Journal of Translational Medicine

RESEARCH

Open Access

# Identification of key genes as diagnostic biomarkers for IBD using bioinformatics and machine learning

Tianhao Li[1,2†], Haoren Jing[1,2,3†], Xinyu Gao[1,2], Te Zhang[1,2], Haitao Yao[1], Xipeng Zhang[1,2,3] and Mingqing Zhang[1,2,3*]

## Abstract

**Background** The pathogenesis of inflammatory bowel disease (IBD) involves complex molecular mechanisms, and achieving clinical remission remains challenging. This study aims to identify IBD-potential biomarkers, analyze their correlation with immune cell infiltration, and identify genes that have a causal relationship with IBD.

**Methods** RNA-seq datasets for IBD were retrieved from GEO, stratified into discovery (GSE75214), validation (GSE36807), and independent testing cohorts (GSE179285, GSE47908). Through comparative expression profiling of the discovery cohort, IBD-associated differentially expressed genes (DEGs) were detected. Core candidate genes were subsequently prioritized using protein-protein interaction network analysis, further refined through machine learning approaches (Random Forest/Support Vector Machines). Immune cell abundance quantification and statistical correlation analyses with IBD-associated transcripts were conducted via the CIBERSORTx deconvolution algorithm. To complement these findings, blood expression quantitative trait loci (eQTL) data from GTExv8.ALL.Whole_Blood were integrated with IBD genome-wide association statistics from the FinnGen consortium. This multi-omics integration framework employed: (1) Bayesian colocalization to assess shared causal variants, (2) HEIDI heterogeneity testing, and (3) summary Mendelian randomization (SMR) for causal inference validation.

**Results** Three genes, IRF1, GBP5, and PARP9, demonstrated significant IBD-promoting effects. The characteristic biomarkers of IBD were associated with immune cell infiltration. SMR analysis based on eQTL data showed that IRF1 was significantly associated with the risk of IBD. Moreover, IRF1 passed the HEIDI test of $> 0.05$ on gene expression, and IRF1 demonstrated the ability to promote the development of IBD.

**Conclusions** This integrative analysis identifies IRF1, GBP5, and PARP9 as potential genes associated with IBD pathogenesis. SMR analysis based on eQTL data revealed that IRF1 was significantly associated with the risk of IBD. The relationship between IRF1 and IBD risk highlights its potential as both a therapeutic target and diagnostic biomarker.

**Keywords** Inflammatory bowel disease, Diagnostic biomarker, Pathogenic mechanism, Bioinformatics, Mendelian randomization

†Tianhao Li and Haoren Jing contributed equally to this work.

*Correspondence:
Mingqing Zhang
zmq@nankai.edu.cn

[1]Department of Colorectal Surgery, Tianjin Union Medical Center, The First Affiliated Hospital of Nankai University, Tianjin 300121, 2755-7131, China
[2]Tianjin Institute of Coloproctology, Tianjin, China
[3]Nankai University School of Medicine, Nankai University, Tianjin, China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

## Introduction

Inflammatory bowel disease (IBD) is characterized as a chronic condition marked by episodes of inflammation that may temporarily subside before recurring, primarily affecting the gastrointestinal tract. Over time, severe inflammation within the intestines can lead to serious complications, including ulcers, narrowing of the intestinal lumen, and even perforations. Notably, IBD can cause considerable damage to the digestive system, resulting in substantial tissue injury, reduced intestinal functionality, associated disability, and a state of systemic inflammation that may impact the individual's overall health [1]. Among the various factors that contributing to susceptibility to IBD, genetic predisposition plays a central role [2].

The causative genes responsible for IBD remain incompletely characterized, highlighting the necessity for a thorough investigation of crucial molecular signatures involved in its development, which may provide potential avenues for effective therapies [2]. Recent advancements in high-throughput sequencing methodologies have been utilized in genome-wide association studies (GWAS) to identify susceptibility regions associated with IBD [3]. However, the complex and multifaceted mechanisms that underlie IBD present substantial challenges in effectively translating these identified risk loci into practical clinical treatments. The development of quantitative trait loci (QTL) data and new statistical methodologies has promoted the expression quantitative trait loci (eQTL) analysis, which supports the rational integration of multidimensional data [4, 5]. Mendelian randomization is based on the essential concept of randomly assigning parental alleles to their offspring, which is in line with the three key assumptions underlying causal reasoning [6]. This approach facilitates the identification of geno-phenotypic causal relationships at the transcriptome level.

Multiple transcriptional datasets were acquired from the Gene Expression Omnibus (GEO) database, sourced from both IBD patients and control subjects for this study. This included including four expression profiles: GSE75214, GSE47906, GSE179285 and GSE36807. The primary analysis was conducted on GSE75214, while the test datasets comprised GSE47906 and GSE179285, which were subsequently processed separately and verified using GSE36807.Initially, an analysis was conducted to identify differentially expressed genes (DEGs) that distinguish between individuals with IBD and healthy controls. Following this, a weighted gene co-expression network analysis (WGCNA) was conducted based on these DEGs, allowing for the identification and filtering of key gene modules that play a significant role in the pathology of the disease. After establishing the relevant gene modules, the key genes identified through the WGCNA and DEGs were further analyzed using GO and KEGG enrichment analyses. Subsequently, the key genes were subjected to protein-protein interaction (PPI) network selection and machine learning-based integration workflows. The generation of hub genes, which exhibited excellent diagnostic results in both training and testing cohorts, was further evaluated alongside their expression and associated patterns with IBD. Combined with transcription factor (TF) prediction, pathways involved in the co-regulation of hub genes in IBD were systematically identified. The infiltration level of immune cells in IBD were quantified using the CIBERSORT method to screen for key immune cell types showing a high correlation with hub genes. Finally, SMR and Bayesian colocalization analyses were conducted based on large-scale GWAS, and eQTL analyses were performed to explore the causal effects of hub genes on IBD risk.

## Methods

### Study design

Based on the GEO database, this study identified the intersection genes related to IBD through WGCNA and DEG analysis. Subsequently, hub genes were further screened using PPI analysis and machine learning. Meanwhile, eQTL data and GWAS data related to IBD were integrated to verify the causal relationship between hub genes and IBD. The workflow of the analysis is shown in Fig. 1.

### Data collection and analysis

Three distinct gene expression profiles (GSE75214, GSE179285, and GSE47908) related to patients with IBD were obtained from the GEO database. GSE75214 was designated as the training group for this study, comprising a sample of 133 IBD patients and 22 non-IBD controls [7]. After removing the batch effects from GSE179285 and GSE47908, the datasets were merged into the test group using the R package "sva" (version 3.50.0). Subsequently, the gene expression profile was subjected to logarithmic transformation, and the data were normalized using the "normalize Between Arrays" function from the R package "limma" (version 3.58.1) [8].

### Weighted gene co-expression network analysis of DEGs

To build a WGCNA for DEGs, we utilized the R package "WGCNA" (version 1.73). The objective was to identify gene modules significantly linked to IBD [9]. Initially, we explored soft threshold powers (β) to establish a scale-free network topology. Subsequently, we generated a weighted adjacency matrix, which was then converted into a topological overlap matrix. For hierarchical clustering, the dissimilarity measure based on TOM was calculated. We applied the dynamic tree cutting method to identify distinct modules by clustering genes based on

