Heliyon 10 (2024) e38101

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Intestinal flora and inflammatory bowel disease: Causal relationships and predictive models

Guan-Wei Bi [a,b,1], Zhen-Guo Wu [a,b], Yu Li [a,b], Jin-Bei Wang [a,b], Zhi-Wen Yao [a,b], Xiao-Yun Yang [b], Yan-Bo Yu [b,c,*]

[a] *First Clinical College, Shandong University, Jinan, Shandong Province, PR China*
[b] *Department of Gastroenterology, Qilu Hospital, Shandong University, Jinan, Shandong, PR China*
[c] *Shandong Provincial Clinical Research Center for Digestive Disease, Qilu Hospital, Shandong University, Jinan, Shandong Province, PR China*

ARTICLE INFO

Keywords:
Inflammatory bowel disease
Mendelian randomization
Transcriptomic analysis
Machine learning
Intestinal flora
Predictive modeling

ABSTRACT

Background: Inflammatory bowel disease (IBD), including Crohn's disease and ulcerative colitis, is significantly influenced by intestinal flora. Understanding the genetic and microbiotic interplay is crucial for IBD prediction and treatment.
Methods: We used Mendelian randomization (MR), transcriptomic analysis, and machine learning techniques, integrating data from the MiBioGen Consortium and various GWAS datasets. SNPs associated with intestinal flora were mapped to genes, with LASSO regression refining gene selection. Differentially expressed genes (DEGs) and immune infiltration patterns were identified through transcriptomic analysis. Six machine learning models were used for predictive modeling.
Findings: MR analysis identified 25 gut microbiota classifications causally related to IBD. SNP mapping and gene expression analysis highlighted 24 significant genes. Drug target MR and colocalization validated these genes' causal relationships with IBD. Key pathways identified included the PI3K-Akt signaling pathway and epithelial-mesenchymal transition. Immune infiltration analysis revealed distinct patterns between high and low LASSO score groups. Machine learning models demonstrated high predictive value, with soft voting enhancing reliability.
Interpretation: By integrating MR, transcriptomic analysis, and sophisticated machine learning approaches, this study elucidates the causal relationships between intestinal flora and IBD. The application of machine learning not only enhanced predictive modeling but also offered new insights into IBD pathogenesis, highlighted potential therapeutic targets, and established a robust framework for predicting IBD onset.

Research in context

Evidence before this study

Although the role of the intestinal flora in the onset and development of IBD has been well established, the related genes, pathways, and mechanisms still require further investigation.

[*] Corresponding author. Department of Gastroenterology, Qilu Hospital, Shandong University, Jinan, Shandong, PR China.
E-mail address: yuyanbo@email.sdu.edu.cn (Y.-B. Yu).

[1] First author of the study.

https://doi.org/10.1016/j.heliyon.2024.e38101
Received 2 August 2024; Received in revised form 17 September 2024; Accepted 17 September 2024
Available online 20 September 2024

Added value of this study

We analyzed the intestinal flora taxa causally related to IBD using Mendelian randomization and identified a new set of genes associated with the intestinal flora and IBD. We validated their efficacy through drug target MR, SMR, and Coloc methods, observed the pathways, functions, and immune infiltration of the related genes through transcriptomic analysis, and constructed a disease model using six different machine learning methods.

Implications of all the available evidence

The identified genes were further tested for their causal relationship with IBD, and their roles in pathways, mechanisms, and immune infiltration were examined. Finally, six different machine learning models were used to construct a disease model.

1. Introduction

Inflammatory bowel disease (IBD) encompasses gastrointestinal inflammatory conditions, primarily Crohn's disease and ulcerative colitis [1]. Historically, IBD has been considered a disease of the Western world and has become a significant global health challenge since the beginning of the 21st century [1,2]. While the precise etiology of IBD remains elusive, numerous studies have sought to uncover its underlying causes [3]. A widely recognized contributing factor is intestinal flora with many investigations highlighting potential causal links [4]. The composition of intestinal flora is strongly associated with genetic susceptibility and genes influencing these intestinal flora may also be linked to IBD [4]. Identifying a set of genes causally related to IBD, analyzing their functions, and constructing a pathogenesis model could be instrumental in predicting IBD onset and developing potential treatments.

Mendelian randomization (MR) is an analytical method that uses genome-wide association studies (GWAS) to investigate causal relationships between two traits [5]. It leverages single nucleotide polymorphisms (SNPs) identified by GWAS along with specific information for each SNP including effect sizes, standard errors, and P-values [6]. These data are utilized to quantitatively analyze potential causal effects between traits via rigorous statistical methods [7]. Recently, drug target MR has gained popularity among researchers as it can directly analyze causal relationships between specific genes and traits via quantitative trait locus (QTL) data [8]. In addition to directly extracting QTL data and performing MR analysis via traditional methods, colocalization analysis and summary-data-based Mendelian randomization (SMR) analysis can also be employed to investigate the causal relationships of specific genes and traits [9,10].

Transcriptomic analysis is a widely used bioinformatics method that integrates clinical information with gene expression matrices from samples to identify differentially expressed genes, immune infiltration patterns, pathways, and functional characteristics among different sample groups [11–16].

Machine learning is a discipline focused on the theory and methods of simulating human learning activities, acquiring knowledge and skills, and improving system performance via computers [17]. By building models with input training data, machine learning can predict disease occurrence on the basis of specific parameters [17]. Currently, the construction of machine learning models is widely used in medical research across various diseases [18–20].

2. Methods

The flowchart of this study is shown in Fig. 1.

2.1. Data source

The GWAS data for intestinal flora were obtained from the MiBioGen Consortium and included approximately 19,000 subjects from 18 different populations. Intestinal flora information and GWAS data were collected through 16S RNA sequencing and whole-genome SNP arrays [21].

The GWAS data for IBD were sourced from 3 studies including the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), FinnGen, and Mbatchou J et al. The IIBDGC dataset included 12,882 cases and 21,770 controls and analyzed 12,716,084 SNPs. The FinnGen dataset comprises 5673 cases and 213,119 controls with a total of 16,380,466 SNPs. Mbatchou J et al. analyzed GWAS data from 4101 cases and 480,497 controls and identified 9,587,836 SNPs. The 3 IBD GWAS datasets were all sourced from the European population [22–24].

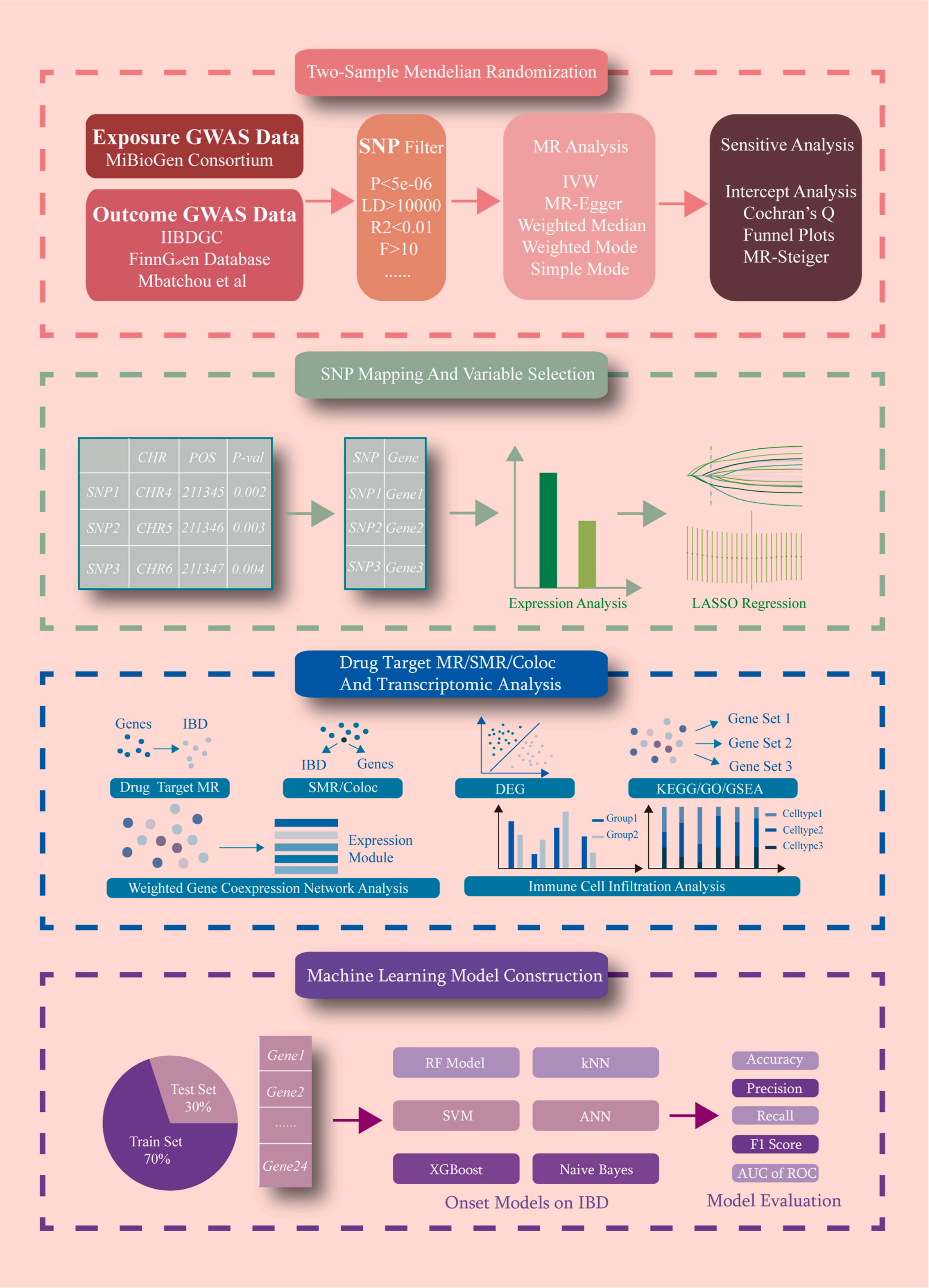
Expression quantitative trait locus (eQTL) data were sourced from Phase I of the eQTLGen Consortium, which aimed to investigate the genetic architecture of blood gene expression and understand the genetic underpinnings of complex traits [25].

