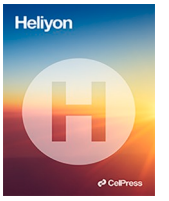
Heliyon 9 (2023) e21154

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Prediction of anti-TNF therapy failure in ulcerative colitis patients by ensemble machine learning: A prospective study

Mohammad Hossein Derakhshan Nazari [a], Shabnam Shahrokh [b], Leila Ghanbari-Maman [a,c], Samaneh Maleknia [a], Mahsa Ghorbaninejad [a], Anna Meyfour [a,*]

[a] Basic and Molecular Epidemiology of Gastrointestinal Disorders, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[b] Research Center for Gastroenterology and Liver Diseases, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[c] Department of Computer Science, Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

ARTICLE INFO

Keywords:
Anti-TNF therapy
Biomarker
Inflammatory bowel disease
Ulcerative colitis
Ensemble machine learning
Inflammation

ABSTRACT

Nowadays, anti-TNF therapy remarkably improves the medical management of ulcerative colitis (UC), but approximately 40 % of patients do not respond to this treatment. In this study, we used 79 anti-TNF-naive patients with moderate-to-severe UC from four cohorts to discover alternative therapeutic targets and develop a personalized medicine approach that can diagnose UC non-responders (UCN) prior to receiving anti-TNF therapy. To this end, two microarray data series were integrated to create a discovery cohort with 35 UC samples. A comprehensive gene expression and functional analysis was performed and identified 313 significantly altered genes, among which *IL6* and *INHBA* were highlighted as overexpressed genes in the baseline mucosal biopsies of UCN, whose cooperation may lead to a decrease in the Tregs population. Besides, screening the abundances of immune cell subpopulations showed neutrophils' accumulation increasing the inflammation. Furthermore, the correlation of KRAS signaling activation with unresponsiveness to anti-TNF mAb was observed using network analysis. Using 50x repeated 10-fold cross-validation LASSO feature selection and a stack ensemble machine learning algorithm, a five-mRNA prognostic panel including *IL13RA2*, *HCAR3*, *CSF3*, *INHBA*, and *MMP1* was introduced that could predict the response of UC patients to anti-TNF antibodies with an average accuracy of 95.3 %. The predictive capacity of the introduced biomarker panel was also validated in two independent cohorts (44 UC patients). Moreover, we presented a distinct immune cell landscape and gene signature for UCN to anti-TNF drugs and further studies should be considered to make this predictive biomarker panel and therapeutic targets applicable in the clinical setting.

1. Introduction

Ulcerative Colitis (UC) is one of the major types of Inflammatory Bowel Disease (IBD), a chronic inflammatory disorder that

* Corresponding author. Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Arabi Ave., Daneshjoo Blvd., Velenjak, Tehran, 1985717413, Iran.
E-mail address: a.meyfour@sbmu.ac.ir (A. Meyfour).

https://doi.org/10.1016/j.heliyon.2023.e21154
Received 12 July 2023; Received in revised form 5 September 2023; Accepted 17 October 2023
Available online 18 October 2023

involves the gastrointestinal tract [1,2]. It is characterized by mucosal inflammation that begins from the rectum and may extend proximally to the colon; thus, it causes severe damage to the bowel wall leading to rectal bleeding [3]. While the exact mechanism of UC pathogenesis is yet unknown, several implications such as immune system dysfunction, gut microbiota dysbiosis, genetic and epigenetic alterations, and environmental factors like smoking have been found to be involved [2,4]. Due to the nature of UC as a progressive disease ending in colorectal cancer [5–7], reliable agents are urgently required to control the disease, improve quality of life, and induce and maintain remission [2]. Although corticosteroids, 5-Aminosalicylates (5-ASA), and thiopurines like azathioprine and 6-mercaptothiopurine are the regular treatments for mild-to-moderate UC patients, treatment of corticosteroid-dependent and -refractory patients are still challenging [1], and more efficacious medications are imperative [8].

Recently, anti-Tumor Necrosis Factor (TNF) monoclonal antibodies (mAb), human or chimeric IgG1 monoclonal antibodies [2,7], have become the main strategy for treating moderate-to-severe steroid-refractory and -dependent UC patients [1,2,7]. TNF-α is an elevated cytokine in UC that contributes to mucosal barrier destruction [9]. Currently, three kinds of anti-TNF mAb, including adalimumab, infliximab, and golimumab, are clinically applied for UC patients [1,2,8]. However, mucosal healing occurs in approximately 60 % of IBD patients after anti-TNF therapy [2], and 30–40 % of cases generally do not respond to these treatments [10]. These non-responders are inevitably subjected to other therapeutic pipelines such as anti-adhesion molecules (Vedolizumab), JAK inhibitors (Tofacitinib), and anti-IL12 or/and IL13 (Ustekinumab) [1,7]. Therefore, developing a personalized medicine approach that can predict the response to anti-TNF mAbs before treatment initiation is valuable to reduce costs and prevent time consumption and disease progression [2,11,12], allowing patients to receive more appropriate first-line therapies.

Currently, pathophysiological biomarkers have failed to act as robust therapeutic response predictors in IBD. Unraveling the molecular mechanisms governing the distinctive response to anti-TNF mAb in pretreatment lesions of UC Responders (UCR) and UC Non-responders (UCN) can be a necessary step to achieve potential biomarkers predicting treatment response. Even though few studies have followed this aim in IBD [11–14], there is still no powerful discriminator panel to accurately classify the response of UC patients to anti-TNF therapy in clinical practice. Besides, understanding the working mechanisms of anti-TNF drugs and monitoring the cellular consequences resulting from these molecular differences further elucidate the pathogenesis and progression of this complex disease and leads to the identification of alternative therapeutic targets and the development of novel therapies with a different mode of action. Nowadays, the integration of high-throughput technologies with advanced bioinformatics and computational methods is increasingly applied to unify and analyze multi-cohort datasets using systems biology approaches, leading to decreased heterogeneity of publicly available single population-based gene expression datasets and the introduction of more robust diagnostic/pathogenic biomarker panels in a personalized manner [15,16].

In the present study, the publicly available gene expression profiles obtained from active lesions of UCN and UCR cases before the initiation of anti-TNF therapy (Infliximab) as discovery cohorts were integrated using Cross-Platform Normalization (CPN) and were analyzed to define a baseline signature of anti-TNF (non-) responsiveness in UC patients. Co-expression/protein-protein interaction (PPI) network construction, functional enrichment, and immune cell landscape analysis were performed to identify the high-ranked regulatory genes and related mechanisms interfering with the response to anti-TNF therapy. Furthermore, feature selection and an ensemble machine learning algorithm were employed using repeated cross-fold validation methods to introduce accurate determinants of anti-TNF response in UC patients. Then, the trained ensemble machine performance was tested on an independent *in silico* cohort. In addition, the predictive power of the introduced multi-mRNA biomarker panel for anti-TNF responsiveness throughout the designed ensemble machine was also experimentally evaluated in our patient cohort.

2. Materials and methods

2.1. Data collection and preprocessing

Using Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), the GSE12251 [11] and GSE16879 [17] generated by the Affymetrix human genome U133 plus 2.0 array chip (GPL570), and GSE73661 [18] generated by the Affymetrix Human Gene 1.0 ST array chip (GPL6244), were identified valid to be analyzed in this study. In these data series, anti-TNF naïve patients with moderate-to-severe UC were diagnosed considering the clinical symptoms and endoscopic Mayo subscore 2 or 3. Beside clinical improvement for the response evaluation, the endoscopic Mayo subscore and histologic grade $\leq$1 were evaluated to determine mucosal healing. The GSE12251 and GSE16879 were integrated (CPN method) to create a discovery cohort for further meta-analysis and the GSE73661 was excluded to be used as an independent test cohort. In this manner, the series matrix of the GSE12251 and GSE16879, and GPL570 annotation files were downloaded from GEO. To eliminate or minimize the technical variability, the *normalizeQuantiles* function in the R *Limma* package was employed to standardize the expression data using the quantile normalization method which is a popular method for microarray data preprocessing and gene expression analysis [19]. Then the *merge* function from the base package, R 4.0.3, was used to construct a metadata file, in which genes were arranged in rows as observations and samples of GSE12251 and GSE16879 were ordered in columns as variables. To alleviate the effect of operators and laboratories in which these data series were generated, the batch effect was removed using the *ComBat* function from the *sva* package in R. The ComBat function in the R language has been indicated as an effective way to remove batches and has been widely used in various research [20]. Then, this preprocessed data was considered for Differentially Expression (DE) analysis and functional annotations, including Immune cell profile analysis, Gene Ontology (GO) analysis, pathway analysis, and network analysis.

2.2. DE analysis and bioinformatics annotation

2.2.1. Immune cell landscape analysis

In order to identify distinct profiles of immune cells in responders/non-responders to anti-TNF therapy, CIBERSORTx (http://cibersortx.stanford.edu) was used. The immune cell landscape of responders and non-responders was estimated using mRNA transcription landscapes from microarray discovery datasets. The metadata file was converted into a mixture file in which patients were identified with their GEO accession numbers, and gene symbols and expression values were arranged in lines. The batch effect was corrected before cell fraction analysis using the B-mode method appropriate for microarray expression data. Analysis of the mixture file was followed using the core LM22 signature containing 547 genes that ultimately discriminate 22 hematopoietic cell phenotypes. The result was reported as relative fractions for all immune cell subtypes and graphically reported by horizontal bar charts. Statistical analysis was performed using SPSS to reveal the significant differences in the immune cell subpopulations in UCR and UCN. Firstly, the normality test, the Kolmogorov-Smirnov, and the Shapiro-Wilk tests were performed to assess the distribution of output values. According to the results, an unpaired two-tailed Mann-Whitney u-test was used to identify significant cell fraction differences.