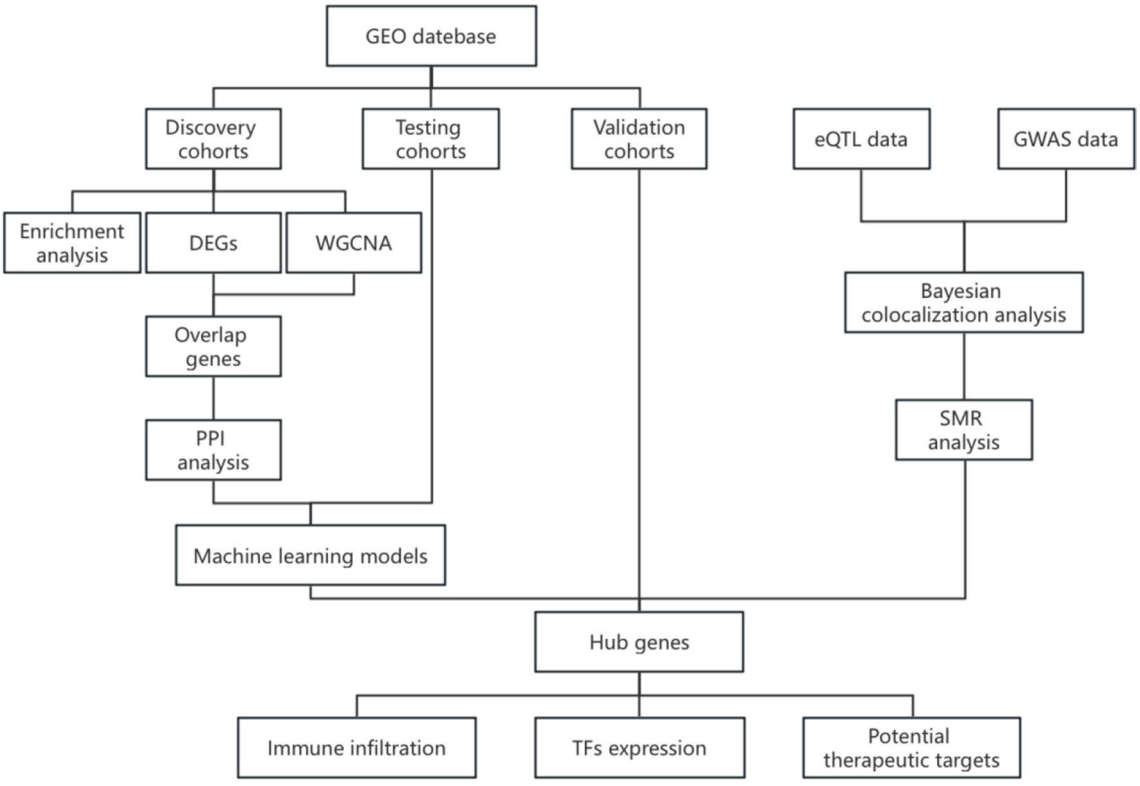



**Fig. 1** Flowchart of the study

their similarity. Genes showing a strong association with IBD were selected for further analysis.

#### GO and KEGG enrichment analysis of key gene

The genes that intersect between the significant co-expressed genes identified through WGCNA and DEG were analyzed for functional and pathway annotations. These annotations related to human diseases, biological functions, and signaling pathways may reveal connections between the overlapping genes associated with IBD [10], thereby clarifying the roles of these genes in IBD. Consequently, a comprehensive enrichment analysis for GO and KEGG was performed on these overlapping genes using the R packages "clusterProfiler" and "DOSE" (version 3.28.2) [11, 12].

#### Construction of protein-protein interaction networks

Using the Search Tool for the Retrieval of Interacting Genes (STRING) database, a network depicting protein-protein interactions (PPI) was formulated. This network specifically focused on the intersection of DEGs and those genes identified through WGCNA. The visualization of these intersections was achieved using Cytoscape software (version 3.9.1) [13, 14]. To further identify the top 20 genes exhibiting the highest binding coefficients within the PPI network, all five sequencing algorithms available in the Cytoscape plug-in cytoHubba were employed. Additionally, the DEGREE algorithm was applied for further filtering of genes within the PPI network [15].

#### Select features through machine learning-based ensembles

The key genes identified through the application of five PPI network analysis algorithms were subsequently integrated into diverse machine learning frameworks [16]. A comprehensive panel of twelve distinct machine learning algorithms was implemented, spanning regularization methods (Lasso, Ridge, ElasticNet), generalized linear model approaches (Stepglm, glmBoost, plsRglm), ensemble learning techniques (Random Forest, GBM, XGBoost), pattern recognition models (SVM, LDA, Naive Bayes), and gradient boosting frameworks. This multi-paradigm design encompassed both linear regression frameworks and advanced nonlinear learning structures to ensure robust analytical coverage. The computational framework incorporated a dual-algorithm architecture: (1) feature ranking through recursive feature elimination for predictor screening, and (2) predictive model construction via stacked generalization. Implemented under stratified 10-fold cross-validation schema, this ensemble strategy generated 113 unique algorithmic permutations through systematic hyperparameter optimization. Following this preparatory step, our comprehensive machine

learning framework commenced with four algorithms capable of performing feature selection: Lasso, Random Forest, Stepglm, and glmBoost. These algorithms were instrumental in refining gene selection, thereby narrowing the focus to a more manageable range. After this initial filtering process, eight additional algorithms were utilized to fit the prediction model based on the gene set selected by the earlier four algorithms. This rigorous approach culminated in testing a total of 113 model combinations, each of which underwent hyperparameter tuning and fitting under the stringent conditions of a 10-fold cross-validation framework. Subsequently, all generated model combinations were evaluated using both training and testing cohorts. The performance of each model was assessed by calculating the AUC scores, which serve as a critical metric for determining model efficacy. The model that demonstrated the highest mean AUC score across the training and testing cohorts was deemed optimal. From this best-performing model, the trait genes were extracted, providing valuable insights into the genetic factors influencing the studied traits.

#### Diagnostic power, expression levels, and associated patterns of hub genes

To examine the diagnostic efficacy of crucial genes in IBD, the AUC score from the ROC analysis was evaluated using the R package "pROC" (version 1.18.5). This score assesses the classification accuracy between the IBD and control groups [17]. In both the training and validation cohorts, genes showing an AUC greater than 0.7 in differentiating IBD from controls were identified as significant genes that could serve as biomarkers for IBD diagnosis. Subsequently, the R package "glmnet" (version 4.1-8) was employed to develop a logistic regression model for assessing the combined diagnostic proficiency of the key genes in both the training and validation cohorts. Correlation analysis and visualization of the key genes in the validation cohort were conducted with the aid of the "PerformanceAnalytics" package (version 2.0.8).

#### Analysis of the expression level of transcription factors and their interaction with hub gene

To investigate the upstream mechanisms governing hub genes, an analysis of the TRRUST v2 database was conducted to identify potential transcription factors (TFs) interacting with these hub genes (www.grnpedia.org/trrust, accessed August 8, 2024) [18]. The 10 predicted TFs with the highest sentence rankings in the database were selected. It is possible that TFs upregulated in IBD actively influence these hub genes. Next, correlation analyses were performed for key TFs and pivotal genes.

#### CIBERSORT evaluated IBD immune cell tags and identified immune cell types that were highly correlated with the hub gene

Immune cell dynamics in IBD were interrogated through CIBERSORT-based deconvolution, which employs linear support vector regression to systematically resolve cellular heterogeneity from complex expression matrices. CIBERSORT provides expression data for 22 types of immune infiltrating cells in CIBERSORTx (https://cibersortx.stanford.edu/), including the different cell types and functions of immune cells. Subsequently, the correlation patterns between hub genes and key immune cell types were assessed, revealing significant upregulation in IBD compared to controls.

#### Bayesian colocalization analysis

Bayesian colocalization analysis evaluates the consistency of causal variation between exposures and outcomes, helping to prevent false signals in MR that may arise from cascade imbalance or pleiotropy [19]. Consequently, the analysis was conducted using the 'COLOC' software package. Five competing causal hypotheses were formulated, categorized by their distinct causal architectures: H0 (null association), H1 (exposure-specific effects), H2 (outcome-restricted linkage), H3 (pleiotropic association with variant heterogeneity), and H4 (shared causal variants). Statistical significance was established when the posterior probability hypothesis for H4 (PPH4) exceeded the 0.8 threshold, thereby confirming colocalization through common pathogenic mechanisms [19].

#### SMR analysis

The SMR platform was deployed on Linux systems to execute cis-regional analyses, incorporating an evaluation of instrumental variable heterogeneity using the HEIDI test [20]. This SNP-focused MR methodology interrogates genetic instruments through single nucleotide polymorphisms. For causal inference, expression quantitative trait loci were modeled as exposures, complemented by genome-wide association study outcome data extracted from the FinnGen biobank. This analysis primarily assesses whether there exists a causal or pleiotropic relationship between gene expression and phenotype, indicating shared causal variants. The analytical pipeline retained default SMR configurations, implementing an FDR-controlled multiplicity adjustment ($\alpha < 0.05$) concurrent with HEIDI-based pleiotropy screening (empirical $p < 0.05$). This dual-threshold framework required concordant evidence thresholds: genome-wide significance ($\alpha < 0.05$) for polygenic signal detection and mechanistic filtering ($p < 0.05$) to exclude mediation chain artifacts from the final causal inferences [20].

### Drug-gene interactions and potential therapeutic targets

Evaluating drug-gene interactions is critical to determine whether a target gene can be used as an actual drug target. In this study, we analysis based on DSigDB database [21]. The cross-module retrieval of relevant genes through an interactive interface that links drugs and other compounds to target genes enables prediction of potential drug candidates, thereby assessing the therapeutic potential of these target genes [2].