Bulk gene expression data for IBD were obtained from datasets GSE36807 [26] and GSE75214 [27] in the Gene Expression Omnibus (GEO) database, which is a public repository that archives and freely distributes microarray gene expression data and other types of genomics data [28].

2.2. Mendelian randomization analysis between intestinal flora and IBD

MR-STROBE [29] is shown in Supplementary Table 1.

The GWAS data for intestinal flora were sourced from the MiBioGen Consortium. To conduct a valid MR analysis, the instrumental



(caption on next page)

Fig. 1. A flowchart of the study. GWAS, genome-wide association studies; IIBDGC, International Inflammatory Bowel Disease Genetics Consortium; SNP, single nucleotide polymorphism; LD, linkage disequilibrium; IVW, inverse variance weighted method; CHR, chromosome; POS, SNP position; LASSO, least absolute shrinkage and selection operator; SMR, summary-based Mendelian randomization; DEG, differentially expressed gene; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, gene set enrichment analysis; RF, random forest; SVM, supporting vector machine; XGBoost, extreme gradient boosting; kNN, k-nearest neighbor; ANN, artificial neural network; AUC, area under the curve; ROC, receiver operating characteristic curve.

variables (SNPs) must meet the following three conditions.1) There must be a strong association between the SNP and the exposure trait (intestinal flora in this study). 2) The SNP must have a weak correlation with the outcome trait (IBD in this study). 3) SNPs can influence the outcome trait only through the exposure trait [30] (Fig. 2).

To meet these conditions, the following criteria were used to screen SNPs.

1) The P-value threshold for the association between SNPs and exposure traits was set at a P-value<5e-06 [31–33]. 2) The SNP linkage disequilibrium (LD) distance was set at 10,000, and the correlation index r^2 was set at 0.001 [33]. 3) The P-value threshold for the association between SNPs and outcome traits was set at a P-value>0.05 [34]. 4) Proxy SNPs were not used. 5) The minor allele frequency (MAF) of each SNP had to be greater than 0.01 [35]. 6) Palindromic SNPs were excluded [35]. 7) The F-statistic for the weak instrument variable test had to be greater than 10 [36].

To calculate the F-statistic, the correlation coefficient R^2 for each SNP with the trait must be determined. The following formulas are used to calculate R^2 and the F-statistic: $R^2 = 2 \times \text{MAF} \times (1-\text{MAF}) \times \text{beta}$ [2]/[$2 \times \text{MAF} \times (1\text{-MAF}) + \text{Se}(\text{beta}$ [2]$) \times 2 \times \text{N} \times \text{MAF} \times (1\text{-MAF})$ [37], $\text{F}=\text{R}^2 \times (\text{N}-2)/(1-\text{R}^2)$ [38], where EAF represents the effect allele frequency, beta represents the effect size between the SNP and exposure trait, N represents the sample size, and Se(beta) represents the variance of beta.

Inverse-variance weighted (IVW) [39], MR-Egger [40], weighted mode [41], simple mode [41], and weighted median [42] methods are used in MR analysis, and if heterogeneity and pleiotropy are detected, then the IVW random effect model is applied [43]. The IVW method is the gold standard and one of the most commonly used MR methods. It assumes that all instrumental variables are valid and there is no horizontal pleiotropy [39]. The IVW method provides efficient and robust causal effect estimates but is sensitive to the assumption of no pleiotropy. MR-Egger regression is an extension of the IVW method that allows for horizontal pleiotropy. While MR-Egger increases the robustness to pleiotropy bias, it has lower efficiency and requires a strong association between instrumental variables and exposure [40]. The weighted median method is a robust approach that assumes at least 50 % of the instrumental variables are valid. Compared with the IVW method, the weighted median method is more robust, although it may be slightly less efficient [42]. The weighted mode method assumes that the majority of instrumental variables have the same effect and permits the presence of invalid instruments. It is particularly useful when there is heterogeneity in the effects of instrumental variables [41]. The simple mode method estimates causal effects without weighting the effects of instrumental variables and identifies the mode of effects. This method is relatively simple and a more robust effect on heterogeneity but it typically has lower efficiency and may perform poorly when the number of instrumental variables is small [41].

For the sensitivity analysis, Cochran's Q test [44], MR-Egger intercept analysis [45] and MR-Steiger [46] were employed. Cochran's Q test was employed to assess heterogeneity in SNP effects within the MR analysis. It compares the contributions of individual SNPs to causal estimates and indicates significant differences in SNP effects when the Q value is significant (P-value>0.05), which suggests potential effect heterogeneity [44]. The MR-Egger intercept analysis is a method used to detect and correct for horizontal pleiotropy. Unlike conventional MR methods, MR-Egger does not require all SNPs to be valid instruments and estimates and

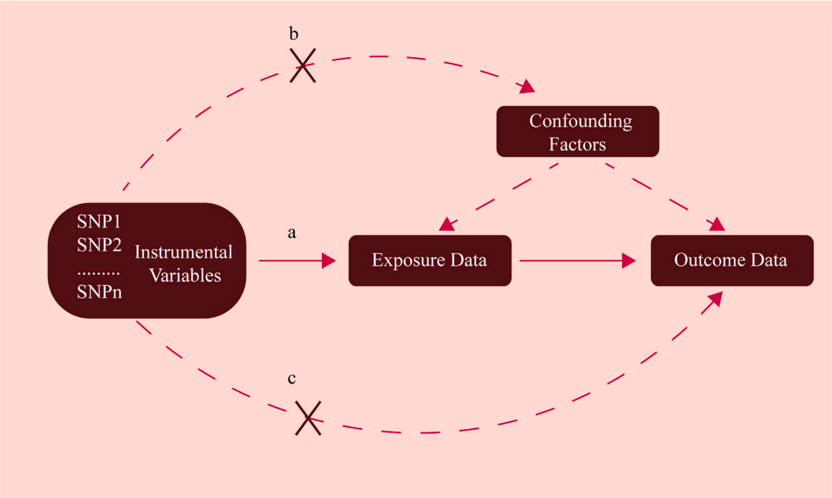


Fig. 2. A schematic plot of Mendelian randomization. (a) Instrumental variables must influence the outcome through exposure. (b) Instrumental variables should not influence outcomes through factors other than exposure. (c) Instrumental variables should not influence outcomes directly. SNP, single nucleotide polymorphism.

adjusts for potential horizontal pleiotropy bias through an intercept term. It also detects bias in MR estimates, which is particularly useful when there is potential horizontal pleiotropy in the data [45]. To assess the possibility of reverse causation effects, we utilized the MR-Steiger method. This method helps researchers assess whether observed associations are more consistent with the exposure causing the outcome, or vice versa, and thereby provides insights into potential reverse causation effects in MR analyses [46].

To control the false negative rate and improve the statistical power of our analysis, we utilized the mRnd website to calculate the statistical power. Then, studies with statistical power greater than 0.50 were included in our analysis [47].

All MR analyses were subjected to P-value correction via the Benjamini−Hochberg method [48]. A result was considered positive if the initial P-value was less than 0.05 and if the adjusted P-value (or P-adjust) was less than 0.10. If a gut microbe shows positive results across all 3 IBD outcome variables, it is selected for further analysis [48–50].

2.3. SNP mapping to genes

We extracted SNPs from the intestinal flora that were identified as positively associated with inflammatory bowel disease (IBD) and used the g:Profiler website [51] and PLINK software [52] to map these SNPs to specific genes. PLINK, a free, open-source whole genome association analysis toolkit, is engineered for basic to large-scale analyses with computational efficiency. It includes a feature that maps SNPs to genes, which indicates the distance between each SNP and its corresponding gene region [52]. We selected results where SNPs are located within gene regions. g:Profiler, which is a tool for functional enrichment and gene translation, also allows SNP mapping to genes via its g:SNPense feature [51]. By combining results from both PLINK and g:Profiler, we compiled a comprehensive list of genes to ensure an integrated understanding of the genetic impacts linked to the gut microbiota in IBD.

2.4. Expression analysis and variable selection

The expression levels of genes previously mapped in the disease and normal groups were compared via the GEO database. The method used for comparison was the Wilcoxon rank-sum test. The Benjamini−Hochberg method was applied for P-value correction, and results with P-value$<$0.05 and P-adj $<$0.10 were selected for variable selection.

The LASSO method was then used to refine the variables among the positive results to identify a set of genes most significantly associated with the onset of IBD [53,54].

LASSO regression involves the addition of a regularization term, which is characterized by a penalty coefficient λ, to minimization of the residual sum of squares. This regularization term is the sum of the absolute values of the model coefficients (L1-norm). Consequently, this method can shrink some of the coefficients to zero and effectively perform variable selection. As the penalty coefficient λ increases, the coefficients of certain variables gradually decrease to zero. The optimal value of λ is selected via 10-fold cross-validation with λ corresponding to the minimum cross-validation error that is used for model construction. The coefficients of variables reflect the influence of each gene on the target variable such as disease status. On the basis of these coefficients, a LASSO score model can be constructed. This model evaluates individuals by calculating the weighted sum of gene expression levels and their corresponding coefficients [54].

2.5. Drug target MR, SMR and colocalization analysis

Drug target MR leverages eQTL data from the eQTLGen Consortium database Phase I as exposure data with IBD GWAS data serving as outcome data. eQTL data for genes that showed positive results in the LASSO analysis were extracted for drug target MR analysis of IBD. Similarly, the IVW method is regarded as the gold standard for testing causal relationships [55].