2.2.2. DE, GO, and pathway analyses

The microarray expression profiles of UCN at baseline were compared to UCR using the R package *Limma*, which has been popular for gene expression analysis and biomarker discovery. The $|\mathrm{Log_2FC}| \geq 1$ and the False Discovery Rate (FDR) < 0.05 were applied as the cut-off for statistical significance to define Differentially Expressed Genes (DEG). To investigate the behavior of the DEGs, functional enrichment annotations, including Gene Ontology (GO) and pathways analyses were carried out using Gene Ontology Resource (GOR) (http://geneontology.org/) and Molecular Signature (MSigDB) databases (https://www.gsea-msigdb.org/gsea/msigdb), respectively. Significant Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC) were obtained using FDR <0.05 as the criteria.

2.2.3. Network construction and annotation

To discover genes likely to play an important function in the molecular processes involved in non-responsiveness to anti-TNF therapy, a network of genes that are significantly co-expressed and/or their encoded proteins have interactions was constructed using Cytoscape software (version 3.8.2). To this end, the Pearson correlation was calculated, and correlation >0.9 or < -0.9 indicated the significant positive or negative co-expressed genes. The interactions of the proteins encoded by these DEGs were explored on the Search Tool for Retrieval of Interacting Genes/Proteins (STRING) database (https://string-db.org/), and the current known interactions that have been experimentally validated were collected and elaborated to the correlation network through a binary approach (0 or 1) as the PPI value. The *CentiScape2.2* plugin was used to evaluate degree eccentricity, closeness, stress, radiality, and betweenness [21]. The degree is the total number of edges to a node in a network and can be interpreted to identify the most interactive protein. The eccentricity of a protein in a network refers to the ease with which it can influence or be reached by other proteins. Eccentricity for each node is calculated through the following formula (Equation (1)), where v and w refer to a couple of nodes:

$$C_{ecc}(v) = \frac{1}{\max\{dist(v,\omega) : \omega \in N\}} \quad (1)$$

Closeness is the sum of the shortest path between each couple of nodes (v, w) in a network and is calculated with equation (2):

$$C_{clo}(v) = \frac{1}{\sum dist(v,\omega)_{\omega \in N}} \quad (2)$$

In biological networks, proteins with a high relative closeness to the network will be central to the regulation of other proteins. Likewise, a protein with high radiality compared to the average radiality of the network is more likely to be a central node. Radiality is measured through the following formula in which ΔG (diameter) is the maximum distance among all the values for the distance between each vertex in graph G (Equation (3)):

$$C_{rad}(v) = \frac{\sum_{\omega \in N}(\Delta G + 1 - dist(v,\omega)}{n-1} \quad (3)$$

The higher the radiality, the closer the node is to the other nodes concerning its diameter. A protein's functional role in connecting nodes in a biological network is determined by its stress sensitivity and betweenness centrality. Stress is calculated by determining the number of shortest paths passing through a node in a graph using the below formula (Equation (4)):

$$C_{str}(v) = \sum_{s \neq v \in N} \sum_{t \neq v \in N} \sigma st(v) \quad (4)$$

While betweenness is computed by dividing the number of shortest paths between v_1 and v_2 that pass through a node n by the total number of shortest paths linking v_1 and v_2 by the following equation (Equation (5)):

$$C_{str}(v) = \sum_{s \neq v \in N} \sum_{t \neq v \in N} \delta st(v) \quad (5)$$

$$\delta st(v) = \frac{\sigma st(v)}{\sigma st}$$

According to these indexes, the *CytoHubba* plugin detects models of the top 100 genes. Then, the final network was built using the overlapping genes in these models. Lastly, pathway analysis was applied to the final network to discover the critical pathways involved in anti-TNF unresponsiveness.

2.3. Identification of predictive biomarkers

To discriminate responders from non-responders to anti-TNF and identify accurate predictors of drug responsiveness, $|\mathrm{Log_2FC}| \geq 2$ was determined to filter DEGs. After the filtration, by using the *cv. glmnet* function from the *glmnet* package in R, the Least Absolute Shrinkage and Selection Operator (LASSO), which can identify specific genes for different samples, was employed with the tuning parameter (λ) estimated using 10-fold cross-validation. This algorithm was repeated 50 times to identify more robust features having LASSO absolute coefficient ≥ 1 in each run. The order of features was then determined based on the Average LASSO Absolute Coefficient (ALAC) to design input gene sets for classification using the n + 1 formula ($n \geq 2$).

2.4. Machine learning approach

In order to achieve an accurate performance estimate, avoid obtaining biased results, and minimize overfitting, a two-layer stack ensemble machine was modeled by recruiting several highly-applied classifiers, including Linear Discriminant Analysis (LDA), Neural Network (NN), Random Forest (RF), and K-nearest Neighbor (KNN) at the first layer, and in the second layer, a Generalized Linear Model (GLM) algorithm was designed based on probabilities obtained from the first-layer machines for each sample using the R packages *caret* and *caretEnsemble*. This ensemble model operates through the below function (Equation (6)):

$$glm_{ensemble} = aLDA \ \pm \ bRF \ \pm cKNN \ \pm dNN \tag{6}$$

The *a*, *b*, *c*, and *d* factors are coefficients of the single machines in the ensemble model. If the outcome of the above equation is positive, the patient is predicted to be UCN. The power of this ensemble machine was measured by the mean accuracy, sensitivity, and specificity for each gene set by using the following formula (Equation (7)):

:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \tag{7}$$

$$Sensetivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Finally, the Receiver Operating Characteristic (ROC) curve was constructed for each gene set, and the Area Under Curve (AUC) was calculated to visualize the predictive performance of the models using the R package *pROC*.

2.4.1. In silico panel development

At first, the performance of the ensemble machine for each gene set was analyzed on the discovery cohort through a 10x repeated 10-fold cross-validation method. In this manner, first, discovery samples were divided into ten folds. Each time, the stack ensemble machine was trained by nine folds and tested on the 10th fold. This process was repeated ten times by new randomly resampled folds to obtain more accurate and reliable results. Afterward, to further test the model performance for each gene set, the stack ensemble machine was trained with all discovery samples and then tested on the independent test cohort (GSE73661). Finally, the ROC curve was constructed, and AUC was calculated to visualize the predictive potentiality of the panels.

2.4.2. Experimentally panel development

In order to explore the predictive potentiality of our multi-mRNA panel and also further evaluate the performance of the stack ensemble machine in patients' classification based on possible response to adalimumab; another type of anti-TNF mAb, a real-life patient's cohort at the gastrointestinal and liver diseases clinic of Taleghani hospital, associated with Shahid Beheshti University of Medical Sciences, was employed. In this cohort, UC patients were diagnosed with moderate-to-sever active disease, had an endoscopic Mayo subscore of 2 or 3, were refractory/dependent to corticosteroids, and had to be naïve for anti-TNF agents. All patients received 160 mg, 80 mg, and 40 mg of adalimumab biosimilar; CinnoRa (CinnaGen, Tehran, Iran), on weeks 0, 2, and 4, respectively. After the 4th week, participants were treated with 40 mg dose every two weeks until the response evaluation. Patients were evaluated after 14 weeks and six months by a gastroenterologist, and those who had improved clinical symptoms, achieved complete mucosal healing with endoscopic Mayo subscore 0 or 1, and the histologic grade of 0 or 1 was determined as the responders to anti-TNF therapy. Subsequently, mucosal biopsies were taken colonoscopically from the active lesions at the baseline before the first treatment and were snap-frozen in liquid nitrogen before storage at −80 °C. Based on the manufacturer's guidelines, the total RNA was isolated from the

mucosal biopsies using TRIzol reagent (Qiagen, USA). Before cDNA synthesis, RNA quality and concentration were determined using the Denovix-DS11 Spectrophotometer/Fluorometer (DeNovix, USA). According to the manufacturer's guidelines, the cDNA was synthesized using the cDNA synthesis kit (Parstous, Tehran, Iran) and stored at −20 °C. Then, quantitative real-time reverse transcription PCR (qRT-PCR) was performed to verify the expression changes of the target genes as well as potential predictive biomarkers using a Rotor-Gene 6000 real-time PCR cycler (Qiagen Corbett, Hilden, Germany). The primers' sequences manufactured by the Pishgam (Tehran, Iran) were shown in Supplementary Table 1. The data were normalized to the housekeeping *GAPDH* gene, and the FC was calculated according to the $2^{-\Delta\Delta CT}$ formula. Then, Kolmogorov-Smirnov and the Shapiro-Wilk tests were applied to analyze the normality of the output data. UCN and UCR groups were compared using the Mann-Whitney u-test, and $p < 0.05$ was considered statistically significant. The data were expressed by the ±standard error of the mean (SEM) and analyzed by the GraphPad Prism. Next, the ensemble machine learning approach was applied through the same 10x repeated 10-fold cross-validation algorithm to the expression profiles of biomarkers in qPCR, and the performance was reported using accuracy, sensitivity, specificity, and AUC which was calculated by ROC curve construction.