### Statistical analysis

Statistical evaluations were conducted with R programming (version 4.3.3). To assess hub gene expression patterns, Pearson correlation coefficients were calculated. Furthermore, Spearman's rank-order method was applied to evaluate potential associations between immune cell infiltration profiles and the identified hub genes' transcriptional activity. In this research, $p < 0.05$ was deemed statistically significant.

## Results

### A total of 1816 differential genes associated with IBD were identified

A total of 1816 DEGs were identified with $\log_2 FC > 0.585$ and FDR < 0.05 between the IBD group and the Control group (Fig. 2).

### WGCNA identified key modules on DEGs that are closely related to IBD

As shown in Fig. 3, MElightyellow, MEblack, and MEgrey60 were significantly positively correlated with IBD ($R > 0.5$, $p < 0.05$). Moreover, the distribution of gene significance (GS) was significantly different between modules ($p < 0.05$). Thereafter, an intra-module analysis was performed, revealing a significant positive correlation between module members (MM) and gene significance (GS) in MElightyellow ($R = 0.7$, $p < 0.05$) (Fig. 3). Finally, the genes in the MElightyellow module ($n = 1094$) were treated as IBD-related genes for further analysis.

### Substantial overlap between key modules indicates a close relationship between the biological functions and pathways associated with IBD

Comparing the WGCNA and DEG significantly positive gene modules, the intersection between these identified modules was obtained. A total of 77 genes were found in both the MElightyellow and DEG modules of WGCNA (Fig. 4(A)). Subsequently, the identified intersection genes were subjected to GO and KEGG enrichment analyses. The GO enrichment analysis suggested that 77 genes were highly correlated with biological membranes, such as the external side of plasma membrane, lysosomal membrane, lytic vacuum membrane, and vacuum membrane (Fig. 4(B)). The KEGG enrichment analysis found that 77 genes were predominantly related to influenza A and Epstein-Barr virus infections. Furthermore, the nod-like receptor signaling pathway and chemokine signaling pathway were identified as the primary molecular pathways implicated in these genes (Fig. 4(C)).

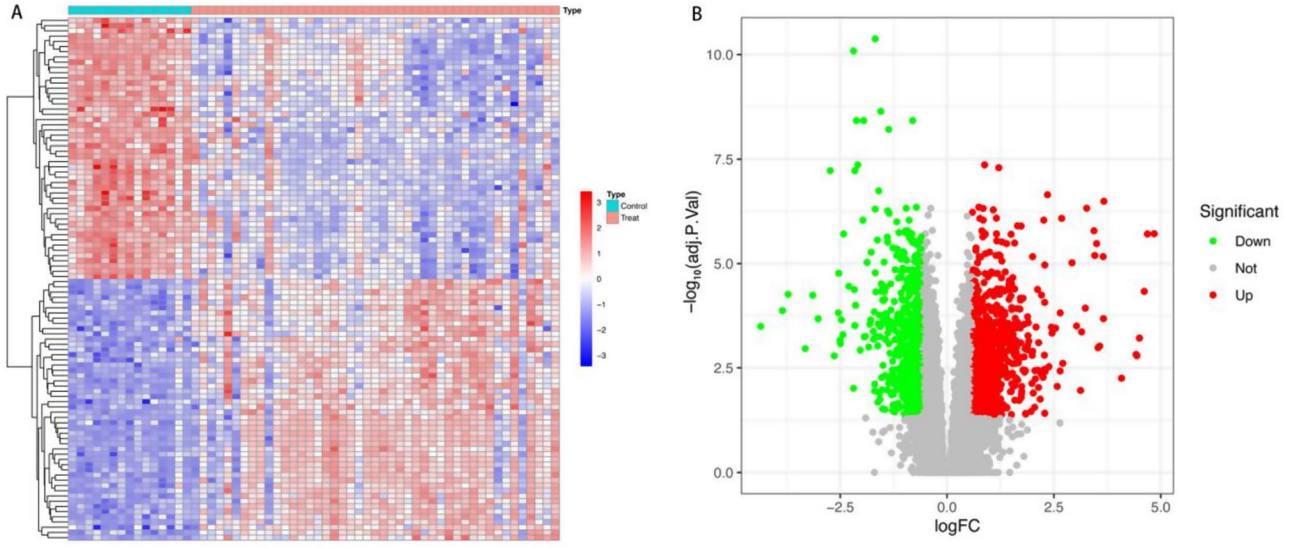



**Fig. 2** DEG analysis to identify IBD from a training cohort of $n = 155$ (133 IBD and 22 non-IBD controls). (**A**) DEGs volcano plot. (**B**) Cluster heat map of DEGs in the IBD group and the control group

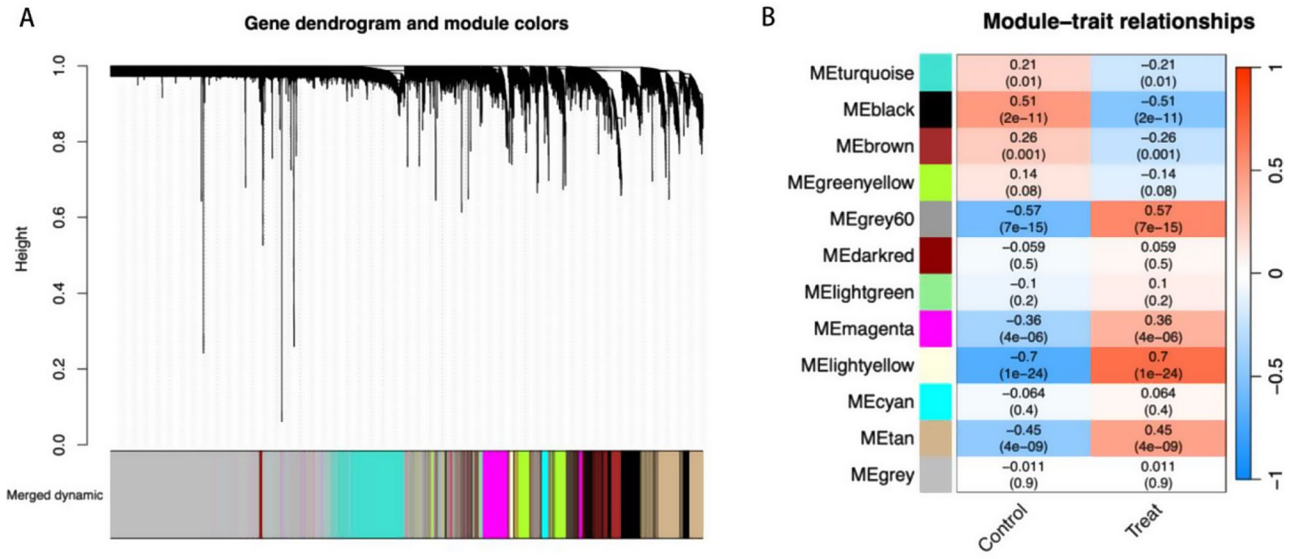



**Fig. 3** Differential analysis and WGCNA Results for display of IBD; (**A**) threshold for WGCNA analysis. (**B**) Module-trait relationships. Each cell contains the corresponding correlation and p-value

## Based on the integration and screening of PPI network analysis and machine learning, 6 characteristic genes of IBD were obtained

The Degree algorithm in cytoHubba was utilized, yielding an interactive network of genetically dominant PPIs and identifying the 15 most important key genes.

Machine learning-based integration of the expression profiles of 15 genes to filter key genes for IBD (Fig. 5(A)). In this research, a total of 113 predictive classification models were constructed and applied to the training and testing datasets utilizing a cross-validation framework. The integration algorithm of RF + NaiveBayes achieved the highest average AUC score of 0.855 in both the training and testing groups, thereby being identified as the most effective classification model. (Fig. 5(B)). Consequently, a set of 6 characteristic genes *STAT1*, *PARP9*, *IRF1*, *GBP5*, *OAS2*, and *TRIM22* were obtained from the RF + NaiveBayes algorithm.

