, Amponsa VO, Warrick JI, et al. On a FOX hunt: functions of FOX transcriptional regulators in bladder cancer. *Nat Rev Urol*. 2017;14(2):98–106. doi:10.1038/nrurol.2016.239
38. Yu L, Gan X, Bai Y, An R. CREB1 protects against the renal injury in a rat model of kidney stone disease and calcium oxalate monohydrate crystals-induced injury in NRK-52E cells. *Toxicol Appl Pharmacol*. 2021;413:115394. doi:10.1016/j.taap.2021.115394
39. Ye G, Hu ML, Xiao L. Forkhead box A2-mediated lncRNA SOX2OT up-regulation alleviates oxidative stress and apoptosis of renal tubular epithelial cells by promoting SIRT1 expression in diabetic nephropathy. *Nephrology*. 2023;28(3):196–207. doi:10.1111/nep.14139
40. van Beelen Granlund A, Østvik AE, Ø B, Torp SH, Gustafsson BI, Sandvik AK. REG gene expression in inflamed and healthy colon mucosa explored by in situ hybridisation. *Cell Tissue Res*. 2013;352(3):639–646. doi:10.1007/s00441-013-1592-z
41. Sun C, Wang X, Hui Y, Fukui H, Wang B, Miwa H. The potential role of REG family proteins in inflammatory and inflammation-associated diseases of the gastrointestinal tract. *Int J mol Sci*. 2021;22(13):7196. doi:10.3390/ijms22137196
42. Lu J, Wang Z, Maimaiti M, Hui W, Abudourexiti A, Gao F. Identification of diagnostic signatures in ulcerative colitis patients via bioinformatic analysis integrated with machine learning. *Hum Cell*. 2022;35(1):179–188. doi:10.1007/s13577-021-00641-w
43. Diener C, Keller A, Meese E. Emerging concepts of miRNA therapeutics: from cells to clinic. *Trends Genet*. 2022;38(6):613–626. doi:10.1016/j.tig.2022.02.006
44. Mishra S, Yadav T, Rani V. Exploring miRNA based approaches in cancer diagnostics and therapeutics. *Crit Rev Oncol Hematol*. 2016;98:12–23. doi:10.1016/j.critrevonc.2015.10.003

45. Hill M, Tran N. miRNA:miRNA interactions: a novel mode of miRNA regulation and its effect on disease. *Adv Exp Med Biol*. 2022;1385:241–257.
46. Jay C, Nemunaitis J, Chen P, Fulgham P, Tong AW. miRNA profiling for diagnosis and prognosis of human cancer. *DNA Cell Biol*. 2007;26 (5):293–300. doi:10.1089/dna.2006.0554
47. Tian Y, Xu J, Li Y, et al. MicroRNA-31 reduces inflammatory signaling and promotes regeneration in colon epithelium, and delivery of mimics in microspheres reduces colitis in mice. *Gastroenterology*. 2019;156(8):2281–2296.e6. doi:10.1053/j.gastro.2019.02.023
48. Liu Z, Jiang H, Yang J, et al. Analysis of altered microRNA expression profiles in the kidney tissues of ethylene glycol-induced hyperoxaluric rats. *mol Med Rep*. 2016;14(5):4650–4658. doi:10.3892/mmr.2016.5833
49. Kumar A, Wong AK, Tizard ML, Moore RJ, Lefèvre C. miRNA_Targets: a database for miRNA target predictions in coding and non-coding regions of mRNAs. *Genomics*. 2012;100(6):352–356. doi:10.1016/j.ygeno.2012.08.006
50. Rennie W, Kanoria S, Liu C, et al. STarMirDB: a database of microRNA binding sites. *RNA Biol*. 2016;13(6):554–560. doi:10.1080/15476286.2016.1182279
51. Wang P, Ning S, Wang Q, et al. mirTarPri: improved prioritization of microRNA targets through incorporation of functional genomics data. *PLoS One*. 2013;8(1):e53685. doi:10.1371/journal.pone.0053685
52. Lambert SA, Jolma A, Campitelli LF, et al. The human transcription factors [published correction appears in Cell 2018 Oct;175(2):598-599. doi: 10.1016/j.cell.2018.09.045]. *Cell*. 2018;172(4):650–665. doi:10.1016/j.cell.2018.01.029
53. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;20(5):257–272. doi:10.1038/s41576-019-0093-7
54. Wang Z, Shen J. The role of goblet cells in Crohn' s disease. *Cell Biosci*. 2024;14(1):43. doi:10.1186/s13578-024-01220-w