3. Results

3.1. Study design and patient information

In this study, the pretreatment mucosal biopsies were analyzed from a total of 79 corticosteroid-refractory or -dependent anti-TNF naïve UC patients who were diagnosed with moderate-to-sever active diseases confirmed by the Mayo subscore 2 or 3. Among them, 31 UC patients were detected as responders according to clinical symptom improvement, endoscopic activity, and histologic scores. Among 79 UC patients, 57 cases from GSE12251, GSE16879, and GSE73661, who were subjected to receive infliximab, were screened via *in silico* approaches, and 22 were monitored experimentally. To find proper datasets, eight keywords, including inflammatory bowel disease, IBD, ulcerative colitis, colitis, anti-TNF, adalimumab, golimumab, and infliximab, were explored in GEO and ArrayExpress databases. The gene expression profiles of UC patients in GSE12251 and GSE16879 were integrated as a discovery cohort (Supplementary Fig. 1 and Supplementary Table 2) to perform functional enrichment analysis, identify potential biomarkers, and develop an ensemble machine learning approach that can predict the response to anti-TNF therapy. Moreover, the colitis samples in GSE73661 (Supplementary Table 3) were used as an independent test cohort to evaluate the designated ensemble machine's performance. To further examine the predictive power and reproducibility of findings, the panel was also experimentally evaluated using gene expression data of the mucosal biopsies obtained from 22 patients before receiving adalimumab at Taleghani hospital (Table 1). The workflow of the study is illustrated in Fig. 1.

3.2. Accumulation of neutrophils and reduction of Tregs in the inflamed tissue of non-responders to anti-TNF treatment

In order to estimate the contribution of distinct immune cell types to anti-TNF unresponsiveness, CIBERSORTx was used to determine the global immune cell landscape of UCN and UCR patients before anti-TNF treatment by deconvolution of gene expression microarray data (Fig. 2A and Supplementary Table 4). It was observed that neutrophils were more infiltrated in the inflamed region of non-responders relative to responders (Fig. 2B), while T regulatory cell populations were substantially reduced (Fig. 2C). Moreover, no significant differences were found between other immune cell subsets of UCR and UCN.

Table 1
Clinical and demographic characteristics of real-life patient cohort at Taleghani hospital.

Clinical Factors	UCN (n = 10)	UCR (n = 12)
Sex (Male/Female)	7 Males/3 Females	6 Males/6 Females
Disease Duration, Median years (IQR)	3 (0.5–9)	5.5 (0.7–16)
Age, Median years (IQR)	35 (18–58)	32.5 (18–62)
Stool Calprotectin, mg/kg, (IQR)	2655.4 (1091–3048.4)	2502 (444–2960.6)
Erythrocyte Sediment Rate, mm/hr, (IQR)	30 (13–65)	25 (5–60)
C-Reactive Protein, mg/dl, (IQR)	17.5 (11–96)	12 (8–47.2)
Background Disease	1 Patient	2 Patients
Hb, g/dl, (IQR)	12.7 (6.8–14)	13.2 (8.9–15.8)
WBC, × 1000/mm^3, (IQR)	7.8 (6.2–9.2)	8.4 (6.6–13.6)
Vitamin D, ng/ml, (IQR)	18.5 (12–36)	29.5 (14.5–54)
Smoking History	1 Patient	1 Patient
Alcohol History	1 Patient	1 Patient
Drug Resistance	3 CS Refractory	4 CS Refractory
	7 CS Dependent	8 CS Dependent

*IQR: Inter Quartile Range.
*UCN: Ulcerative Colitis Non-responder (to anti-TNF therapy).
*UCR: Ulcerative Colitis Responder (to anti-TNF therapy).
*CS: Corticosteroids.

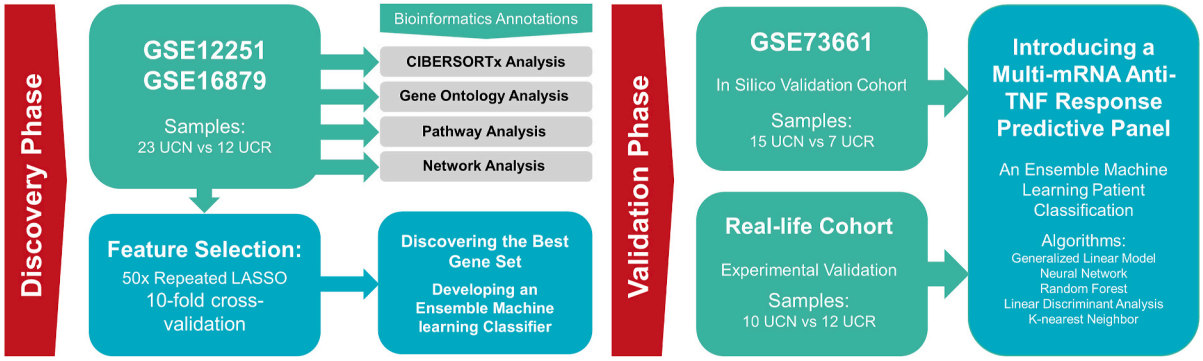


Fig. 1. The workflow. The pretreatment biopsies of 35 ulcerative colitis patients were analyzed in a discovery cohort to identify alternative therapeutic targets and predictive biomarkers of anti-TNF response using feature selection and machine learning algorithms. In addition, an independent *in silico* cohort with 22 UC samples was used to test the predictive performance of multi-mRNA gene sets. Finally, the introduced anti-TNF response predictive panel was experimentally validated on a real-life cohort. UCR: Ulcerative Colitis Responders and UCN: Ulcerative Colitis Non-responders (to anti-TNF therapy).

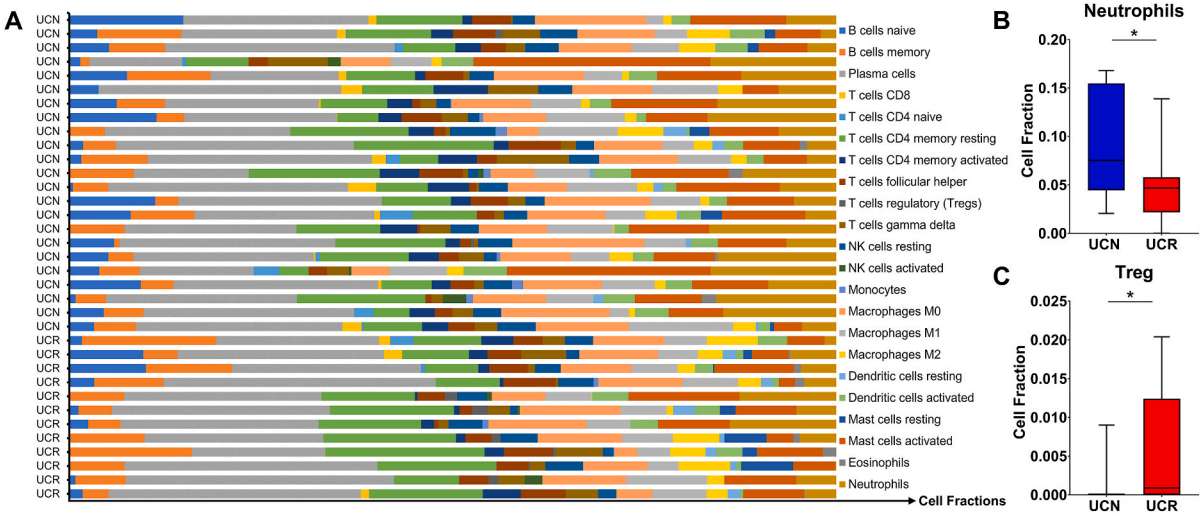


Fig. 2. Immune cell subpopulation analysis in UC patients prior to anti-TNF therapy. (A) The immune cell landscape of UC patients prior to commencing anti-TNF therapy was identified using the CIBERSORTx database. **(B)** Neutrophil accumulation and **(C)** T cell regulatory reduction were significantly observed in UCN patients. UCR: Ulcerative Colitis Responders and UCN: Ulcerative Colitis Non-responders (to anti-TNF therapy).

3.3. Identification of statistically significant altered genes

Using the discovery cohort that was previously created by preprocessing, removing batch effect, and integrating GSE12251 and GSE16879, gene expression profiles of UCN were compared to UCR, and 313 DEGs were identified using $|\mathrm{Log_2FC}| > 1$ and FDR <0.05 as the criteria, among which 282 genes were upregulated while the expression levels of 31 ones decreased. Among the DEGs, the mRNA levels of *CSF3* and *TFPI2* were detected more than nine times in UC baseline tissues that belong to anti-TNF non-responders. Furthermore, more than six-fold changes in the expression of *IL11*, *INHBA*, *PROK2*, *IL24*, *PI15*, and *IL13RA2* were observed in these patients. On the other hand, *CHP2*, *BEST4*, *GUCA2B*, and *GUCA2A* were illustrated to be suppressed at least three times in UCN compared to the UCR mRNA profiles (Fig. 3A). A correlation heatmap coupled with hierarchical clustering was constructed based on the expression values of the 313 altered genes to show the similarity of samples (Fig. 3B). We noticed a remarkable correlation and a high degree of consistency among intragroup samples. However, few UCN samples whose responses to the drug had been assessed at week 4 or 6 were placed among UCR which could be due to too early response evaluation or disease severity.