The SMR (Summary-data-based Mendelian randomization) software tool, which was initially developed to implement the SMR & HEIDI methods, tests for pleiotropic associations between the expression level of a gene and a complex trait via summary-level data from GWAS and eQTL studies. The SMR test integrates eQTL and GWAS data to estimate the effect size of the exposure variable on the outcome variable and statistically tests this effect size. The HEIDI (Heterogeneity in Dependent Instruments) test examines whether associations identified by the SMR test could be due to multiple causal effects or collinearity. If the HEIDI test shows inconsistent effects (P-value$<$0.05), then the association may not be driven by a single causal pathway but could be influenced by multiple causal effects or other complex genetic regulatory mechanisms [10].

For colocalization analysis, the Coloc R package was used. Coloc analysis is based on 5 hypotheses that are used to assess whether two traits share the same genetic variant. The hypotheses are as follows. 1) H0: Neither trait has a causal association in the genomic region. 2) H1: The first trait has a causal association in the genomic region, whereas the second trait does not. 3) H2: The second trait has a causal association in the genomic region, whereas the first trait does not. 4) H3: Both traits have causal associations in the genomic region, but these signals are driven by different genetic variants. 5) H4: Both traits have causal associations in the genomic region, and these signals are driven by the same genetic variant and this indicates a shared genetic variant. Coloc analysis calculates the posterior probability for each hypothesis and helps to determine which hypothesis is most likely to explain the observed data. If the posterior probability of H3+H4 is high, the two traits are both likely to have a causal relationship to the outcome traits. This information is crucial for understanding the genetic linkage and potential biological mechanisms between these two traits [9].

2.6. Transcriptomic analysis

The gene expression patterns of IBD patients were analyzed using the weighted gene co-expression network analysis (WGCNA) method. WGCNA is a method for analyzing gene co-expression patterns. This method assumes that gene expression networks in biological systems follow a scale-free topology, which is characterized by a few hub genes with extensive connections and most other genes with relatively few connections. In these types of networks, the number of gene connections follows a negative exponential function with respect to their probability of occurrence. When using WGCNA, Pearson correlation coefficients between expression levels of all gene pairs are calculated first. Next, an appropriate soft threshold is selected to transform the correlation matrix into an adjacency matrix, which represents the connection strength between genes. Then, the topological overlap matrix (TOM) is constructed to measure the interconnectedness of genes. Finally, hierarchical clustering is performed on the basis of the TOM to group genes in modules of co-expressed genes. These steps enable the identification of gene modules with similar expression patterns and facilitate the discovery of gene interactions and potential biological functions. Genes selected via LASSO are then compared with the hub genes identified via WGCNA. If a selected gene is present in a specific WGCNA gene expression module, then all genes within that corresponding module will be extracted for further analysis [15].

Moreover, differentially expressed gene (DEG) analysis was conducted between the disease group/control group and between the high-score/low-score groups on the basis of LASSO results [11] via the Limma R package [56]. Through DEG analysis, we identified genes that were differentially expressed in respective groups. Due to the cascading amplification effect in gene expression, some genes with relatively small log2FC values may have significant impacts. Therefore, DEG analysis results are then filtered based only on P-adj<0.10 without applying a log2FC threshold [56].

The intersection of genes identified from the DEG analysis was examined and the union of these intersecting genes with genes identified from the WGCNA was determined. These intersecting genes were further subjected to gene enrichment analyses, including Gene Ontology (GO) analysis [14], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis [13], and Gene Set Enrichment Analysis (GSEA) [16], using Hallmark gene sets. These analyses provided insights into the functional and pathway characteristics of genes.

Additionally, immune infiltration differences between the high- and low-LASSO score groups were examined via CIBERSORT [12], which is an approach to characterize the immune cell composition of complex tissues from their gene expression profiles. This analysis explored the differences in immune functions between the two groups and enhanced the understanding of their roles in IBD.

2.7. Construction of an IBD onset model based on machine learning

The GEO IBD dataset was randomly divided into training and testing sets at a 70:30 ratio. In the training set, machine learning models, including random forest (RF) [57], support vector machine (SVM) [58], extreme gradient boosting (XGBoost) [59], artificial neural networks (ANN) [60], naive Bayes [61], and k-nearest neighbors (kNN) [62], were constructed. The parameters of all models were tuned using the R package ‘caret’ to identify those that minimized the error.

RF is an ensemble learning method that constructs multiple decision trees and aggregates their results for classification or regression [57]. SVM finds the optimal hyperplane that separates data points into different classes in a high-dimensional space. The algorithm maximizes the margin between the classification boundary and the nearest data points and enhances the model’s generalization ability [58]. XGBoost is an efficient implementation of the boosting tree algorithm. It iteratively trains multiple weak learners and optimizes the loss function at each iteration to build a strong predictive model [59]. ANN is computational models inspired by biological neural networks and consist of input layers, hidden layers, and output layers; each layer contains multiple neurons connected by weighted links. By adjusting these weights, neural networks can learn and approximate complex nonlinear functions [60]. Naive Bayes is a probabilistic classifier that is based on Bayes’ theorem and assumes independence among predictors. kNN is an instance-based classification and regression algorithm. For each test sample, the algorithm calculates the Euclidean distance to all training samples, selects the k closest neighbors, and makes a prediction [62].

Each model is then evaluated in the test set on the basis of metrics such as precision, accuracy, F1 score, recall, and the area under the receiver operating characteristic (ROC) curve (AUC). [63]To ensure the robustness of the model, we performed 400 times of 10-fold cross-validation for each model and draw a box plot of AUC.

Each machine learning model expresses an output of a probability between 0 and 1 where values closer to 1 indicate a higher propensity for disease onset. The final probability result is determined via a soft voting mechanism, which averages the probabilities provided by each machine learning model [64].

2.8. Statistical analysis

Statistical analyses were conducted via R 4.3.1 or corresponding software and websites. The hypothesis testing method typically employs the Wilcoxon rank-sum test, and the Benjamini−Hochberg method is the P-value adjustment method that is generally used.

3. Results

3.1. Mendelian randomization between intestinal flora and IBD

According to the selection criteria in Method 2.1, 8780 SNPs from 211 intestinal flora classifications (including 9 phyla, 16 classes, 20 orders, 35 families, and 131 genera) were extracted. After MR analysis and P-value adjustment, 25 gut microbiota classifications (including 4 phyla, 3 classes, 5 orders, 2 families, and 11 genera) yielded positive results across three IBD outcome variables. These 25 intestinal flora classifications have a causal effect on IBD. Sensitivity analyses were conducted for all 25 classifications. Detailed results can be found in Table 1 and Supplementary Tables 2–4.

3.2. SNP mapping to genes

A total of 4006 SNPs related to intestinal flora were extracted. In the gene mapping process and for PLINK results, we extracted all SNPs located within genes and assigned a GeneSymbol name, which was marked as “GeneSymbolName(0)." SNPs without a gene symbol were excluded. For g:Profiler results, if a SNP was mapped to two genes, the first gene was chosen for further analysis as the second gene is often a subtype of the first. The genes identified by both tools were then merged and deduplicated, which resulted in a final set of 174 genes.

3.3. Expression analysis and variable selection

Gene expression analysis was conducted via data from the GSE36807 and GSE75214 datasets in the GEO database. The ulcerative colitis and Crohn’s disease groups were combined into a single IBD group. After P-value adjustment, 76 genes were identified as significantly different between the disease and control groups (detailed results can be found in Fig. 3A–D and Supplementary Table 5).

In the LASSO analysis, 10-fold cross-validation was used to determine the minimum λ as 0.012 ($\ln\lambda = -4.423$), which resulted in 24 genes selected for further analysis. The LASSO score for each sample was calculated via the formula score $= \Sigma$ (gene coefficient $\times$ gene expression) (see Fig. 3E–F and Supplementary Table 5).

3.4. Drug target MR, SMR and colocalization analysis

Due to limitations in eQTL data, only 21 of 24 genes were selected for drug target MR, SMR, and Coloc analyses. Because of data constraints, only IIBDGC GWAS data were used for the analysis.

In the drug target MR analysis, 15 of these genes were significantly and differentially expressed after drug target MR analysis and P-value adjustment. In the SMR analysis and after NA values were removed from the results, 13 SNPs of 3 genes were determined significant (P-SMR $<$0.10, P-HEIDI $>$0.05). In the Coloc colocalization analysis and of the 21 genes analyzed, 12 genes had P(H3) + P(H4) $>$ 0.80, and among these genes, 2 genes had P(H4) $>$ 0.75. Detailed results are presented in Table 2.