### *IRF1*, *GBP5*, and *PARP9* are diagnostic biomarkers for IBD

To evaluate the diagnostic ability of the six characteristic genes, ROC curve analysis was performed in both the training and testing cohorts. The results showed that the AUC value for *IRF1*, GBP5, and *PARP9* in the training cohort were greater than 0.95 (Fig. 6(A)), indicating high sensitivity and specificity to distinguish between IBD and control. Therefore, *IRF1*, *GBP5*, and *PARP9* were identified as central genes of IBD and can serve as potential diagnostic biomarkers. In addition, a model was constructed based on the combination of the three central genes, yielding significantly improved diagnostic performance in both the training (AUC = 0.994) and testing cohort (AUC = 0.843, 0.728). The evaluation of the diagnostic accuracy, relevance, and expression patterns of *IRF1*, *GBP5*, and *PARP9* was conducted across three different datasets (GSE179285 and GSE47908 as test sets and GSE36807 as external validation set) (Table 1). Despite the variation in sample sources among these datasets, the strong diagnostic capabilities and increased expression levels of these genes were reaffirmed, particularly for *IRF1* (all AUC > 0.75). A significant pattern of upregulation of the three central genes in IBD was observed in external validation cohorts ($p < 0.001$) (Fig. 6(B)).

## CIITA, STAT3, RELA, STAT2 are transcription factors that actively target the characteristic genes *IRF1* and *STAT1*

Based on the TRRUST v2 database, 6 TFs targeting characteristic genes were recruited (STAT2, CREBBP, CIITA, STAT3, TP53, RELA) (Table 2). In the training and testing cohorts, a total of 4 TFs (CIITA, STAT3, RELA, STAT2) were increased in IBD and were considered key TFs in IBD progression. The above TFs were all related to *IRF1* and *STAT1*. Figure 7(A) further illustrates the correlation pattern of 2 characteristic genes and 4 key TFs in IBD, with STAT2 significantly correlating with *IRF1* and *STAT1*.

The regulatory network involving *IRF1* and *STAT1*-related TFs is shown in Fig. 7(B). The top three enrichment functions were cellular response to interferon-gamma, response to interferon-gamma, and response to type I interferon.

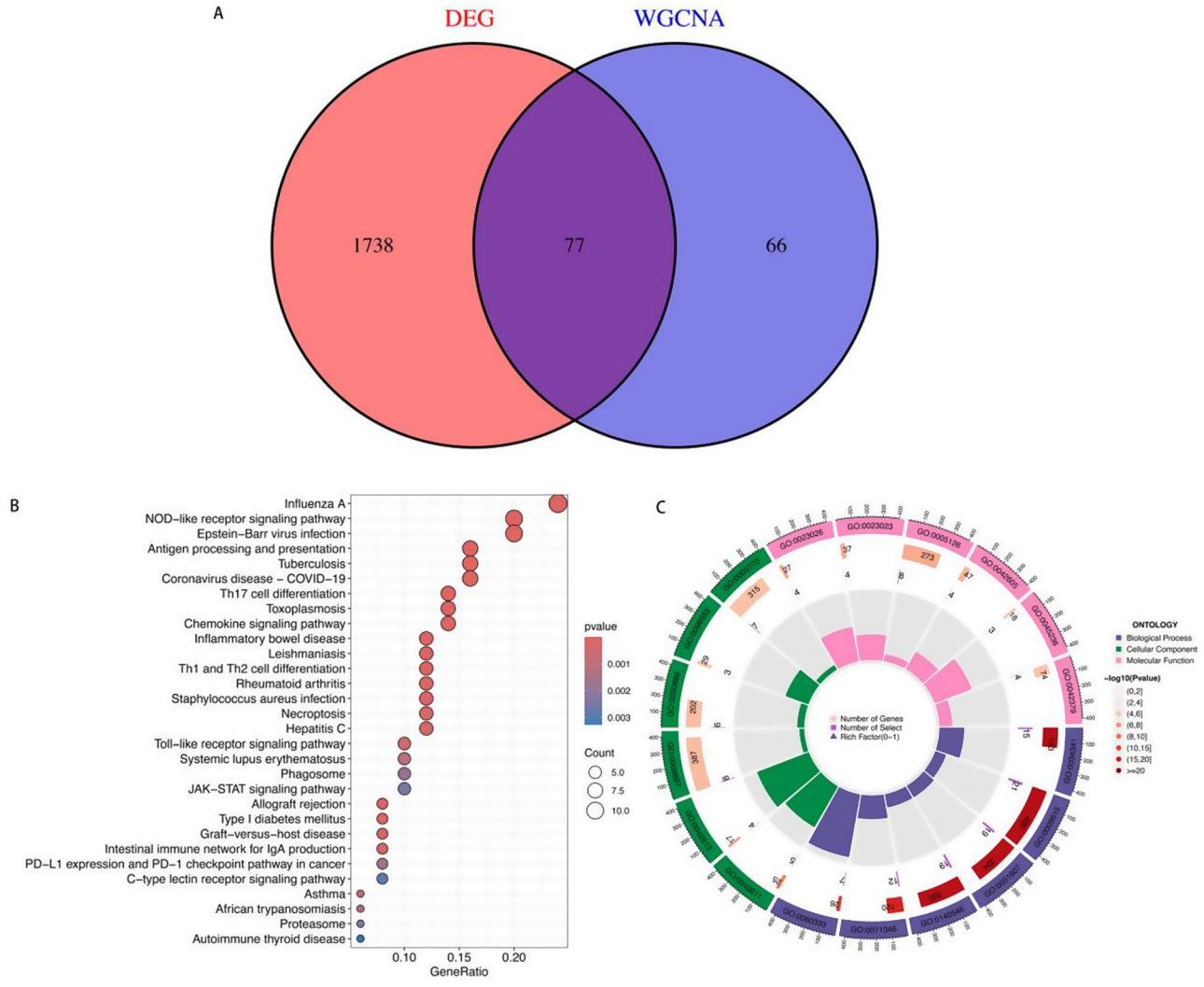



**Fig. 4** Overlap genes in WGCNA and DEG significantly positive gene modules. (**A**) The Venn diagram of genes among DEGs and co-expression modules. (**B**) GO enrichment analysis. (**C**) KEGG enrichment analysis

## Elevated macrophage M1 cell infiltration and decreased plasma cell infiltration are highly correlated with central genes

M1 macrophages demonstrated a notable increase in infiltration, along with a positive correlation to central genes in both the IBD training and validation cohorts. In addition, plasma cells, T cells CD8, and activated NK cells showed significantly downregulated infiltration and negative correlation with central genes in the IBD training cohort. Moreover, M2 macrophages, plasma cells, and resting mast cells showed significantly downregulated infiltration and negative correlation with central genes in the IBD validation cohort, indicating that M1 macrophages and plasma cells play a key role in immune-related pathways regulated by hub genes (Fig. 8).

## Elevated *IRF1* in IBD patients has strong diagnostic value and has a positive causal relationship with IBD risk

To investigate the potential of genes biomarkers associated with genetic variation, SMR analysis was performed based on eQTL data and large-scale GWAS data (Finngen database), followed by HEIDI (Supplementary Material SM1). After Bonferroni correction and colocalization analysis, 133 genes were found to be associated with IBD risk (Fig. 9(A)). Among them, *IRF1* was also the central gene for IBD. Ultimately, a positive correlation was found between *IRF1* and IBD risk ($P < 0.05$) (Fig. 9(B), Table 3). Therefore, a higher *IRF1* may correspond to an increased risk of IBD.