3.4. IL6, the key protein in UC tissue lacking response to anti-TNF treatment

Enrichment analysis was conducted to understand the biological actions of genes whose expression levels are associated with anti-TNF responsiveness. GO analysis in GOR indicated secretory granule membrane, cytokine activity, and cytokine-mediated signaling

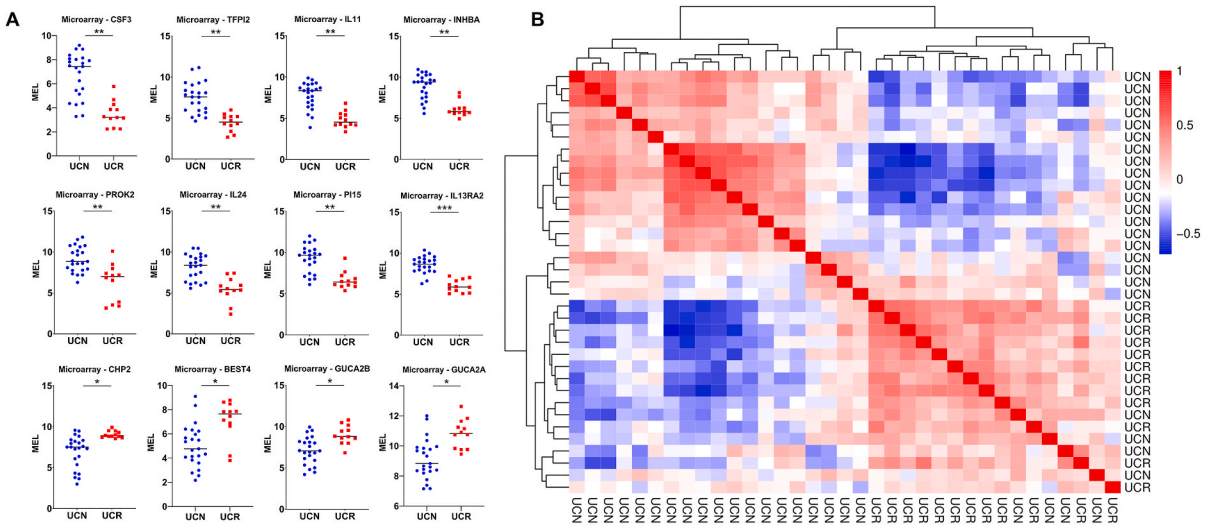


Fig. 3. Distinct gene expression profiles between UCN and UCR patients. (A) The most significant differentially expressed genes in UCN and UCR patients (*, **, and *** refer to the adjusted *p*-value <0.05, <0.01, and <0.001, respectively). **(B)** Pearson correlation analysis, along with the hierarchical clustering of the 313 differentially expressed genes, revealed a remarkable correlation among intragroup samples. UCR: Ulcerative Colitis Responders and UCN: Ulcerative Colitis Non-responders (to anti-TNF therapy).

pathway as hallmarks for CC, MF, and BP, respectively (Fig. 4A). Subsequently, signature pathway analysis demonstrated that inflammatory response and TNF-α signaling via NF-kB are two critical pathways involved in unresponsiveness to anti-TNF mAb (Fig. 4B). Among all DEGs, *IL6* was found in the most enriched pathways (Supplementary Table 5 and Table 6). Analyzing the microarray mRNA

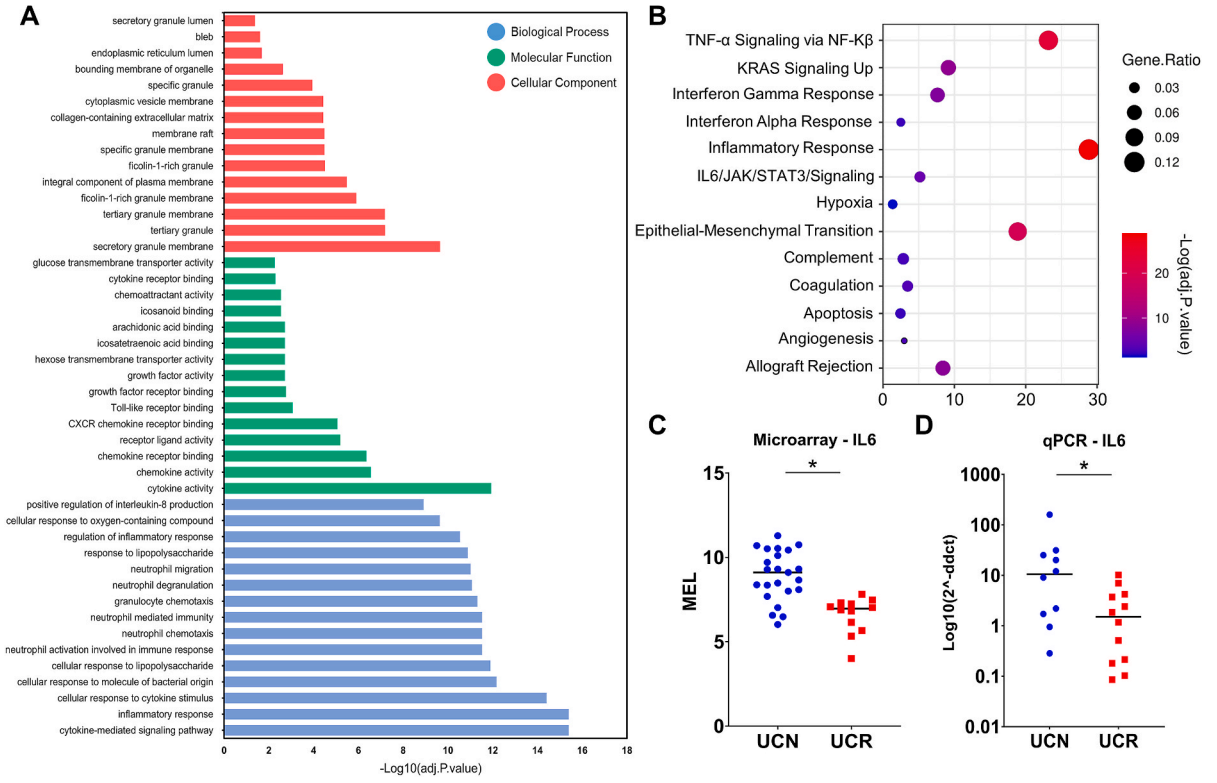
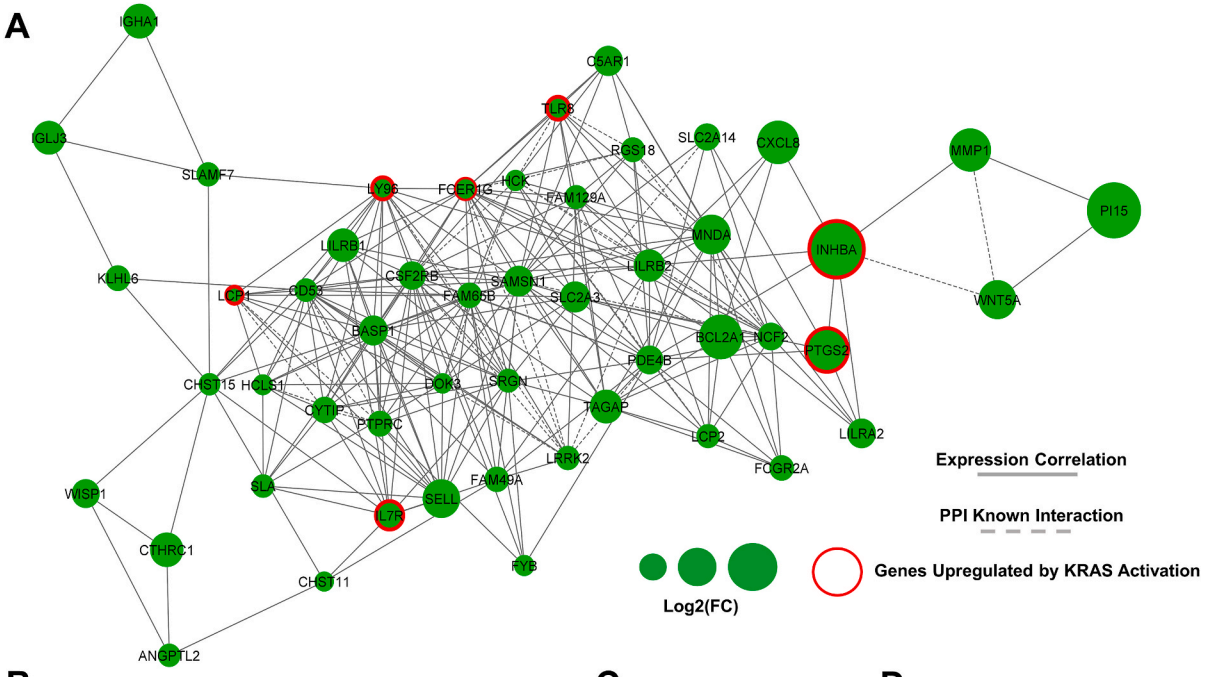


Fig. 4. Functional enrichment analysis. (A) Enriched gene ontologies and **(B)** pathways in biopsies of anti-TNF non-responders using GOR and MSigDB databases, respectively. **(C)** Microarray and **(D)** qRT-PCR analysis showed the overexpression of IL6 in UCN patients. The * refers to adjusted *p*-value <0.01, and *p*-value <0.05. UCR: Ulcerative Colitis Responders and UCN: Ulcerative Colitis Non-responders (to anti-TNF therapy), MEL: Microarray Expression Level.

profiles of discovery patients showed a 5-fold increase in *IL6* expression in anti-TNF non-responders (Fig. 4C). Likewise, the over-expression of *IL6* in our patient cohort was experimentally analyzed using qRT-PCR and was confirmed ($log_2FC = 4.07$, $p < 0.05$) (Fig. 4D). Furthermore, GO analysis illustrated that IL6 is involved in the cytokine-mediated signaling pathway (BP) as a protein with cytokine activity (MF). Regarding these outcomes, the upregulation of IL6 as a cytokine may contribute to the lack of response to anti-TNF mAb therapy in UC patients.