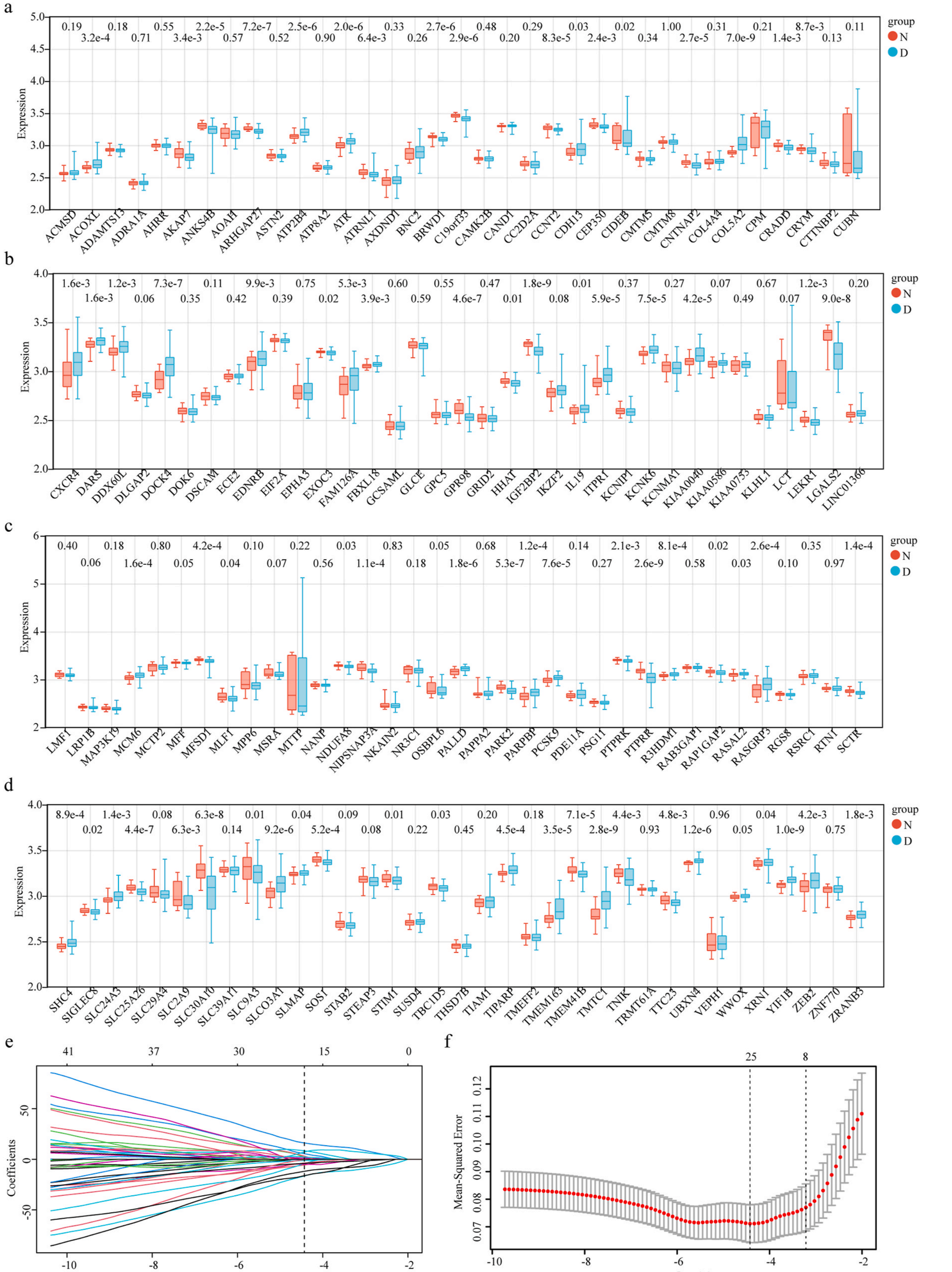
3.5. Transcriptomic analysis

In the WGCNA, genes with zero expression were filtered out, and sample hierarchical clustering was performed to remove outlier samples. A soft threshold power of 20 was chosen to achieve a scale-free topology model fit (R^2) of 0.85, and an average connectivity plot was generated. The adjacency matrix and topological overlap matrix (TOM) were calculated and dissimilarity (1-TOM) was also

Table 1
Mendelian randomization results of inverse variance weighted method.

Trait	IVW-OR (95%CI, P-value)		
	IIBDGC	Finngen	Mbatchou et al.
class.Actinobacteria.id.419	0.805(0.788–0.823,0)	0.83(0.810–0.851,0)	0.906(0.890–0.922,0)
class.Bacteroidia.id.912	1.161(1.092–1.236,0)	1.188(1.102–1.281,0)	1.192(1.138–1.248,0)
class.Negativicutes.id.2164	0.784(0.665–0.924,0.004)	0.633(0.484–0.829,0.001)	1.234(1.042–1.461,0.015)
family.Bifidobacteriaceae.id.433	0.784(0.766–0.802,0)	0.782(0.761–0.803,0)	0.939(0.923–0.955,0)
family.Peptostreptococcaceae.id.2042	1.322(1.232–1.419,0)	0.729(0.649–0.819,0)	1.14(1.083–1.200,0)
genus.Anaerostipes.id.1991	1.09(1.005–1.183,0.036)	2.067(1.875–2.279,0)	1.23(1.164–1.299,0)
genus.Bifidobacterium.id.436	0.806(0.788–0.824,0)	0.789(0.768–0.811,0)	0.935(0.920–0.951,0)
genus.Lachnoclostridium.id.11308	0.885(0.816–0.960,0.003)	0.508(0.460–0.560,0)	0.721(0.680–0.766,0)
genus.Odoribacter.id.952	1.187(1.073–1.314,0.001)	1.545(1.368–1.745,0)	0.798(0.744–0.857,0)
genus.Parasutterella.id.2892	1.085(1.020–1.154,0.01)	0.904(0.840–0.973,0.007)	1.083(1.038–1.130,0)
genus.Peptococcus.id.2037	1.166(1.137–1.195,0)	0.912(0.885–0.939,0)	0.909(0.892–0.926,0)
genus.Romboutsia.id.11347	1.546(1.384–1.726,0)	0.489(0.425–0.563,0)	1.13(1.047–1.219,0.002)
genus.Ruminiclostridium9.id.11357	0.943(0.896–0.992,0.023)	1.313(1.232–1.399,0)	1.164(1.123–1.206,0)
genus.RuminococcaceaeUCG011.id.11368	1.043(1.020–1.067,0)	1.071(1.044–1.098,0)	0.972(0.955–0.988,0.001)
genus.unknowngenus.id.2071	1.155(1.030–1.296,0.014)	0.676(0.601–0.759,0)	0.904(0.846–0.966,0.003)
genus.Victivallis.id.2256	1.086(1.003–1.175,0.041)	0.856(0.778–0.942,0.001)	1.121(1.051–1.196,0.001)
order.Bacillales.id.1674	0.877(0.823–0.935,0)	0.778(0.711–0.852,0)	0.882(0.841–0.925,0)
order.Bacteroidales.id.913	1.161(1.092–1.236,0)	1.188(1.102–1.281,0)	1.192(1.138–1.248,0)
order.Bifidobacteriales.id.432	0.784(0.766–0.802,0)	0.782(0.761–0.803,0)	0.939(0.923–0.955,0)
order.Lactobacillales.id.1800	0.856(0.768–0.955,0.005)	1.152(1.020–1.301,0.023)	1.118(1.038–1.205,0.003)
order.Selenomonadales.id.2165	0.784(0.665–0.924,0.004)	0.633(0.484–0.829,0.001)	1.234(1.042–1.461,0.015)
phylum.Actinobacteria.id.400	0.721(0.695–0.748,0)	0.765(0.734–0.797,0)	0.906(0.881–0.931,0)
phylum.Bacteroidetes.id.905	1.166(1.098–1.239,0)	1.188(1.104–1.279,0)	1.185(1.133–1.240,0)
phylum.Proteobacteria.id.2375	1.537(1.367–1.728,0)	0.735(0.642–0.842,0)	0.842(0.779–0.909,0)
phylum.Verrucomicrobia.id.3982	0.706(0.658–0.757,0)	0.778(0.701–0.862,0)	0.87(0.828–0.915,0)

IVW, inverse variance weighted method; OR, odds ratio; CI, confidential interval; a P-value of 0 means the P-value$<$0.0001.



(caption on next page)

Fig. 3. Expression analysis and variable selection. (a–d) Bar chart of gene expression levels determined via GEO data. (e) LASSO coefficient path plot. (f) LASSO cross-validation error plot. LASSO, least absolute shrinkage and selection operator.

computed. Modules with a minimum size of 30 genes were identified, and modules with a merging distance of less than 0.25 were merged and resulted in 12 modules. Then, a gene dendrogram was plotted, module eigengenes were calculated, and an eigengene dendrogram was created. After all the hub genes were extracted, three of the 24 genes (PARPBP, CXCR4, and TNIK) were found in the pink, lightcyan, and turquoise modules. All 1313 node genes from these modules were extracted for further analysis (see Fig. 4A–F).

With a threshold of P-adj <0.10, DEG analysis was performed for both the IBD group vs. the normal group and the high-score group vs. the low-score group. From each DEG result, the top 1000 genes ranked by log2FC were selected, and union genes with the 1313 genes identified from the WGCNA were selected, which resulted in a final set of 1723 genes. Enrichment analysis of these 1723 genes via KEGG, GO, and GSEA hallmark datasets revealed significant enrichment in response to chemicals, regulation of response to stimulus, the PI3K−Akt signaling pathway, epithelial−mesenchymal transition, and other related pathways (see Fig. 5A–E).

In the high- and low-LASSO score groups, an immune infiltration analysis was performed via CIBERSORT, and a rainbow plot, grouped bar chart, and immune cell−gene correlation heatmap were generated. Results revealed that in the low-LASSO-score group, plasma cells, $CD8^+$ T cells, Treg cells, NK cells, M2 macrophages, and mast cells had relatively high infiltration levels. In the high-LASSO-score group, $CD4^+$ T cells, M0 macrophages, M1 macrophages, dendritic cells, and neutrophils had relatively high infiltration levels (see Fig. 5F–G).

3.6. *Construction of an IBD onset model based on machine learning*

The RF model was constructed via the randomForest package with the following hyperparameters: mtry (number of features selected at each split) was set to 4, ntree (number of trees) was set to 500, and other parameters were set to their default values. The model achieved an accuracy of 0.957, a precision of 0.952, a recall of 1, an F1 score of 0.975, and an AUC of 0.978 for the test set.

The SVM model was built via the e1071 package with the following hyperparameters: type set to ‘eps-regression’, cost (C-penalty parameter for support vectors) set to 10, gamma (parameter for radial basis kernel function) set to 0.01, epsilon set to 0.1, and other parameters set to their defaults. The SVM model achieved an accuracy of 0.975, a recall of 1, an F1 score of 0.975, and an AUC of 0.987.

The ANN model was constructed via the neuralnet package with the following hyperparameters: 2 hidden layers with 2 and 6 neurons and other parameters set to their defaults. The ANN model achieved an accuracy of 0.911, a precision of 0.930, a recall of 0.889, an F1 score of 0.909, and an AUC of 0.914.