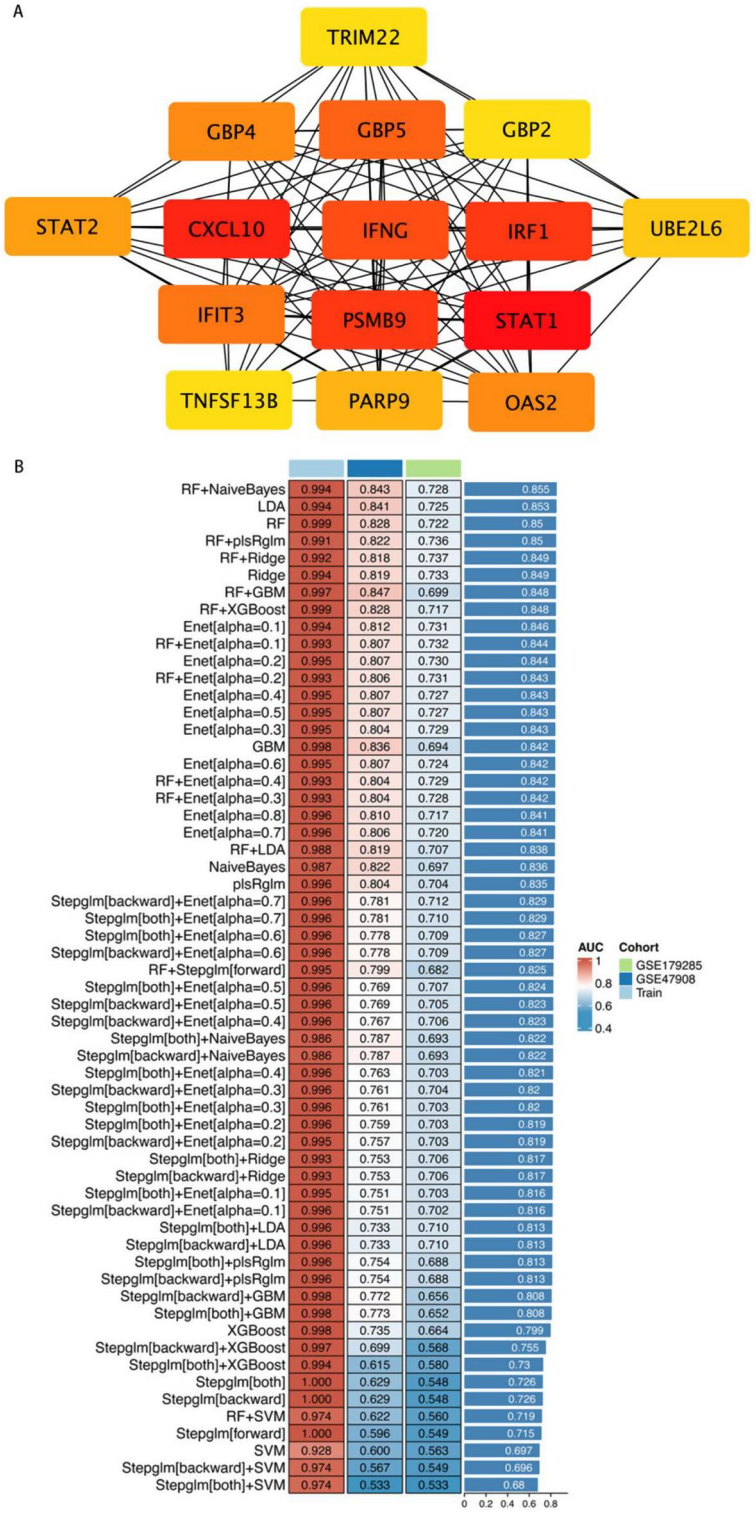



**Fig. 5** (**A**) The top 15 pivotal genes in the PPI network. (**B**) 113 machine learning algorithm combinations evaluated via 10-fold cross-validation

## Discussion

This study has made innovative breakthroughs in the integration of multi-dimensional methods and the analysis of immune regulatory networks. First, we integrated the systematic validation framework of bioinformatics, machine learning and Mendelian randomization (SMR, HEIDI test) to systematically verify the importance of IRF1, GBP5 and PARP9 in IBD from a multi-omics

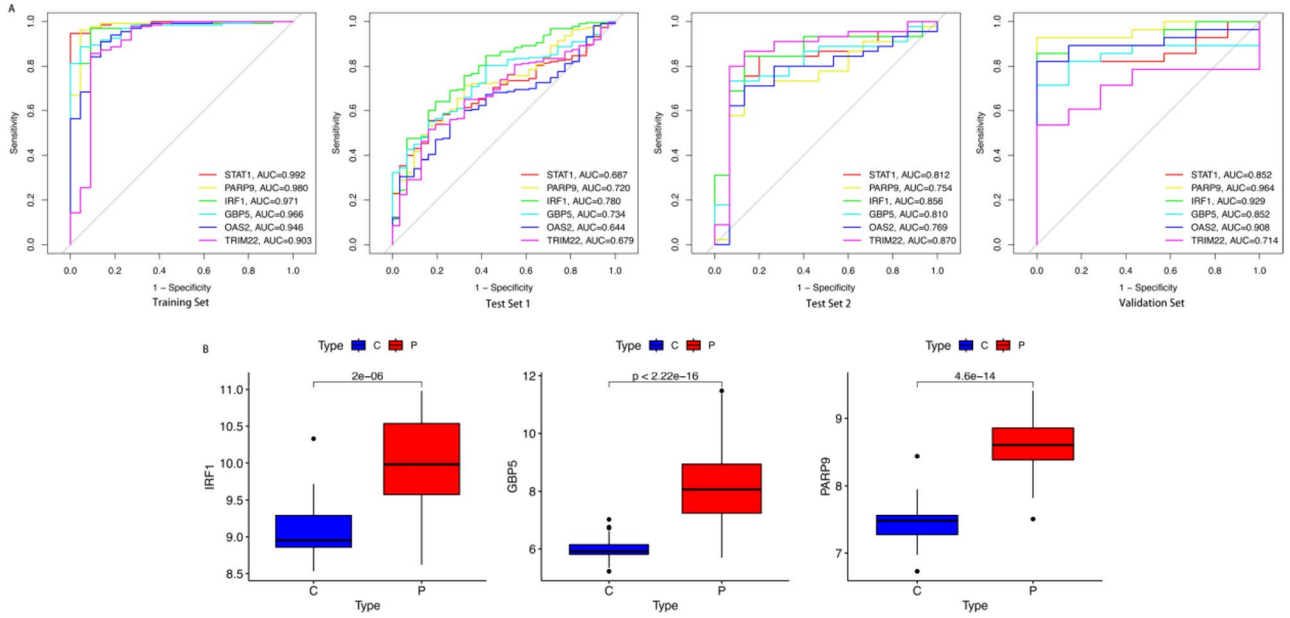



**Fig. 6** ROC curve and expression level of characteristic genes. (**A**) ROC curves of 6 characteristic genes in the training set, test sets and external validation set. (**B**) Expression levels of three hub genes in external validation sets

**Table 1** AUC of six characteristic genes in the training set, test sets, and external validation set

| Gene | Training | Test1 | Test2 | Average AUC | Validation |
|---|---|---|---|---|---|
| IRF1 | 0.97 | *0.78* | *0.856* | *2.606* | 0.929 |
| GBP5 | 0.966 | 0.734 | 0.81 | 2.51 | 0.852 |
| STAT1 | *0.992* | 0.687 | 0.812 | 2.491 | 0.852 |
| PARP9 | 0.98 | 0.72 | 0.754 | 2.454 | *0.964* |
| TRIM22 | 0.903 | 0.679 | 0.87 | 2.452 | 0.714 |
| OAS2 | 0.946 | 0.644 | 0.769 | 2.359 | 0.908 |

**Table 2** TFs targeting characteristic genes

| Key TF | Description | *P* value | FDR |
|---|---|---|---|
| STAT2 | signal transducer and activator of transcription 2, 113 kDa | 4.63E-06 | 2.78E-05 |
| CREBBP | CREB binding protein | 2.52E-05 | 7.56E-05 |
| CIITA | class II, major histocompatibility complex, transactivator | 3.90E-05 | 7.80E-05 |
| STAT3 | signal transducer and activator of transcription 3 (acute-phase response factor) | 0.000827 | 0.00124 |
| TP53 | tumor protein p53 | 0.0011 | 0.00132 |
| RELA | v-rel reticuloendotheliosis viral oncogene homolog A (avian) | 0.00365 | 0.00365 |

perspective. The causal association between IRF1 genetic variation and IBD was confirmed for the first time by SMR analysis (b_SMR = 1.5029, $p = 2.59E-09$), which broke through the limitations of traditional functional prediction studies and provided genetic evidence for the pathogenesis of IRF1. Secondly, through CIBERSORT and regulatory network analysis, the positive interaction between M1 macrophage infiltration and IRF1/GBP5 was revealed. By further constructing the "gene-immune cell/transcription factor" cooperative regulatory network, the deficiency of single gene function research is supplemented.

The methodological system of this study presents unique advantages in both single method and integrated analysis. At the level of individual methods, this study reduced the randomness of single gene analysis through modular screening through bioinformatics (WGCNA/PPI), improved diagnostic specificity through cross-validation through machine learning (RF + NaiveBayes), and established the causal relationship between gene and disease by using genetic variation to avoid environmental confusions. SMR combined with eQTL data to distinguish the direct genetic regulatory mechanisms of gene expression. At the integration level, we achieved the leap from correlation to mechanism analysis through the formation and verification of multi-omics evidence chains from biomarkers, immune interaction to genetic causation, combined with HEIDI test to eliminate linkage interference. This multi-omics strategy effectively avoids the limitations of a single technique and provides a new systems-biology perspective for the pathogenesis of IBD.

Considering the high incidence of IBD, early diagnosis and early treatment are crucial [22, 23]. The identification of IBD typically relies on endoscopic examination and histological analysis, although significant delays in clinical diagnosis remain prevalent [24, 25]. Consequently, it is essential to discover biomarkers in individuals with IBD. This research aims to identify potential diagnostic biomarkers for IBD and to investigate the connections between them.