3.5. Upregulation of INHBA and KRAS signaling in non-responsive patients

A co-expression-based biological network, which also included protein interactions, was constructed in Cytoscape, containing 154 nodes communicated through 716 interactions. The models of the top 100 nodes were then constructed based on degree, radiality, and centrality factors using *CentiScape2.2* and *CytoHubba* plugins. By intersecting these models, 51 essential genes were represented with 258 interactions, containing co-expression and PPI (Fig. 5A). In this network, SAMSN1 and CSF2RB were illustrated as key genes due to the highest number of interactions and their central role (Supplementary Table 7). Pathway analysis was applied to the final network using MSigDB to discover the critical pathways contributing to the anti-TNF treatment failure (Fig. 5B). The involvement of over-expressed INHBA, PTGS2, IL7R, LCP1, LY96, FCER1G, and TLR8 genes in the network as KRAS downstream targets indicates the remarkable activation of KRAS signaling in non-responders. Also, the inflammatory response was manifested as another critical pathway in UCN patients. According to enrichment analysis, *INHBA,* a member of the TGF-beta (transforming growth factor-beta)



Pathways	Gene Ratio
KRAS Signaling Up	0.137254902
Inflammatory Response	0.137254902
TNF-alpha Signaling via NF-kB	0.117647059
Epithelial Mesenchymal Transition	0.117647059
Allograft Rejection	0.098039216
Interferon Gamma Response	0.098039216

Fig. 5. Systematic network construction and analysis. (A) The network of crucial genes was constructed using CytoHubba and CentiScape2.2 plugins, containing 51 genes that were communicated through 258 interactions. **(B)** and **(C)** represent the KRAS expression among UCR and UCN patients in both discovery and real-life validation cohorts, respectively. No significant differences were identified. UCR: Ulcerative Colitis Responders and UCN: Ulcerative Colitis Non-responders (to anti-TNF therapy), MEL: Microarray Expression Level.

superfamily of proteins, was found to be involved in both KRAS activation and inflammatory response pathways highlighting its possible role in unresponsiveness to anti-TNF mAb treatment. These findings suggest a correlation between the severe inflammation in non-responders and their resistance to anti-TNF therapy and KRAS-induced function, which may be the output of overexpression or point mutations of the *KRAS* gene. Therefore, the expression level of this gene was examined in the *in silico* discovery cohort, revealing that KRAS was not significantly overexpressed among non-responders (Fig. 5C). To ensure this finding, the expression of KRAS was evaluated in our patient cohort by qRT-PCR, and no significant change in KRAS expression was observed between the two groups. (Fig. 5D). Hence, the potential role of point mutations in KRAS overactivation and anti-TNF unresponsiveness is suggested to be considered in further studies.

3.6. Identification of a multi-mRNA panel to predict anti-TNF response

To nominate specific mRNAs that could predict the response of UC patients to anti-TNF therapy, different cut-offs of adjusted *p*-value and FC were selected and applied to gene expression data (data not shown). Principle component analysis (PCA) showed that $|\mathrm{Log_2FC}| \geq 2$ could be the best cut-off by which 31 genes were detected to able to properly separate UCR from UCN patients of the discovery cohort (Supplementary Fig. 2 and Supplementary Table 8). Then, a 50x repeated 10-fold cross-validation LASSO algorithm was employed to lessen these 31 signature genes and find accurate predictors of anti-TNF therapy failure. In this regard, five upregulated genes were identified with the ALAC value ≥ 1, including *IL13RA2*, *MMP1, HCAR3*, *CSF3*, and *INHBA* (Table 2).

IL13RA2 was the most powerful indicator of the UCN class that was observed in 49 times run of 10-fold cross-validation LASSO with ALAC of 3.577. According to the ALAC value, different combinations of 3–5 gene sets were developed based on the n + 1 formula (top 3, top 4, and top 5 features). Then, a stack ensemble machine learning approach was used, in which in the first layer LDA, NN, KNN, and RF operated to generate probabilities for the second layer in which a GLM machine operated. At first, the ensemble machine functioned in a 10x repeated 10-fold cross-validation manner to monitor the performance of three-, four-, and five-mRNA panels in predicting the response to infliximab therapy in discovery cohort patients. For all gene sets, the GLM indicated higher accuracy and reliability compared to the single-machine classifications (Fig. 6A). Having mentioned that, GLM showed the five-mRNA panel with more powerful and reproducible predictive potentiality with an accuracy of 97 % (sensitivity 97 %, specificity 95 %, AUC of 0.991), compared to four - and three-biomarker sets with accuracies of 89 % (sensitivity 91 %, specificity 86 %, AUC of 0.969) and 92 % (sensitivity 94 %, specificity 88 %, AUC of 0.982), respectively (Fig. 6B). In order to further evaluate the predictive performance of three-, four-, and five-gene sets, the ensemble machine was trained using all of the samples in the discovery cohort and then tested on an independent cohort (GSE73661) with 22 infliximab pretreatment mucosal biopsies. The five-featured panel was consistently observed to predict the response to anti-TNF with higher accuracy of 95 % (sensitivity 100 %, specificity 88 %, AUC of 0.981), in comparison with four - and three-mRNA panels which showed accuracies of 73 % (sensitivity 85 %, specificity 56 %, AUC of 0.819) and 77 % (sensitivity 86 %, specificity 62 %, AUC of 0.848), respectively (Fig. 6C).

3.7. Experimental validation of the predictive performance of mRNA panels in a real-life patient cohort

To authenticate the expression alterations and discriminative potentiality of the predictive biomarkers, qRT-PCR was employed for mucosal biopsies of patients referred to Taleghani hospital. At first, the statistically significant higher expression of *IL13RA2*, *HCAR3*, *CSF3*, *INHBA*, and *MMP1* was verified in active lesions of non-responsive UC patients (Fig. 7A). To explore the capacity of gene sets for predicting the response to adalimumab, another type of anti-TNF mAb, the ensemble classifier was used through a 10x repeated 5-fold cross-validation algorithm. The heterogeneity in the type of anti-TNF mAb in the *in silico* and experimental cohorts can help identify a more robust and accurate panel. Anticipatedly, compared to each single machine learning method, the ensemble showed higher accuracy (Fig. 7B) which is in line with previous results in this study. The GLM findings indicated the five-featured set as the optimal predictive panel with an accuracy of 94 % (sensitivity 89 %, specificity 98 %, AUC of 0.948). The accuracy of three- and four-featured gene sets were 91 % (sensitivity 89 %, specificity 93 %, AUC of 0.964) and 79 % (sensitivity 76 %, specificity 81 %, AUC of 0.878), respectively, which are not promising as the five-mRNA panel (Fig. 7C).

Table 2
Five robust predictive biomarkers are achieved through 50x repeated 10-fold cross-validation LASSO.

Gene	50x-AC	ALAC	$\mathrm{Log_2FC}$	FDR
IL13RA2	49	3.577101	2.598922418	2.23E-05
MMP1	43	3.082295	2.032687021	0.01018
HCAR3	34	2.007405	2.44543932	0.012008
CSF3	35	1.424827	3.214592423	0.002381
INHBA	9	1.1094	2.793163178	0.001794

*50x-AC: 50x Appearance Number.
*ALAC: Average LASSO Absolute Coefficient.
*FC: Fold Change.
*FDR: False Discovery Rate.