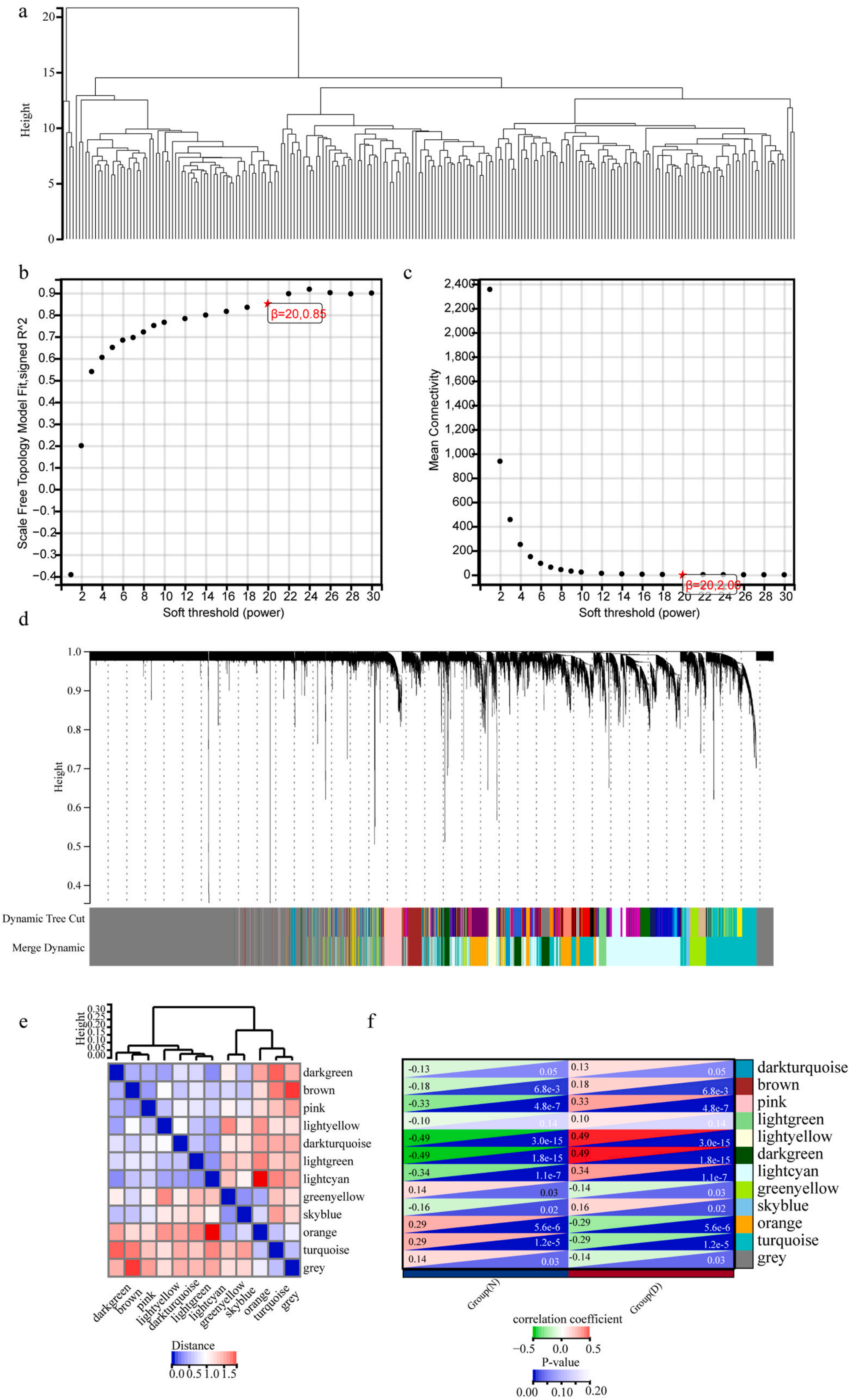
The XGBoost model was built via the XGBoost package with the following hyperparameters: nrounds (number of iterations) set to 300, max_depth (maximum tree depth) set to 6, eta (learning rate) set to 0.1, gamma (minimum loss reduction required for tree splitting) set to 0.1, colsample_bytree (feature subsample ratio) set to 0.8, and other parameters set to their defaults. The XGBoost model achieved an accuracy of 0.967, a precision of 0.937, a recall of 1, an F1 score of 0.967, and an AUC of 0.972.

Table 2
Drug target MR, colocalization and SMR results.

Gene	OR (95%CI), P-value	Colocalization Analysis P(H3) + P(H4)	P-SMR	P-HEIDI	Number of Significant SNPs in SMR[a]
ARHGAP27	1.232(1.221–1.242), 0	0.968	5.58e-03	0.04	0
C19orf33	N/A				
CCNT2	0.913(0.905–0.922), 0	1	0.221	0.425	0
CDH13	1.085(1.045–1.126), 0	1	0.958	0.747	0
CRADD	1.595(0.821–3.100), 0.200	1	N/A	N/A	0
CXCR4	1.307(1.086–1.573), 0.006	1	0.294	0.457	0
DDX60L	0.99(0.985–0.996), 0.001	1	0.564	0.348	0
EXOC3	1.087(1.083–1.091), 0	1	0.003	0.403	6
FBXL18	0.991(0.93–1.057), 0.788	1	N/A	N/A	N/A
GPR98	N/A	0.461	N/A	N/A	N/A
HHAT	0.985(0.952–1.02), 0.442	1	0.426	0.454	0
LEKR1	N/A				
MFSD1	1.043(1.031–1.054), 0	1	0.107	0.348	0
PALLD	1.122(1.089–1.157), 0	1	0.295	0.859	0
PARPBP	0.859(0.829–0.89), 0	1	0.986	N/A	0
PTPRR	N/A				
SLC25A26	0.87(0.844–0.896), 0	1	0.002	0.104	1
SLC2A9	0.925(0.919–0.932), 0	1	0.061	0.338	0
TBC1D5	1.277(1.216–1.341), 0	1	0.347	0.980	0
TIPARP	0.995(0.958–1.032), 0.788	0.997	0.126	N/A	0
TNIK	0.935(0.915–0.957), 0	1	0.135	0.802	0
TTC23	0.948(0.94–0.956), 0	1	0.201	0.380	0
XRN1	1.011(0.994–1.027), 0.247	N/A	0.368	0.783	0
YIF1B	0.878(0.869–0.886), 0	N/A	0.002	0.709	6

OR, odds ratio; CI, confidential interval.

[a] Significant SNPs must be with a P-SMR<0.05 and P-HEIDI>0.05; a P-value of 0 means the P-value<0.0001.



(caption on next page)

Fig. 4. Weighted gene co-expression network analysis. (a) Cluster dendrogram of the sample. (b) Scale independence plot. (c) Mean connectivity plot. (d) Cluster dendrogram of genes. (e) Clustering dendrogram of module eigengenes. (f) Module−trait relationship heatmap.

The kNN model was built via the e1071 package with the following hyperparameters: k (number of nearest neighbors) was set to 2 and other parameters were set to their defaults. The kNN model achieved an accuracy of 0.920, a precision of 0.947, a recall of 0.889, an F1 score of 0.917, and an AUC of 0.923.

The naive Bayes model was constructed via the e1071 package and the laplace (Laplace smoothing parameter) hyperparameter was set to 0. The naive Bayes model achieved an accuracy of 0.959, a precision of 0.923, a recall of 1, an F1 score of 0.960, and an AUC of 0.970.

The results of 400 times 10-fold cross-validation demonstrated the robustness of the model. The average AUC of the RF model was 0.9982435, the SVM model was 0.997, the ANN model was 0.947, the XGBoost model was 0.996, the kNN model was 0.974, and the naive Bayes model was 0.962.

Given the satisfactory performance of all models, soft voting can be applied to the final prediction by averaging the probabilities from the six models to predict the likelihood of IBD.

All outcomes can be found in Fig. 6 and Supplementary Fig. 1.

4. Discussion

Intestinal flora plays a crucial role in the occurrence and development of IBD [4]. Among the 25 intestinal flora classifications selected in this study, 8 classifications exhibit protective effects against IBD (OR<1 in three outcome GWASs), 4 classifications exhibit pathogenic effects against IBD (OR>1 in three outcome GWASs), and 13 classifications may have either protective or pathogenic effects on IBD (OR>1 in some GWASs and OR<1 in others).

Regardless of whether the role of these intestinal flora classifications is clear or unclear, the significant results from the three-outcome GWAS provide sufficient evidence to indicate a causal relationship between these intestinal flora classifications and IBD. Therefore, we extracted all SNPs from the GWASs of these intestinal flora classifications and mapped them to the corresponding genes.

We employed LASSO for feature selection. As previously mentioned, the choice of λ significantly impacts the results of LASSO. When using cross-validation for model selection, the results usually present the λ-min, where the error is minimized, and λ-min plus one standard error (λ-min + 1se). In this process, we chose λ-min, primarily considering its minimal error and because this paper is innovative, providing more valuable genes for further research is necessary. However, this also results in a larger number of genes, which brings potential inconvenience for clinical translation. After expression analysis and LASSO regression, we ultimately identified 24 genes potentially associated with IBD.

These 24 genes vary in their importance to IBD and perform differently in downstream analyses. In the drug target MR, SMR, and Coloc analyses, only SLC25A26 and EXOC3 had positive results in all three analyses. In contrast, C19orf33, PTPRR, and LEKR1 could not be analyzed further because of a lack of eQTL data.

Among the three analyses, SMR explores whether there are common driver SNPs between genes and traits. However, since an SNP is just a single base pair in the gene region, the absence of a common driver SNP does not necessarily indicate the absence of a causal relationship, which makes its importance relatively lower. Drug target MR and Coloc analysis, particularly P(H3+H4), better assess the causal relationships between genes and traits.