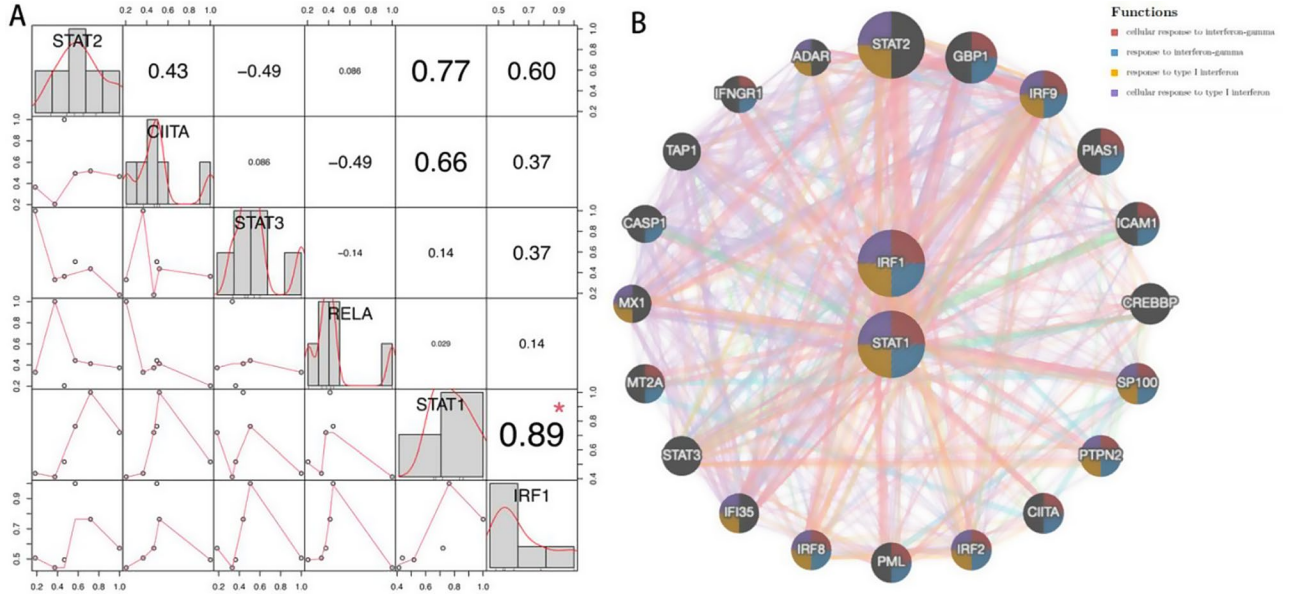



**Fig. 7** Interaction between TFs and characteristic gene-related pathways (**A**) Heatmap of the correlation between TFs and characteristic gene expression in HF (**B**) Regulatory network of TFs associated with characteristic genes

Among these three key genes, the interferon regulatory factor (IRF) family serves as a transcriptional regulator of the interferon system, which is curcial in modulating the body's immune response. *IRF1* was identified by Miyamoto et al. as a protein that binds to the regulatory elements of the human IFNβ gene. Under normal conditions, IRF is ubiquitous in the body and is expressed at low levels; however, in cases of viral infection or inflammation, IRF, especially IRF1 is significantly up-regulated [26]. Beyond its canonical roles, the molecule exhibits transcription factor activity that modulates fundamental processes including but not limited to inflammatory signaling and proliferative dynamics in tumorigenesis [26, 27]. *IRF1* is predominantly expressed in structural cells, such as epithelial cells, endothelial cells, fibroblasts [28], intestinal epithelial cells [29], and skin structural cells [28]. As a key transcription factor, *IRF1* promotes the shedding of intestinal epithelial cells triggered by tumor necrosis factor α (TNF-α) through a Caspase-3-dependent pathway, which is closely associated with the pathogenesis of IBD [30]. Therefore, inhibition of *IRF1* expression may represent a potential therapeutic strategy for IBD [31]. This is consistent with our findings.

*GBP5* belongs to the family of GTPases and is notably induced by interferon gamma (IFN-γ). This protein play a significant role in mediating various inflammatory responses within cells. Through its activation by IFN-γ, *GBP5* participates in the complex signaling pathways associated with inflammation, highlighting its crucial role in the body's immune response mechanisms, which include the activation of intracellular inflammasomes and the release of cytokines [32]. Furthermore the expression of *GBP5* is elevated in the colonic mucosa of patients with IBD. It is predominantly expressed in immune cells, including T lymphocytes, antigen-presenting cells, and macrophages [33].

Recent studies have confirmed the pro-inflammatory effects of *GBP5* in IBD. When *GBP5* is induced, the expression of caspase 1, gasdermin D and IL-1β is elevated. The NLRP3 inflammasome comprises the NLRP3 (inflammasome sensor), ASC (adaptor protein), and caspase 1 (effector). Activation of the NLRP3 inflammasome leads to the production of active forms of the pro-inflammatory cytokines IL-1β and IL-18. Therefore, it is plausible that the pro-inflammatory effects of *GBP5* may be mediated by the NLRP3 inflammasome [33, 34].

As a member of the PARPS family, *PARP9* participates in the interferon IFN-γ signal transducer and activator of the *STAT1* pathway, thereby promoting cell growth, viability, and motility [35]. In addition, it plays an essential role in the inflammatory process, exerting anti-inflammatory effects and participating in DNA repair and regulation of gene transcription. According to research, *PARP9* is involved in the disease processes associated with IBD, rheumatoid arthritis (RA), and other inflammatory diseases. Nonetheless, the exact mechanism remains unclear. For example, excess NAD + in IBD may lead to increased intestinal inflammation by increasing the activity of enzymes that consume NAD+, such as *PARP9* [36].

In our research, the CIBERSORT algorithm was used to investigate the connection between IBD characteristic genes and immune cell infiltration. Numerous studies have indicated that the immune microenvironment plays a crucial role in the development of IBD. Mitsialis

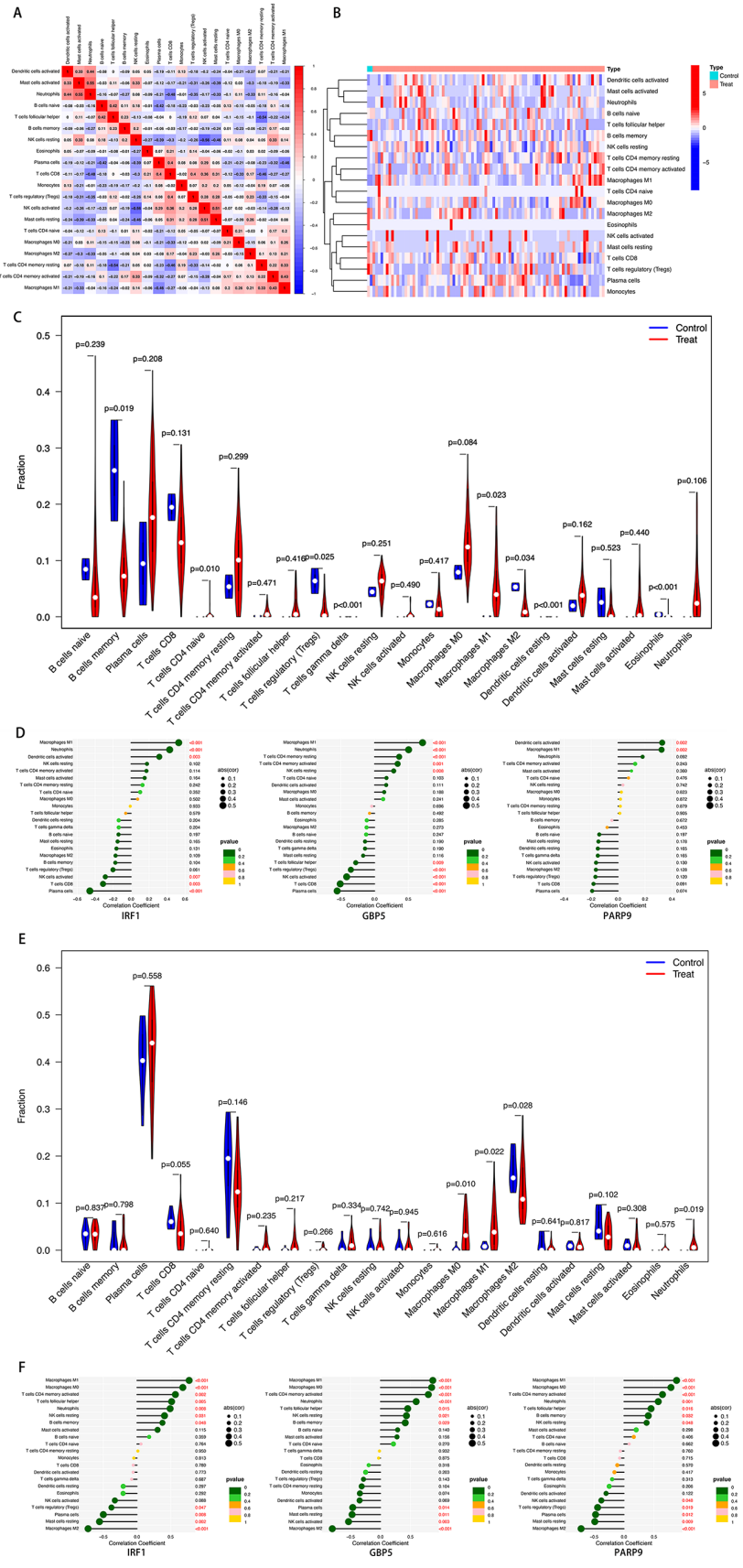



**Fig. 8** (See legend on next page.)

(See figure on previous page.)