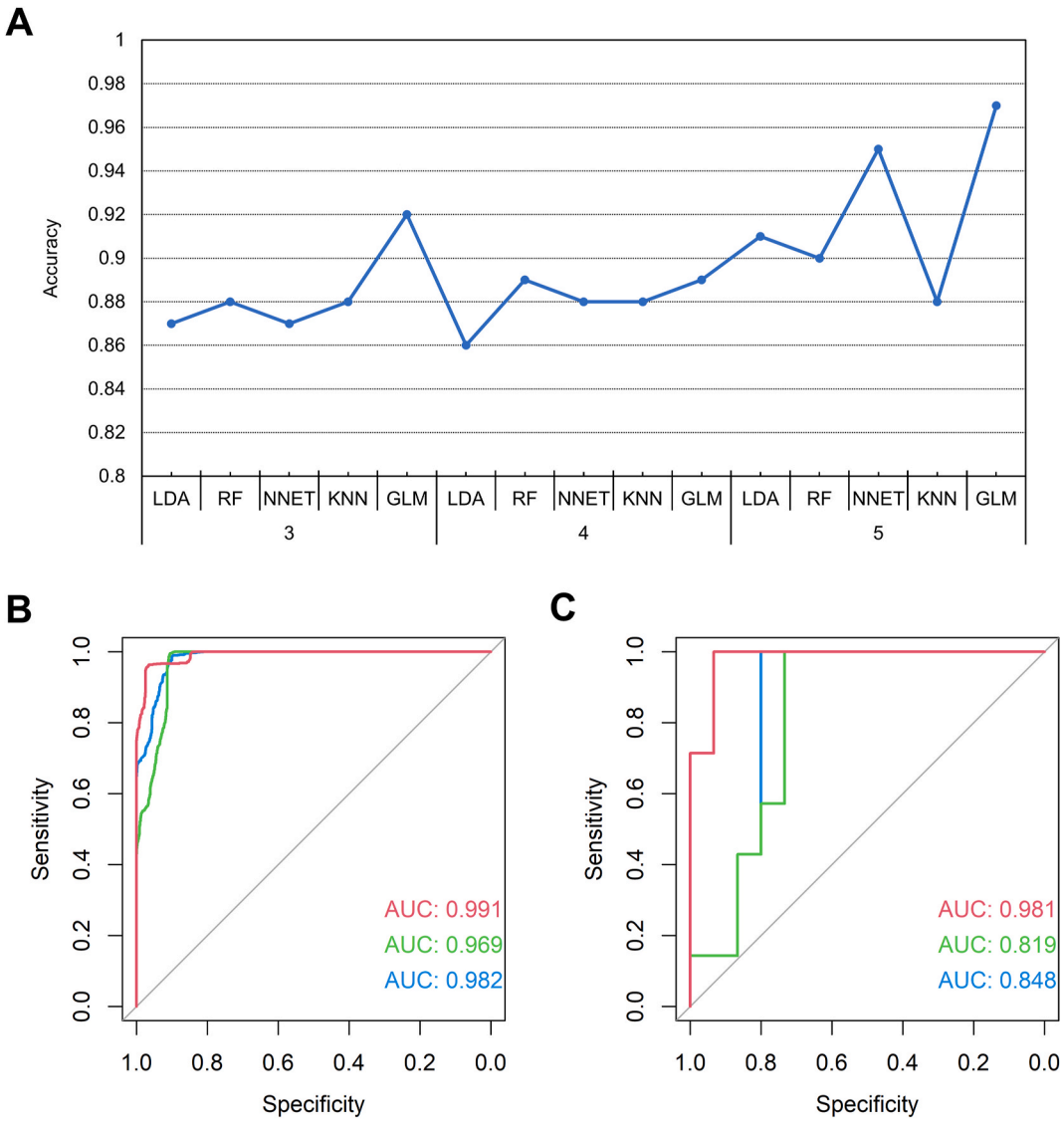


Fig. 6. Predictive potentiality of three-, four-, and five-mRNA panels in microarray datasets based on accuracy and ROC curve using the ensemble machine learning. (A) The accuracy for every single machine and the ensemble machine for three-, four-, and five-featured panels showed improved performance for the ensemble machine learning approach compared to single machines in the discovery cohort. The ROC curve and AUC of the three (blue line), four- (green line), and five-featured (red line) panels in the **(B)** discovery and **(C)** *in silico* impendent test cohorts. The AUC was shown in the plot. ROC: Receiver Operating Characteristic curve. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4. Discussion

Although anti-TNF mAbs are considered efficient medications for treating UC, many cases do not respond to them, and the inflammation progresses in this condition [10]. Therefore, it is imperative to diagnose non-responders before treatment initiation and subject them to other therapeutic guidelines to avoid time and cost consumption. In the present study, the CPN method was followed by which transcriptome profiles of two *in silico* cohorts from two different hospitals, including 35 infliximab naïve UC biopsies, were integrated and considered as a discovery cohort to identify alternative therapeutic targets and predictive biomarkers. Integrating datasets using the CPN method has been shown not only to improve sample size and overall heterogeneity and reduce the individual study bias [22] but also lead to the identification of more specific biomarkers, as integrating genomics data has been proven to boost clinical diagnosis and precision medicine [23]. At first, the immune cell landscape of UCN and UCR patients was analyzed by CIBERSORTx which showed the significant accumulation of neutrophils and reduction of Tregs in UCN patients. Similarly, an increase in neutrophils subpopulation has been reported in UC inflamed mucosa compared to uninflamed and healthy intestinal tissues [24]. Neutrophils are the cells known to increase intestinal inflammation and tissue damage in UC [25]. Not only has the essentiality of

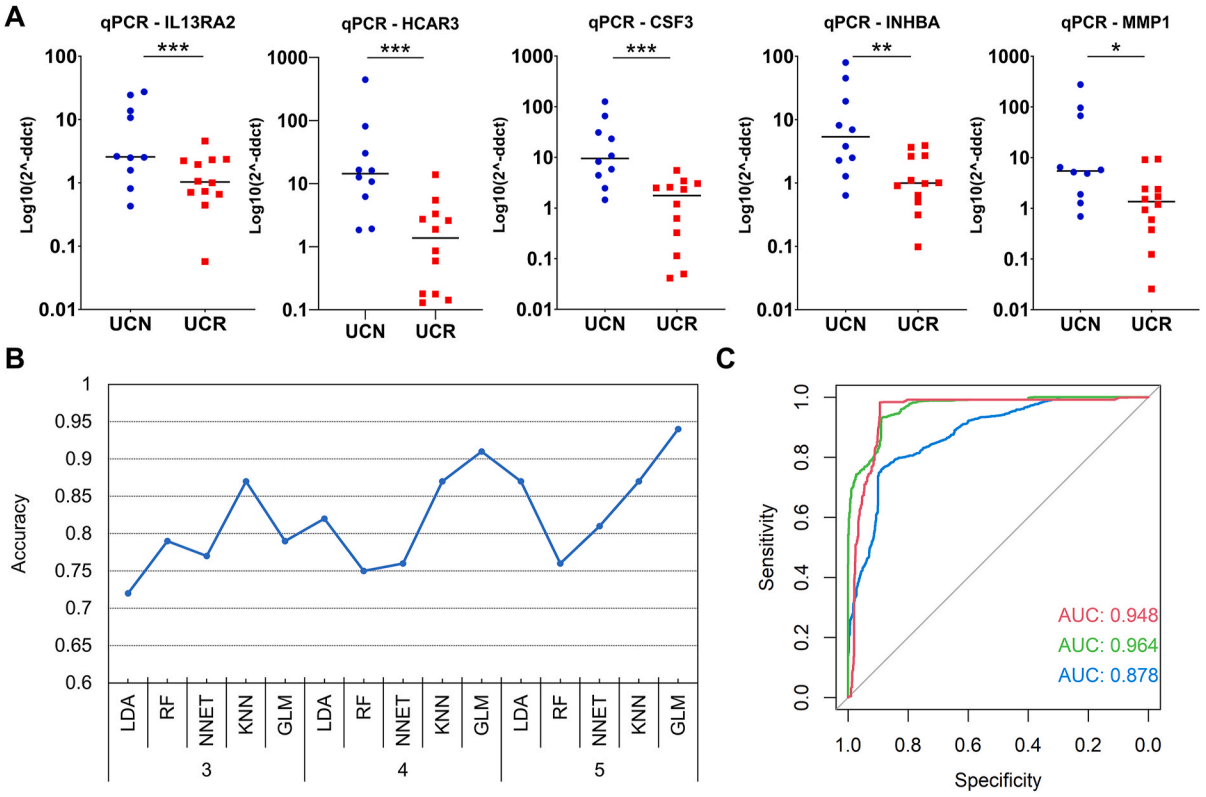


Fig. 7. Real-life cohort evaluation of predictive biomarkers and the ensemble machine learning method. (A) Experimental validation of the predictive biomarkers in the real-life cohort. The expression levels of biomarker candidates were measured in UCN and UCR patients of Taleghani hospital by RT-qPCR. *, **, and *** refer to the $p < 0.05$, <0.01, and <0.001, respectively. **(B)** The accuracy for the single machines and the ensemble machine for three-, four-, and five-featured panels showed improved performance for the ensemble machine learning approach compared to single machines in the real-life cohort. **(C)** The ROC curve for the three- (blue line), four- (green line), and five-mRNA (red line) gene sets for the discrimination of UCN from UCR patients in the real-life validation. The AUC was shown in the plot. ROC: Receiver Operating Characteristic curve. UCR: Ulcerative Colitis Responders and UCN: Ulcerative Colitis Non-responders (to anti-TNF therapy). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

neutrophils in UC pathogenesis been proven [25], but also previous research demonstrated that anti-TNF therapy reduces the neutrophil infiltration into the inflamed mucosa of IBD patients [26]. Likewise, the accumulation of neutrophils has been recognized to intensively correlate with more severe inflammatory conditions in UC [25]. Interestingly, the decreased population of Treg cells reinforces neutrophile function to augment the inflammation in the colon mucosa since Treg cells negatively regulate inflammation in different conditions [27].

To identify the molecular mechanisms involved in non-responders prior to anti-TNF therapy, gene expression analysis was performed using baseline discovery mucosal samples, and 313 significant DEGs were detected. Functional annotation revealed that secretory granule membrane, cytokine activity, and cytokine-mediated signaling pathway were hallmarks of CC, MF, and BP, respectively. Furthermore, inflammatory response and TNF-α signaling via NF-kB were identified as two remarkable pathways involved in the unresponsiveness to anti-TNF therapy, which may be in part due to disease severity and duration of inflammation in UCN patients as reported in clinical studies [28,29]. IL6, a cytokine that is secreted at sites of inflammation, was found to be a common member among most of the enriched pathways associated with anti-TNF unresponsiveness. IL6 is a mediator of fever and acute phase immune responses [30], playing major roles in the survival and differentiation of T cells, cytokine production, and regulation of Th17/Treg balance in a dose-dependent manner [31]. The involvement of IL6 in the pathophysiology of IBD with a profound influence on the disease progression has been demonstrated in numerous previous studies [32], one of which has reported that the expression level of IL6 as the predominant central cytokine significantly elevates and controls the pro-inflammatory cytokine and immune responses in UC [33]. In line with the well-recognized role of IL6 in the inflammation progression, it seems that IL6 overexpression would be a main regulatory factor in anti-TNF therapy failure.