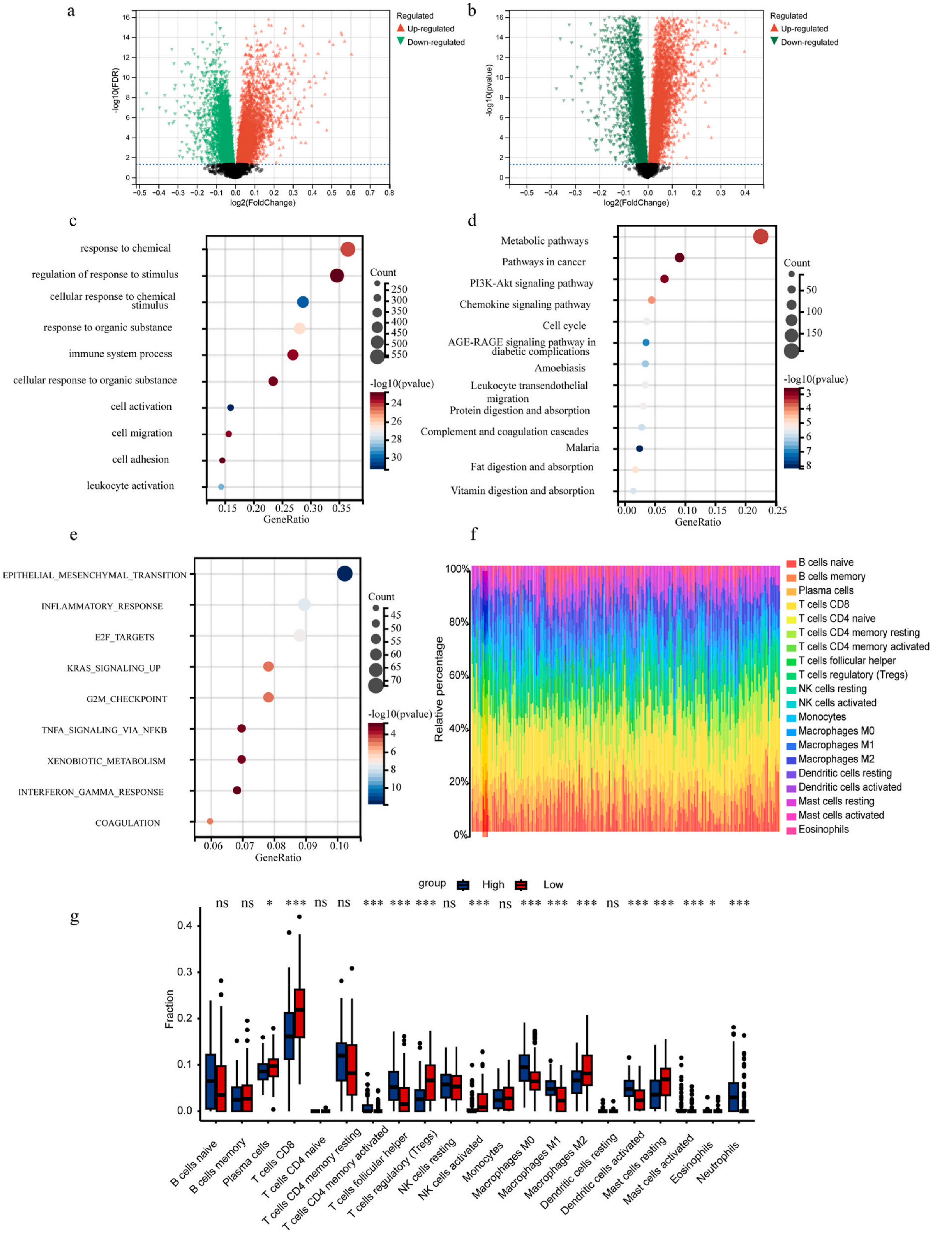
By excluding three genes lacking eQTL data, 15 out of the 20 genes presented positive results in both drug target MR and Coloc analyses. Five genes (FBXL18, GPR98, HHAT, TIPARP, and YIF1B) were positive in only one analysis, and only CRADD and XRN1 were negative in both analyses. This finding indicates that the initial gene selection was highly robust and sensitive.

Among these genes, CXCR4 is well studied in relation to IBD. The CXCR4 gene encodes GPCR CD184, which is involved in immune cell trafficking and hematopoiesis. It also interacts with CXCL12 and CXCR7 and influences cell survival, chemotaxis, and gene transcription [65]. The role of CXCR4 in immune cell migration has potential implications for gut immunity and IBD [66]. Most of the other genes have not been studied in relation to IBD, which highlights potential promising research directions for IBD.

For the transcriptomic analysis, we used the intersection of DEG results because we wanted to identify genes related to the LASSO score (and thus to the positive intestinal flora) that are also associated with the development of IBD. WGCNA serves more as an unsupervised learning method that distinguishes expression modules by constructing TOM matrices and hierarchical clustering. Genes in these modules are not necessarily differentially expressed but are closely related and so we combined these genes.

In the KEGG analysis, the top two pathways were the broad metabolic pathway and cancer pathways. These findings strongly suggest that metabolic factors play a significant role in intestinal flora-induced IBD and cancer pathways imply that intestinal flora might be a common cause of both IBD and certain gastrointestinal cancers. The PI3K-Akt pathway is relatively underappreciated in its relationship with IBD. Some studies suggest that the PI3K-Akt pathway can increase glycolysis and thereby confer mesenchymal stem cells with anti-IBD capabilities [67]. Another study indicated that *Bifidobacterium infantis* can inhibit the PI3K-Akt-mTOR pathway through PD-L1 and exert an immunosuppressive effect [68]; in our MR analysis, Bifidobacterium was shown to have a protective effect against IBD (OR<1). Overall, the role of the PI3K-Akt signaling pathway in IBD warrants further exploration.

In the GSEA Hallmark enrichment analysis, epithelial-mesenchymal transition (EMT) ranked first. In addition, numerous studies have shown that EMT can promote the development of IBD through intestinal fibrosis and damage to the mucosal barrier. Chronic inflammation promotes both IBD and EMT, which makes EMT a likely intermediate factor in the causal chain from the intestinal flora to IBD [69–72].



(caption on next page)

Fig. 5. Transcriptomic analysis. (a) Volcano plot of differentially expressed genes between the inflammatory bowel disease group and the normal group. (b) Volcano plot of differentially expressed genes between the high-LASSO-score group and the low-LASSO-score group. (c) Bubble plot of the GO analysis results. (d) Bubble plot of the KEGG analysis results. (e) Bubble plot of GSEA using the hallmark gene set. (f) Rainbow plot of CIBERSORT analysis. (G) Grouped bar chart of CIBERSORT. LASSO, least absolute shrinkage and selection operator; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, gene set enrichment analysis; ns, not significant; *, P-value<0.05; **, P-value<0.01; ***, P-value<0.001.

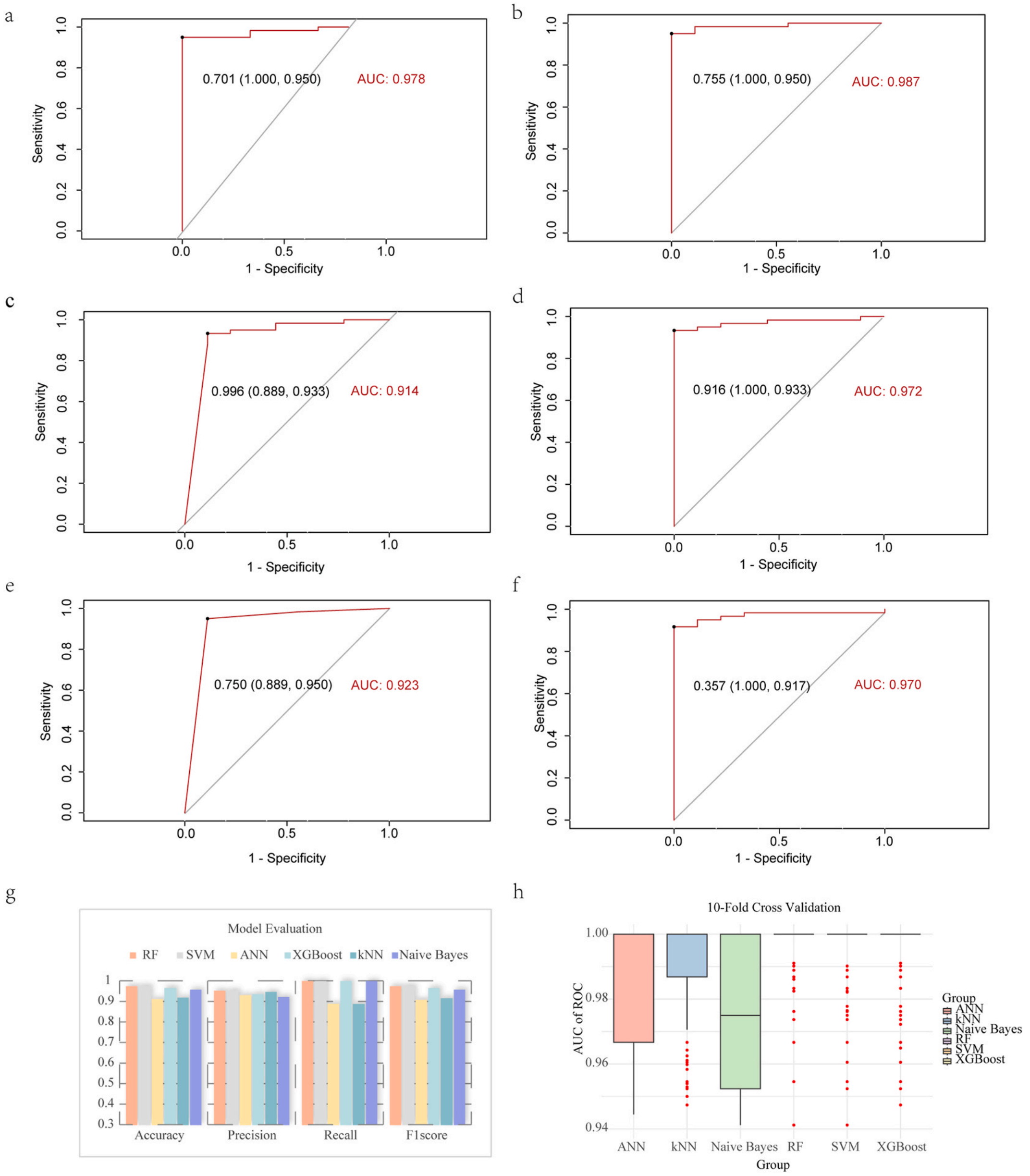


Fig. 6. Model evaluation of machine learning models. (a–f) AUC of the ROC curve for RF, SVM, ANN, XGBoost, kNN, and naive Bayes classifiers. (g) Accuracy, precision, recall and F1 score of the six models. (h) Box plot of AUC of the ROC curve in 400 times 10-fold cross-validation. AUC, area under the curve; ROC, receiver operator characteristic curve; RF, random forest; SVM, supporting vector machine; ANN, artificial neural network; XGBoost, extreme gradient boosting; kNN, k-nearest neighbor.