**Fig. 8** Investigating the immune traits related to IBD and the immune infiltrating cells associated with central genes. (**A**) Correlation analysis of key immune cells in the training cohort. (**B**) Heat map of immune cells. (**C**) Analysis of 22 important immune-infiltrating cells in the training cohort and (**E**) validation cohort. (**D**) Lollipop plot showing the correlation between the central genes and immune infiltrating cells in the training set and (**F**) validation set

and colleagues identified notable variations in certain immune cell populations within mucosal and blood specimens of patients with UC [37]. Our study revealed that *IRF1*, *GBP5*, *PARP9*, and M1 macrophages were positively regulated in IBD, whereas plasma cells were negatively regulated. Nevertheless, these findings require further research to validate the role of immune response modulation in IBD.

The results revealed that three hub genes (*IRF1*, *PARP9*, *GBP5*) and four key TFs (CIITA, STAT3, RELA, STAT2) may be potential therapeutic targets for IBD. Additionally, several small molecule drugs targeting these hub genes were predicted, including digitoxin niclosamide, digoxin, budesonide, celecoxib, hexachlorophene, podophyllotoxin, doxorubicin hydrochloride, rubidomycin hydrochloride targeting STAT3;and vincristine sulfate, vinblastine sulfate, pyrimethamine, voriconazole, podophyllotoxin, flouridine, dexamethasone, trifluridine, and doxorubicin hydrochloride targeting RELA. Further exploration of the drug actions and molecular mechanisms in cellular and animal studies is necessary.

Previous analyses reported that the three hub genes may hold promising potential as molecular diagnostic biomarkers and therapeutic targets for IBD. To further investigate the potential of these three hub genes as biomarkers for IBD, the current study employed SMR and HEIDI to assess the pleiotropic association between gene expression levels and IBD based on GWAS and eQTL data. As a result, a significant causal association was found between *IRF1* and an increased risk of IBD. Notably, genetic variants in *IRF1* have been shown to promote the development of IBD on large-scale GWAS data, implying the great potential of *IRF1* as a biomarker for IBD.

Despite the findings, it is important to recognize the limitations of the current research. Firstly, both the IBD GWAS dataset and the blood eQTL data are characterized by limited sample sizes, which may undermine the reliability of the results. Therefore, future research should incorporate larger GWAS and QTL datasets, as well as complementary analyses such as TWAS and colocalization [38], to yield more dependable and meaningful findings. Additionally, the study's population was exclusively European, which restricts the applicability of the outcomes across diverse global populations. Thus, it is crucial to gather a broader spectrum of demographic data to validate our findings. Furthermore, this investigation exclusively examined eQTL data from blood samples and did not include protein data. Given the complex mechanisms by which biomarkers influence IBD pathogenesis, further molecular investigations, including gene knockout studies and pharmacological assessments in cellular or animal models, are essential to support our conclusions.

Through a combination of different levels of bioinformatics, Mendelian randomization, and SMR analysis, our study reveals the relationship between *IRF1* and IBD, offering new insights into the mechanisms of the disease and identifying potential therapeutic targets. Future research should delve deeper into its function and mechanism of action, thereby paving the way for new avenues in IBD treatment.

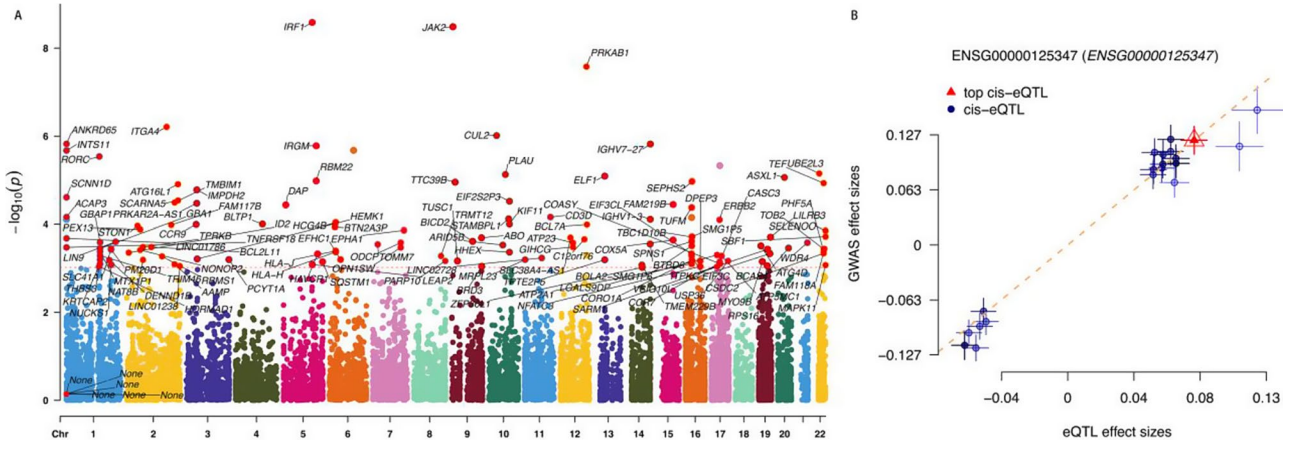



**Fig. 9** SMR analysis based on eQTL data and large-scale GWAS data (Finngen database). (**A**) Manhattan plot showing genes associated with IBD. (**B**) Scatter plot showing that IRF1 is positively associated with IBD risk

**Table 3** IRF1 SMR analysis

| Gene | b_SMR | se_SMR | p_SMR | p_HEIDI | nsnp_HEIDI | FDR |
|---|---|---|---|---|---|---|
| IRF1 | 1.5029 | 0.252361 | 2.59E-09 | 0.351 | 20 | 2.57E-06 |

## Conclusion

Several potential genes associated with IBD were identified, including *IRF1*, *GBP5*, and *PARP9*. SMR analysis based on eQTL data revealed that *IRF1* was significantly associated with the risk of IBD. Additional functional research is necessary to clarify the mechanisms involved, which would aid in the advancement of targeted molecular therapies and immunotherapies for IBD.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-06531-1.

Supplementary Material 1

Supplementary Material 2

**Acknowledgements**

We express our gratitude to TCGA and the GEO database for their platforms and the contributors who uploaded significant datasets. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

**Author contributions**

The concept for the study was developed by TL, MZ, and HJ. Planning and supervision of data collection were carried out by TL and MZ. Data collection was executed by HJ and XG. Analysis of the data was performed by HJ, TZ, and HY. The manuscript was initially drafted by TL and HJ. Contributions to the writing of the manuscript were made by MZ, TL, and XZ. All authors have engaged in discussions and reached a consensus that this is the final version of the manuscript, and they have agreed to its publication.

**Funding**

This study was funded by Tianjin Key Medical Discipline (Specialty) Construction Project, Key Discipline Special Project of Tianjin Health Science and Technology Program (TJWJ2023XK015), Tianjin Science and Technology Bureau Science and Technology Popularization Project (23KPHDRC00220).