In addition, the construction of a network based on the expression correlation and experimentally confirmed protein interactions illustrated that KRAS activation and inflammatory response were crucial enriched pathways in non-responders to anti-TNF therapy. Inhibin-β A or INHBA, which had the highest expression change amongst other network genes, was determined to be involved in KRAS activation and inflammatory response, suggesting its possible role in anti-TNF unresponsiveness. INHBA is a member of the TGF-β superfamily [34], known as pleiotropic cytokines, which have both anti- and pro-inflammatory activities, depending on the cellular

and environmental condition. It has been reported that blockade of TGF-β signaling in both TGF-β1 knockout and T cell-specific TGF-β receptor II knockout mice resulted in the development of severe autoimmunity, highlighting its modulatory effect on T cells' activities [35,36]. Furthermore, TGF-β induces peripheral tolerance by maintaining the survival of naïve T cells and inhibiting the proliferation and differentiation of $CD4^+$ and $CD8^+$ self-reactive T cells [37]. However, in the presence of inflammation and overexpressed IL6, TGF-β acts differently and leads to an increase in T helper 17 (Th17) cell differentiation, reduction of Treg cells, further inflammation [38], and exacerbation of autoimmune conditions [37]. Our findings are in accordance with these studies, and it seems that INHBA, a regulator of the TGF-β pathway, in cooperation with IL6, would increase the severity of the inflammation in the mucosa of UCN patients by reducing the population of Treg cells, which ultimately leads to anti-TNF therapy failure. Future mechanistic studies should be performed to further investigate the role of these potential therapeutic targets in anti-TNF unresponsiveness. In addition to INHBA, six other genes of the network, including PTGS2, TLR8, IL7R, LCP1, LY96, and FCER1G, were involved in activated KRAS signaling in non-responders. Overexpression of these genes has been previously reported in both Crohn's disease (CD) and UC and even pediatric IBD patients [39–42].

Prostaglandin endoperoxide synthase 2 or PTGS2 also known as cyclooxygenase 2 (COX2) is an enzyme that increases prostaglandin production that generates inflammatory responses and can cause cardinal signs of acute inflammation [43,44]. Higher expression of *PTGS2* in the mucosa of UC patients leads to a more severe inflammation that might be associated with unresponsiveness in these patients. Toll-like receptor 8 or TLR8 not only has been reported to be upregulated in active UC but also has been discovered as an X-chromosome linked IBD susceptibility gene with both common predisposing and protecting haplotypes in both CD and UC [45,46]. Besides, TLR8 has been identified to induce mucosal inflammation in UC patients by enhancing the expression of TNF-α and IL-1β [47]. Another overexpressed gene, IL7R acts as a receptor for IL7 cytokine and is involved in maintaining T cell survival through the inhibition of mitochondrial death, upregulation of the anti-apoptotic protein BCL-2, and an increase in glucose uptake and metabolism. It has been proven as a fuel for autoimmunity and inflammatory response [48]. In this study, observing the overexpression of IL7R and IL6 as T cells saviors together with suppression of Treg subpopulations and involvement of inflammatory response together in UCN mucosa may indicate lymphocytosis condition in which Th17 cells are accumulated [49,50].

Lymphocyte cytosolic protein 1 or LCP1 has previously been reported as a diagnostic biomarker for IBD [51]. Also, an association between the expression alteration of this gene and unresponsiveness to anti-TNF therapy in IBD patients was observed by Wang et al. [52]. The function of this protein and how it contributes to inflammation have not yet been fully determined; however, a recent study by Mahat et al. suggested an inflammatory action by observing neutropenia that is caused by I232F missense mutation in LCP1 [53]. Further *in vitro* and *in vivo* studies are needed to explore the role of LCP1 in inflammation, especially in autoimmune diseases such as IBD. Lymphocyte antigen 96 or LY96, also known as myeloid differentiation 2 (MD2), plays an important role in inflammatory diseases such as CD [54]. LY96, a lymphocytic marker has been detected to activate the NF-kB pathway by interacting with TLR4 and inducing the production of pro-inflammatory cytokines in colon cancer [55]. Our enrichment analysis also showed the contribution of LY96 to TNF-α signaling via the NF-kB pathway in non-responders, suggesting the same role for this gene as discovered in colon cancer.

FCER1G gene, which encodes a high affinity Fc fragment of IgE receptor I for gamma polypeptide, is widely expressed in immune cells and connects adaptive and innate immunity [56]. In the inflammatory condition in the intestinal mucosa, TLRs and FCER1G co-stimulate and induce the Th17 polarizing macrophage phenotype, causing the production of inflammatory cytokines [57]. In such inflammatory conditions, FCER1G cross-linking with anti-commensal IgG leads to Th17 cell activation and immunity which is observed in the development of intestinal inflammation in UC. FCER1G was identified as the key regulator of KRAS signaling in the network with the highest number of degrees. The upregulation of *INHBA*, *PTGS2*, *TLR8*, *IL7R*, *LCP1*, *LY96*, and *FCER1G* together in the KRAS signaling pathway significantly aggravates inflammation in the colonic mucosa of UC patients and may cause resistance to anti-TNF mAbs.

The KRAS protein belongs to the RAS small GTPase superfamily that controls various cellular processes such as cell polarization, adhesion, and proliferation [58,59]. Among RAS superfamily proteins, KRAS is predominant in colon tissue and accounts for almost 90 % of RAS proteins in colonocytes [60,61]. Although the central role of KRAS signaling in maintaining the colon homeostasis and regeneration of intestinal stem cells has been recently reported [62], to the best of our knowledge, this is the first study that suggests a key role of KRAS signaling in anti-TNF treatment failure in UC patients. In this regard, the expression of KRAS was examined in UC patients; however, no significant differences were detected between UCN and UCR, suggesting an activating mutation in the KRAS gene that should be monitored in further research in a larger population. The inflammatory effects of KRAS mutation have been previously observed in lung tumors [63]. Furthermore, KRAS mutations are present in 45 % of patients with colorectal cancers [64]. Interestingly, it has been reported that KRAS mutational status is associated with poor response to anti-EGFR therapy in colorectal cancer [65].

Computational methods were applied to develop a platform based on the expression of several mRNA biomarkers as input by which the response to anti-TNF therapy in UC patients could be predicted. In biomedical research, several algorithms and machine learning methods have been applied to subset and classify biomolecules in order to identify predictive/diagnostic biomarkers of various pathological conditions [66,67]. In this study, the LASSO algorithm, which increases the interpretability and accuracy of linear models [68], was applied and repeated 50 times in a 10-fold cross-validation manner. This repetition with random folding of samples allowed LASSO to illustrate five specific indicators of UCN class, including *IL13RA2*, *HCAR3*, *CSF3*, *INHBA*, and *MMP1* with a more accurate absolute coefficient which later was used to develop predictive gene sets.

Among these predictors, *IL13RA2,* which had the greatest ALAC indicating higher potentiality in classifying UCN and UCR patients, has been previously reported as a predictive biomarker for the response to anti-TNF mAb in both UC and CD [11–13,69]. Interleukin 13 Receptor Alpha 2 or IL13RA2 acts as a decoy receptor for IL13 which is a type 2 immunity cytokine. IL31RA2 binds IL13 with higher affinity than IL4RA/IL13RA1, thus declining the available IL-13 to drive STAT6-dependent signaling [70]. This signaling has been

indicated to be involved in mucosal repair in IBD using murine models [71]. As a result, upregulation of this gene in the active lesions of UCN may impair the type 2 immune response mediated by IL13, thus impeding intestinal tissue remodeling.

MMP1 was also introduced as one of the predictors of anti-TNF treatment failure. Elevated expression of Matrix metalloproteinases or MMPs has been reported in various inflammatory conditions in human diseases, and this intemperate inflammation not only does not help host defense but also causes further tissue injury [72]. Besides, MMPs are able to cleave IgG antibodies [73]; hence, their induced expression could prevent the neutralizing function of anti-TNF IgGs on TNF molecules [74], which may cause anti-TNF mAb therapy failure. Of note, the potential role of *MMP3*, another member of the matrixin family, in predicting the response to infliximab in IBD patients has been previously reported [75].

Hydroxy carboxylic acid receptor-3 or HCAR3, another predictor of anti-TNF treatment failure, contributes to the regulation of lipolysis during increased β-oxidation and is probably involved in the interaction of gut microbiome-derived metabolites and immune cells [76,77]. Until now, evidence is still lacking to indicate a pro-inflammatory function for HCAR3; however, a recent study has reported the anti-inflammatory effect of HCAR3 expression in macrophages [78]. While few studies have reported an association between the overexpression of HCAR3 and UC pathogenesis [79–81], it is the first time this gene is related to unresponsiveness to anti-TNF therapy failure.

Colony-Stimulating Factor 3 or CSF3, also called Granulocyte-CSF (G-CSF), is a member of the IL-6 superfamily whose receptor is mainly expressed on neutrophils [25]. G-CSF, which increases inflammatory conditions [82], inhibits apoptosis in neutrophils *in vivo* and *in vitro* [83]. Therefore, *CSF3* overexpression in the mucosa of UCN patients prevents apoptosis of neutrophils and explains the possible reason for neutrophil accumulation in the inflamed lesions, leading to inflammation progression and treatment failure. This finding is intriguingly consistent with the higher number of recruited neutrophils in the inflamed colonic mucosa of UCN patients that we observed in the immune cell subpopulation analysis. *INHBA*, which was discussed earlier in KRAS signaling activation, was the final significant predictive marker with the lower ALAC. Apart from IL13RA2, four other predictive biomarkers were introduced for the first time, and their role in the unresponsiveness of UC patients to anti-TNF mAbs needs to be experimentally clarified.