Notably, immune infiltration analysis revealed that neutrophils, M1 macrophages, and CD8 T cells were more enriched in the high-LASSO-score group, whereas Treg cells and M2 macrophages were more enriched in the low-LASSO-score group. The roles of these cells in IBD have long been recognized. Studies indicate that CD8$^+$ T-cell exhaustion plays a crucial role in the onset of IBD [73], have reviewed the significant role of neutrophils in intestinal flora and IBD [74], and noted that B. adolescentis affects IBD by influencing Treg cells [75]. Overall, immune infiltration results align with those of previous studies and enhance the robustness of our findings.

In the machine learning analysis, six different models were employed and all of them surprisingly demonstrated a favorable performance. The final model integration method that was chosen was soft threshold voting, which is a method that aligns well with machine learning principles. Machine learning is a popular approach in biomedical informatics research as it allows for effective data integration and analysis. The constructed models have the potential to predict the onset of IBD.

Meanwhile, these predictive models and the 24 genes they utilize hold potential clinical translational value. Further investigation of these genes could provide a deeper understanding of their roles in the development of IBD and their interactions within the human-intestinal flora. This knowledge could be valuable for developing novel therapeutics for IBD and identifying methods to modulate the intestinal flora in IBD patients. In combination with clinical testing techniques, these predictive models could aid in the early screening and prevention of IBD, which would be beneficial to the health of IBD patients.

The strengths of this study are the following. 1) Broad data sources, including IIBDGC, FinnGen database, eQTLGen, GEO database, and original literature. 2) Diverse analytical methods employing various methods at different stages were used to verify conclusions from multiple perspectives. 3) Machine learning techniques were used to construct models to predict the onset of IBD on the basis of gene expression levels. 4) Inspiring results identifying many genes and pathways with little previous research highlight potential new directions for future investigations into the relationship between the gut microbiota and IBD.

However, this study has several limitations. 1) Lack of experimental validation: although data for bioinformatics analysis were obtained from reliable public databases, experimental validation remains crucial in medical research. Furthermore, experimental evidence can significantly enhance the reliability of results. 2) eQTL data for three genes were missing, which resulted in a lack of information and discussion regarding these genes in the study. 3)In this study, the predictive model was developed based on data from two GEO databases, both of which are derived from European populations. Therefore, the model may be influenced by confounding factors such as ethnicity. Its generalizability to other populations requires further investigation. 4) These models are based on 24 genes and are not convenient for clinical testing. If further clinical translation work is to be carried out, it is necessary to further verify the functions and importance of these genes, or design a convenient and fast detection method with the support of technology.

In conclusion, this study analyzed the causal relationship between intestinal flora and IBD via two-sample Mendelian randomization. The SNPs from positive intestinal flora GWASs were mapped to corresponding genes via the g:Profiler and Plink tools. A set of genes was selected through LASSO. Further analysis of relationships between these genes and IBD was conducted via drug target MR, SMR, colocalization, and transcriptomic analyses. Finally, a disease prediction model was constructed via six machine learning methods and integrated via a soft voting method, which suggests potential translational value.

Data sharing statement

All the data can be obtained from public databases, for more information please refer to method 2.1.

Ethics statement

Data used in this study were all obtained from public databases, and all participants in the original studies signed informed consent forms. For more information, please refer to the original publications.

Funding

This study was supported by the National Natural Science Foundation of China [NSFC 82070540], the Taishan Scholars Program of Shandong Province (tsqn202211309) and the ECCM Program of Clinical Research Center of Shandong University (NO.2021SDUCRCB003).

CRediT authorship contribution statement

Guan-Wei Bi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhen-Guo Wu:** Investigation, Formal analysis, Data curation. **Yu Li:** Writing – original draft, Data curation. **Jin-Bei Wang:** Writing – original draft. **Zhi-Wen Yao:** Writing – original draft. **Xiao-Yun Yang:** Methodology. **Yan-Bo Yu:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We appreciate the data provided by IEU OpenGWAS, IIBDGC, GEO database, eQTLGen Consortium, FinnGen database, and Mbatchou J et al. We are grateful for the software support from g:Profiler and PLINK, and for the analysis and plotting assistance provided by the Sangerbox website. We thank all the researchers for their hard work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e38101.

References

[1] S.C. Ng, H.Y. Shi, N. Hamidi, et al., Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies, Lancet 390 (10114) (2017) 2769–2778.

[2] G.G. Kaplan, The global burden of IBD: from 2015 to 2025, Nat. Rev. Gastroenterol. Hepatol. 12 (12) (2015) 720–727.

[3] M. Agrawal, K.H. Allin, F. Petralia, J.F. Colombel, T. Jess, Multiomics to elucidate inflammatory bowel disease risk factors and pathways, Nat. Rev. Gastroenterol. Hepatol. 19 (6) (2022) 399–409.

[4] J. Ni, G.D. Wu, L. Albenberg, V.T. Tomov, Gut microbiota and IBD: causation or correlation? Nat. Rev. Gastroenterol. Hepatol. 14 (10) (2017) 573–584.

[5] P. Sekula, M.F. Del Greco, C. Pattaro, A. Köttgen, Mendelian randomization as an approach to assess causality using observational data, J. Am. Soc. Nephrol. : JASN (J. Am. Soc. Nephrol.) 27 (11) (2016) 3253–3265.

[6] J. Bowden, M.V. Holmes, Meta-analysis and Mendelian randomization: a review, Res. Synth. Methods 10 (4) (2019) 486–496.

[7] E. Birney, Mendelian randomization, Cold Spring Harb Perspect Med. 12 (4) (2022 May 17) a041302, https://doi.org/10.1101/cshperspect.a041302. PMID: 34872952; PMCID: PMC9121891.

[8] S. Chauquet, Z. Zhu, M.C. O'Donovan, J.T.R. Walters, N.R. Wray, S. Shah, Association of antihypertensive drug target genes with psychiatric disorders: a mendelian randomization study, JAMA Psychiatr. 78 (6) (2021) 623–631.

[9] B. Simovski, C. Kanduri, S. Gundersen, et al., Coloc-stats: a unified web interface to perform colocalization analysis of genomic features, Nucleic Acids Res. 46 (W1) (2018) W186, w93.

[10] Z. Zhu, F. Zhang, H. Hu, et al., Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets, Nat. Genet. 48 (5) (2016) 481–487.

[11] R. Zhang, H.Y. Ou, C.T. Zhang, DEG: a database of essential genes, Nucleic Acids Res. 32 (Database issue) (2004) D271–D272.

[12] A.M. Newman, C.L. Liu, M.R. Green, et al., Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (5) (2015) 453–457.

[13] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.

[14] Gene Ontology Consortium, Gene ontology Consortium: going forward, Nucleic Acids Res. 43 (Database issue) (2015 Jan) D1049–D1056, https://doi.org/10.1093/nar/gku1179. Epub 2014 Nov 26. PMID: 25428369; PMCID: PMC4383973.

[15] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (2008) 559.

[16] A. Subramanian, P. Tamayo, V.K. Mootha, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U.S.A. 102 (43) (2005) 15545–15550.

[17] J.G. Greener, S.M. Kandathil, L. Moffat, D.T. Jones, A guide to machine learning for biologists, Nat. Rev. Mol. Cell Biol. 23 (1) (2022) 40–55.

[18] N. Peiffer-Smadja, T.M. Rawson, R. Ahmad, et al., Machine learning for clinical decision support in infectious diseases: a narrative review of current applications, Clin. Microbiol. Infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 26 (5) (2020) 584–595.

[19] Y.W. Lee, J.W. Choi, E.H. Shin, Machine learning model for predicting malaria using clinical information, Comput. Biol. Med. 129 (2021) 104151.

[20] R.C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.

[21] K.J. van der Velde, F. Imhann, B. Charbon, et al., MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians, Bioinformatics 35 (6) (2019) 1076–1078.

[22] J.A. Cavanaugh, IBD international genetics Consortium: international cooperation making sense of complex disease, Inflamm. Bowel Dis. 9 (3) (2003) 190–193.

[23] M.I. Kurki, J. Karjalainen, P. Palta, et al., FinnGen provides genetic insights from a well-phenotyped isolated population, Nature 613 (7944) (2023) 508–518.

[24] J. Mbatchou, L. Barnard, J. Backman, et al., Computationally efficient whole-genome regression for quantitative and binary traits, Nat. Genet. 53 (7) (2021) 1097–1103.

[25] U. Võsa, A. Claringbould, H.J. Westra, et al., Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression, Nat. Genet. 53 (9) (2021) 1300–1310.