**Data availability**

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Ethics approval and consent to participate**

GEO is a public database. Since our study relies on open-source data, it encounters no ethical concerns or conflicts of interest.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

Received: 8 March 2025 / Accepted: 23 April 2025
Published online: 03 July 2025

**References**

1. Abraham C, Cho JH. Inflammatory bowel disease. N Engl J Med. 2009;361:2066–78. https://doi.org/10.1056/NEJMra0804647
2. Luo Q, Wang J, Ge W, Li Z, Mao Y, Wang C, Zhang L. Exploration of the potential causative genes for inflammatory bowel disease: Transcriptome-wide association analysis, Mendelian randomization analysis and bayesian colocalisation. Heliyon. 2024;10:e28944. https://doi.org/10.1016/j.heliyon.2024.e28944.
3. Brudek T. Inflammatory bowel diseases and Parkinson's disease. J Parkinsons Dis. 2019;9:S331–44. https://doi.org/10.3233/JPD-191729.
4. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, Ming C, Neff R, Ma W, Fullard JF, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018;5:180185. https://doi.org/10.1038/sdata.2018.185.
5. Montaner J, Ramiro L, Simats A, Tiedt S, Makris K, Jickling GC, Debette S, Sanchez JC, Bustamante A. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. Nat Rev Neurol. 2020;16:247–64. https://doi.org/10.1038/s41582-020-0350-6.
6. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7. https://doi.org/10.7554/eLife.34408.
7. Flam E, Jang C, Murashige D, Yang Y, Morley MP, Jung S, Kantner DS, Pepper H, Bedi KC Jr., Brandimarto J, et al. Integrated landscape of cardiac metabolism in end-stage human nonischemic dilated cardiomyopathy. Nat Cardiovasc Res. 2022;1:817–29. https://doi.org/10.1038/s44161-022-00117-6.
8. Kerr MK. Linear models for microarray data analysis: hidden similarities and differences. J Comput Biol. 2003;10:891–901. https://doi.org/10.1089/106652703322756131.
9. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.
10. Cao J, Liu Z, Liu J, Li C, Zhang G, Shi R. Bioinformatics analysis and identification of genes and pathways in ischemic cardiomyopathy. Int J Gen Med. 2021;14:5927–37. https://doi.org/10.2147/IJGM.S329980.
11. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.
12. Yu G, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015;31:608–9. https://doi.org/10.1093/bioinformatics/btu684.
13. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504. https://doi.org/10.1101/gr.1239303.
14. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447–452. https://doi.org/10.1093/nar/gku1003.
15. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. CytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8(Suppl 4):S11. https://doi.org/10.1186/1752-0509-8-S4-S11.
16. Liu Z, Liu L, Weng S, Guo C, Dang Q, Xu H, Wang L, Lu T, Zhang Y, Sun Z, Han X. Machine learning-based integration develops an immune-derived LncRNA signature for improving outcomes in colorectal cancer. Nat Commun. 2022;13:816. https://doi.org/10.1038/s41467-022-28421-6.
17. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.
18. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46:D380–6. https://doi.org/10.1093/nar/gkx1013.
19. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen, C., Psychiatric Genomics, C., Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control, C., Duncan, L., et al. An atlas of

genetic correlations across human diseases and traits. Nat Genet. 2015; 47: 1236–1241. https://doi.org/10.1038/ng.3406
20. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7. https://doi.org/10.1038/ng.3538.
21. Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, Jeon M, Kang J, Tan AC. DSigDB: drug signatures database for gene set analysis. Bioinformatics. 2015;31:3069–71. https://doi.org/10.1093/bioinformatics/btv313.
22. Noor NM, Sousa P, Paul S, Roblin X, Early Diagnosis. Early stratification, and early intervention to deliver precision medicine in IBD. Inflamm Bowel Dis. 2022;28:1254–64. https://doi.org/10.1093/ibd/izab228.
23. Parente P, Pastore M, Grillo F, Fassan M, Francalanci P, Dirodi A, Rossi C, Arpa G, De Angelis P, Gullo I, et al. Very early Onset-IBD: evidence for the need of a multidisciplinary approach. Pathologica. 2022;114:3–11. https://doi.org/10.32074/1591-951X-336.
24. Ricciuto A, Mack DR, Huynh HQ, Jacobson K, Otley AR, deBruyn J, El-Matary W, Deslandres C, Sherlock ME, Critch JN, et al. Diagnostic delay is associated with complicated disease and growth impairment in paediatric Crohn's disease. J Crohns Colitis. 2021;15:419–31. https://doi.org/10.1093/ecco-jcc/jjaa197.
25. Ajbar A, Cross E, Matoi S, Hay CA, Baines LM, Saunders B, Farmer AD, Prior JA. Diagnostic delay in pediatric inflammatory bowel disease: A systematic review. Dig Dis Sci. 2022;67:5444–54. https://doi.org/10.1007/s10620-022-07452-5.
26. Taniguchi T, Ogasawara K, Takaoka A, Tanaka N. IRF family of transcription factors as regulators of host defense. Annu Rev Immunol. 2001;19:623–55. https://doi.org/10.1146/annurev.immunol.19.1.623.
27. Chen F, Song J, Di J, Zhang Q, Tian H, Zheng J. IRF1 suppresses Ki-67 promoter activity through interfering with Sp1 activation. Tumour Biol. 2012;33:2217–25. https://doi.org/10.1007/s13277-012-0483-3.
28. Geng F, Chen J, Song B, Tang Z, Li X, Zhang S, Yang T, Liu Y, Mo W, Zhang Y, et al. Chaperone- and PTM-mediated activation of IRF1 tames radiation-induced cell death and the inflammatory response. Cell Mol Immunol. 2024;21:856–72. https://doi.org/10.1038/s41423-024-01185-3.
29. Tang R, Yang G, Zhang S, Wu C, Chen M. Opposite effects of interferon regulatory factor 1 and osteopontin on the apoptosis of epithelial cells induced by TNF-alpha in inflammatory bowel disease. Inflamm Bowel Dis. 2014;20:1950–61. https://doi.org/10.1097/MIB.0000000000000192.
30. Giese-Kim N, Wu M, Dehghan M, Sceats LA, Park KT. Home Infliximab infusions are associated with suboptimal outcomes without cost savings in inflammatory bowel diseases. Am J Gastroenterol. 2020;115:1698–706. https://doi.org/10.14309/ajg.0000000000000750.
31. Tan G, Huang C, Chen J, Chen B, Shi Y, Zhi F. An IRF1-dependent pathway of TNFalpha-induced shedding in intestinal epithelial cells. J Crohns Colitis. 2022;16:133–42. https://doi.org/10.1093/ecco-jcc/jjab134.
32. Fujiwara Y, Hizukuri Y, Yamashiro K, Makita N, Ohnishi K, Takeya M, Komohara Y, Hayashi Y. Guanylate-binding protein 5 is a marker of interferon-gamma-induced classically activated macrophages. Clin Transl Immunol. 2016;5:e111. https://doi.org/10.1038/cti.2016.59.
33. Li Y, Lin X, Wang W, Wang W, Cheng S, Huang Y, Zou Y, Ke J, Zhu L. The Proinflammatory role of Guanylate-Binding protein 5 in inflammatory bowel diseases. Front Microbiol. 2022;13:926915. https://doi.org/10.3389/fmicb.2022.926915.
34. Gao S, Li Y, Wu D, Jiao N, Yang L, Zhao R, Xu Z, Chen W, Lin X, Cheng S, et al. IBD Subtype-Regulators IFNG and GBP5 identified by causal inference drive more intense innate immunity and inflammatory responses in CD than those in UC. Front Pharmacol. 2022;13:869200. https://doi.org/10.3389/fphar.2022.869200.
35. Iwata H, Goettsch C, Sharma A, Ricchiuto P, Goh WW, Halu A, Yamada I, Yoshida H, Hara T, Wei M, et al. PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. Nat Commun. 2016;7:12849. https://doi.org/10.1038/ncomms12849.
36. Wnorowski A, Wnorowska S, Kurzepa J, Parada-Turska J. Alterations in kynurenine and NAD(+) salvage pathways during the successful treatment of inflammatory bowel disease suggest HCAR3 and NNMT as potential drug targets. Int J Mol Sci. 2021;22. https://doi.org/10.3390/ijms222413497.
37. Mitsialis V, Wall S, Liu P, Ordovas-Montanes J, Parmet T, Vukovic M, Spencer D, Field M, McCourt C, Toothaker J, et al. Single-Cell analyses of Colon and blood reveal distinct immune cell signatures of ulcerative colitis and Crohn's disease. Gastroenterology. 2020;159:591–e608510. https://doi.org/10.1053/j.gastro.2020.04.074.
38. Mai J, Lu M, Gao Q, Zeng J, Xiao J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol. 2023;6:899. https://doi.org/10.1038/s42003-023-05279-y.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.