[26] T. Montero-Meléndez, X. Llor, E. García-Planella, M. Perretti, A. Suárez, Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling, PLoS One 8 (10) (2013) e76235.

[27] M. Vancamelbeke, T. Vanuytsel, R. Farré, et al., Genetic and transcriptomic bases of intestinal epithelial barrier dysfunction in inflammatory bowel disease, Inflamm. Bowel Dis. 23 (10) (2017) 1718–1729.

[28] T. Barrett, S.E. Wilhite, P. Ledoux, et al., NCBI GEO: archive for functional genomics data sets–update, Nucleic Acids Res. 41 (Database issue) (2013) D991–D995.

[29] V.W. Skrivankova, R.C. Richmond, B.A.R. Woolf, et al., Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration, BMJ (Clinical research ed) 375 (2021) n2233.

[30] E. Sanderson, M.M. Glymour, M.V. Holmes, et al., Mendelian randomization, Nature reviews Methods primers 2 (2022).

[31] B. Liu, L. Lyu, W. Zhou, et al., Associations of the circulating levels of cytokines with risk of amyotrophic lateral sclerosis: a Mendelian randomization study, BMC Med. 21 (1) (2023) 39.

[32] D. Ji, W.Z. Chen, L. Zhang, Z.H. Zhang, L.J. Chen, Gut microbiota, circulating cytokines and dementia: a Mendelian randomization study, J. Neuroinflammation 21 (1) (2024) 2.

[33] G. Cui, S. Li, H. Ye, et al., Gut microbiome and frailty: insight from genetic correlation and mendelian randomization, Gut Microb. 15 (2) (2023) 2282795.

[34] N.M. Davies, M.V. Holmes, G. Davey Smith, Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians, BMJ (Clinical research ed) 362 (2018) k601.

[35] P. Luo, Q. Yuan, X. Wan, M. Yang, P. Xu, A two-sample Mendelian randomization study of circulating lipids and deep venous thrombosis, Sci. Rep. 13 (1) (2023) 7432.

[36] Y. Long, L. Tang, Y. Zhou, S. Zhao, H. Zhu, Causal relationship between gut microbiota and cancers: a two-sample Mendelian randomisation study, BMC Med. 21 (1) (2023) 66.

[37] A. Kumar, C. Gupta, J. Thomas, A. Pereira, Genetic dissection of grain yield component traits under high nighttime temperature stress in a rice diversity panel, Front. Plant Sci. 12 (2021) 712167.

[38] R. Feng, M. Lu, J. Xu, et al., Pulmonary embolism and 529 human blood metabolites: genetic correlation and two-sample Mendelian randomization study, BMC genomic data 23 (1) (2022) 69.

[39] S. Burgess, A. Butterworth, S.G. Thompson, Mendelian randomization analysis with multiple genetic variants using summarized data, Genet. Epidemiol. 37 (7) (2013) 658–665.

[40] S. Burgess, S.G. Thompson, Interpreting findings from Mendelian randomization using the MR-Egger method, Eur. J. Epidemiol. 32 (5) (2017) 377–389.

[41] F.P. Hartwig, G. Davey Smith, J. Bowden, Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption, Int. J. Epidemiol. 46 (6) (2017) 1985–1998.

[42] J. Bowden, G. Davey Smith, P.C. Haycock, S. Burgess, Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator, Genet. Epidemiol. 40 (4) (2016) 304–314.

[43] J. Bowden, M.F. Del Greco, C. Minelli, G. Davey Smith, N. Sheehan, J. Thompson, A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization, Stat. Med. 36 (11) (2017) 1783–1802.

[44] M.F. Greco, C. Minelli, N.A. Sheehan, J.R. Thompson, Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome, Stat. Med. 34 (21) (2015) 2926–2940.

[45] J. Bowden, G. Davey Smith, S. Burgess, Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression, Int. J. Epidemiol. 44 (2) (2015) 512–525.

[46] G. Hemani, K. Tilling, G. Davey Smith, Orienting the causal relationship between imprecisely measured traits using GWAS summary data, PLoS Genet. 13 (11) (2017) e1007081.

[47] M.J. Brion, K. Shakhbazov, P.M. Visscher, Calculating statistical power in Mendelian randomization studies, Int. J. Epidemiol. 42 (5) (2013) 1497–1501.

[48] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Stat. Soc. B 57 (1) (2018) 289–300.

[49] M. Liu, D. Yu, Y. Pan, et al., Causal roles of lifestyle, psychosocial characteristics, and sleep status in sarcopenia: a mendelian randomization study, The journals of gerontology Series A, Biological sciences and medical sciences 79 (1) (2024).

[50] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proc. Natl. Acad. Sci. U.S.A. 100 (16) (2003) 9440–9445.

[51] L. Kolberg, U. Raudvere, I. Kuzmin, P. Adler, J. Vilo, H. Peterson, g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update), Nucleic Acids Res. 51 (W1) (2023 Jul 5) W207–W212, https://doi.org/10.1093/nar/gkad347. PMID: 37144459; PMCID: PMC10320099.

[52] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4 (2015) 7.

[53] H. Ali, M. Shahzad, S. Sarfraz, K.B. Sewell, S. Alqalyoobi, B.P. Mohan, Application and impact of Lasso regression in gastroenterology: a systematic review, Indian J. Gastroenterol. : official journal of the Indian Society of Gastroenterology 42 (6) (2023) 780–790.

[54] P.I. Johansson, H.H. Henriksen, S.T. Karvelsson, et al., LASSO regression shows histidine and sphingosine 1 phosphate are linked to both sepsis mortality and endothelial damage, Eur. J. Med. Res. 29 (1) (2024) 71.

[55] B.A. Ference, Interpreting the clinical implications of drug-target mendelian randomization studies, J. Am. Coll. Cardiol. 80 (7) (2022) 663–665.

[56] M.E. Ritchie, B. Phipson, D. Wu, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.

[57] J.S. Rhodes, A. Cutler, K.R. Moon, Geometry- and accuracy-preserving random forest proximities, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 10947–10959.

[58] H. Wang, Y. Shao, S. Zhou, C. Zhang, N. Xiu, Support vector machine classifier via L(0/1) soft-margin loss, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2022) 7253–7265.

[59] H. Shin, XGBoost regression of the most significant photoplethysmogram features for assessing vascular aging, IEEE journal of biomedical and health informatics 26 (7) (2022) 3354–3361.

[60] M. Hofmann, P. Mader, Synaptic scaling-an artificial neural network regularization inspired by nature, IEEE Transact. Neural Networks Learn. Syst. 33 (7) (2022) 3094–3108.

[61] M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, M. Valdes-Sosa, Fast Gaussian Naïve Bayes for searchlight classification analysis, Neuroimage 163 (2017) 471–479.

[62] J Wang, X Geng, Large margin weighted k -nearest neighbors label distribution learning for classification, IEEE Transact. Neural Networks Learn. Syst. 12 (2023) 1–13.

[63] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, Sci. Rep. 14 (1) (2024) 6086.

[64] O.O. Awe, G.O. Opateye, C.A.G. Johnson, O.T. Tayo, R. Dias, Weighted hard and soft voting ensemble machine learning classifiers: application to anaemia diagnosis, in: O.O. Awe, E.A. Vance (Eds.), Sustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network, vol. 2023, Cham: Springer Nature Switzerland, Ghana, 2022, pp. 351–374.

[65] M.A. Nengroo, M.A. Khan, A. Verma, D. Datta, Demystifying the CXCR4 conundrum in cancer biology: beyond the surface signaling paradigm, Biochimica et biophysica acta Reviews on cancer 1877 (5) (2022) 188790.

[66] L. Werner, H. Guzner-Gur, I. Dotan, Involvement of CXCR4/CXCR7/CXCL12 interactions in inflammatory bowel disease, Theranostics 3 (1) (2013) 40–46.

[67] C. Xu, C. Feng, P. Huang, et al., TNFα and IFNγ rapidly activate PI3K-AKT signaling to drive glycolysis that confers mesenchymal stem cells enhanced anti-inflammatory property, Stem Cell Res. Ther. 13 (1) (2022) 491.

[68] L. Zhou, Y. Xie, Y. Li, Bifidobacterium infantis promotes Foxp3 expression in colon cells via PD-L1-mediated inhibition of the PI3K-Akt-mTOR signaling pathway, Front. Immunol. 13 (2022) 871705.

[69] K. Rajamäki, A. Taira, R. Katainen, et al., Genetic and epigenetic characteristics of inflammatory bowel disease-associated colorectal cancer, Gastroenterology 161 (2) (2021) 592–607.

[70] Q. Shen, Z. Huang, J. Yao, Y. Jin, Extracellular vesicles-mediated interaction within intestinal microenvironment in inflammatory bowel disease, J. Adv. Res. 37 (2022) 221–233.

[71] S. Lovisa, G. Genovese, S. Danese, Role of epithelial-to-mesenchymal transition in inflammatory bowel disease, Journal of Crohn's & colitis 13 (5) (2019) 659–668.

[72] M. Scharl, N. Huber, S. Lang, A. Fürst, E. Jehle, G. Rogler, Hallmarks of epithelial to mesenchymal transition are detectable in Crohn's disease associated intestinal fibrosis, Clin. Transl. Med. 4 (2015) 1.

[73] A. Noble, L. Durant, L. Hoyles, et al., Deficient resident memory T cell and CD8 T cell response to commensals in inflammatory bowel disease, Journal of Crohn's & colitis 14 (4) (2020) 525–537.

[74] C. Danne, J. Skerniskyte, B. Marteyn, H. Sokol, Neutrophils: from IBD to the gut microbiota, Nat. Rev. Gastroenterol. Hepatol. 21 (3) (2024) 184–197.

[75] L. Fan, Y. Qi, S. Qu, et al., B. adolescentis ameliorates chronic colitis by regulating Treg/Th2 response and gut microbiota remodeling, Gut Microb. 13 (1) (2021) 1–17.