While various studies have been conducted to identify predictors of the response to anti-TNF mAbs [13,40,69], few studies have introduced a panel that can perform more accurately compared to single biomarkers in anti-TNF therapy response prediction. Using the ALAC of these predictive biomarkers given by repeated LASSO algorithm, a combination of three-, four-, and five-mRNA sets were defined to develop an input panel for machine learning. Then, a two-layer stack ensemble machine learning was designated, at the first layer of which neural network as a function-based classifier, k-nearest neighbor and random forest as rule-based classifiers, and linear discriminant analysis as a dimension reducer were employed, and at the second layer, a generalized linear model operated in a 10-fold cross-validation manner with 10 times random sampling to produce the most reliable results. Previous computational studies on ensemble algorithms have proven that in classification and prediction, not only such an engineered ensemble machine is able to deal with small sample sizes and high dimensionality in microarray cohorts better than single machine-learning models [84], but also generally has a better performance in terms of accuracy and stability than a single machine learning algorithm [85]. Consistently, our ensemble machine had a more accurate performance compared to individual machines in both *in silico* and real-life cohorts. Moreover, since ensemble models are computationally complex, they can overcome some of the problems associated with small sample sizes and large dimensionality, which are frequently encountered in bioinformatics research [84]. Although few studies have been accomplished to develop predictive panels for the response to anti-TNF mAbs in IBD patients, such an engineered ensemble machine, which is able to generate reliable accurate models, has not been previously designated. An *in silico* recent study by Sakaram et al. [12] has reported a prognostic panel of seven genes that predict the response to infliximab and adalimumab with an accuracy of 86 %, based on an Anti-TNF Response (ATR) scoring system in conjunction with the ground truth response adjudication. Another study by Arijs et al. [11] reported a five-mRNA gene set capable of predicting infliximab therapy response with 89 % accuracy using supervised linear modeling. In the current research, for the first time, a five-mRNA panel and an ensemble machine learning algorithm which were designated and evaluated using four different cohorts were introduced with the capacity of predicting the response to two types of anti-TNF mAbs, infliximab and adalimumab, with an average accuracy of 95.3 % (sensitivity 97.6 %, specificity 92 %). However, there are some limitations to this study such as small sample-sized cohorts. The performance of this panel in predicting the response to anti-TNF mAbs should be further investigated in larger cohorts to be applied for clinical purposes.

5. Conclusion

In this study, we presented a statistical methodology, a combination of high-throughput gene expression analysis and determination of immune cell subpopulation, to decipher pathogenic signals in UCN patients that were mostly concealed from view. We suggested important pathological functions for *IL6* and *INHBA* in anti-TNF unresponsiveness in UC patients through decreasing local Tregs, promoting distinct inflammatory conditions in the colon mucosa. We additionally proposed the contribution of KRAS signaling to anti-TNF treatment failure in UC patients for the first time, which can be monitored in further research. Furthermore, through a systematic and comprehensive study design including discovery and validation phases, we introduced an accurate ensemble machine with a five-mRNA biomarker panel capable of predicting the response to anti-TNF mAbs in colitis patients that could be developed for clinical use.

Ethics approval and consent to participants

This research was approved by the ethical review board of Shahid Beheshti University of Medical Sciences (approval number IR.SBMU.RIGLD.REC.1402.001). All the participants were diagnosed in the gastroenterology and liver diseases clinic and consented to be

involved in the current research.

Author contribution statement

Mohammad Hossein Derakhshan Nazari: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.
Shabnam Shahrokh: Performed the experiments; Contributed reagents, materials, analysis tools or data.
Leila Ghanbari-Maman: Performed the experiments; Analyzed and interpreted the data.
Samaneh Maleknia: Analyzed the data.
Mahsa Ghorbaninejad: Performed the experiments.
Anna Meyfour: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

Data included in article/supp. material/referenced in article.

Funding

This work was financially supported by the Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Anna Meyfour, Mohammad Hossein Derakhshan Nazari, and Shabnam Shahrokh have registered a patent related to the predictive panels introduced in this article in the Intellectual Property Center of the Islamic Republic of Iran on 10/28/2022 under the number ID 140150140003005566. All other authors declare that they have no competing interests.

Acknowledgment

We thank Dr. Fayaz Zadeh for her valuable comments on feature selection and network analysis. In addition, we thank the Shahid Beheshti University of Medical Science and the CinnaGen company to support this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e21154.

List of Abbreviations

Ulcerative Colitis UC
Inflammatory Bowel Disease IBD
5-Aminosalicylates 5-ASA
Tumor Necrosis Factor TNF
monoclonal antibodies mAb
UC Responders UCR
UC Non-responders UCN
Cross-Platform Normalization CPN
Gene Expression Omnibus GEO
False Discovery Rate FDR
the Differentially Expressed Genes DEG
Gene Ontology GO
Gene Ontology Resource GOR
Molecular Signature Databases MSigDB
Biological Processes BP
Molecular Functions MF
Cellular Components CC
Search Tool for Retrieval of Interacting Genes/Proteins STRING
Least Absolute Shrinkage and Selection Operator LASSO
ALAC Average LASSO Absolute Coefficient
Neural Network NN
K-nearest Neighbor KNN

Random Forest RF
Generalized Linear Model GLM
Linear Discriminant Analysis LDA
Receiver Operating Characteristic curve ROC
Area Under Curve AUC
Quantitative Real-time Reverse Transcription PCR qRT-PCR
Standard Error Mean SEM
Principle Component Analysis PCA
Crohn's disease CD
T helper 17 Th17
Granulocyte-Colony Stimulating Factor G-CSF
Myeloid Differentiation 2 MD2
Cyclooxygenase 2 COX2
Microarray Expression Level MEL
Differentially Expression DE

References

[1] T. Kobayashi, et al., Ulcerative colitis, Nat. Rev. Dis. Prim. 6 (1) (2020) 74.
[2] G. Cui, et al., Evaluation of anti-TNF therapeutic response in patients with inflammatory bowel disease: current and novel biomarkers, EBioMedicine 66 (2021), 103329.
[3] D.T. Rubin, et al., ACG clinical guideline: ulcerative colitis in adults, Am. J. Gastroenterol. 114 (3) (2019) 384–413.
[4] L. Du, C. Ha, Epidemiology and pathogenesis of ulcerative colitis, Gastroenterol. Clin. N. Am. 49 (4) (2020) 643–654.
[5] R. Ungaro, et al., A Treat-to-target update in ulcerative colitis: a systematic review, Am. J. Gastroenterol. 114 (6) (2019) 874–883.
[6] J.F. Colombel, et al., Outcomes and strategies to support a Treat-to-target approach in inflammatory bowel disease: a systematic review, J Crohns Colitis 14 (2) (2020) 254–266.
[7] G. Fiorino, et al., Medical therapy versus surgery in moderate-to-severe ulcerative colitis, Dig. Liver Dis. 53 (4) (2021) 403–408.
[8] M. Vulliemoz, et al., TNF-alpha blockers in inflammatory bowel diseases: practical recommendations and a user's guide: an update, Digestion 101 (1) (2020) 16–26, suppl 1.
[9] D. Lissner, et al., Monocyte and M1 macrophage-induced barrier defect contributes to chronic intestinal inflammation in IBD, Inflamm. Bowel Dis. 21 (6) (2015) 1297–1305.
[10] K. Papamichael, et al., Role for therapeutic drug monitoring during induction therapy with TNF antagonists in IBD: evolution in the definition and management of primary nonresponse, Inflamm. Bowel Dis. 21 (1) (2015) 182–197.
[11] I. Arijs, et al., Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis, Gut 58 (12) (2009) 1612–1619.
[12] S. Sakaram, et al., A multi-mRNA prognostic signature for anti-TNFα therapy response in patients with inflammatory bowel disease, Diagnostics 11 (10) (2021).
[13] B. Verstockt, et al., Low TREM1 expression in whole blood predicts anti-TNF response in inflammatory bowel disease, EBioMedicine 40 (2019) 733–742.
[14] S.E. Telesco, et al., Gene expression signature for prediction of golimumab response in a phase 2a open-Label Trial of patients with ulcerative colitis, Gastroenterology 155 (4) (2018) 1008–1011 e8.
[15] N. Jiang, et al., Methods for evaluating gene expression from Affymetrix microarray datasets, BMC Bioinf. 9 (1) (2008) 284.
[16] R. Torres, R.L. Judson-Torres, Research Techniques made simple: feature selection for biomarker discovery, J. Invest. Dermatol. 139 (10) (2019) 2068–2074.e1.
[17] I. Arijs, et al., Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment, PLoS One 4 (11) (2009) e7984.
[18] I. Arijs, et al., Effect of vedolizumab (anti-α4β7-integrin) therapy on histological healing and mucosal gene expression in patients with UC, Gut 67 (1) (2018) 43–52.
[19] Y. Zhao, L. Wong, W.W.B. Goh, How to do quantile normalization correctly for gene expression data analyses, Sci. Rep. 10 (1) (2020), 15534.
[20] C. Chen, et al., Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods, PLoS One 6 (2) (2011), e17238.
[21] G. Scardoni, C. Laudanna, Centralities based analysis of complex networks, New frontiers in graph theory (2012) 323–348.
[22] S. Maleknia, et al., An integrative Bayesian network approach to highlight key drivers in systemic lupus erythematosus, Arthritis Res. Ther. 22 (1) (2020) 156.
[23] J.S. Hamid, et al., Data integration in genetics and genomics: methods and challenges, Hum. Genom. Proteonomics 2009 (2009).
[24] H.M. Penrose, et al., Ulcerative colitis immune cell landscapes and differentially expressed gene signatures determine novel regulators and predict clinical response to biologic therapy, Sci. Rep. 11 (1) (2021) 9010.
[25] D. Muthas, et al., Neutrophils in ulcerative colitis: a review of selected biomarkers and their potential therapeutic implications, Scand. J. Gastroenterol. 52 (2) (2017) 125–135.
[26] C. Zhang, et al., Anti-TNF-α Therapy Suppresses Proinflammatory Activities of Mucosal Neutrophils in Inflammatory Bowel Disease, 2018, Mediators of inflammation, 2018, 3021863, 3021863.
[27] J. van der Veeken, et al., Memory of inflammation in regulatory T cells, Cell 166 (4) (2016) 977–990.
[28] A. Oussalah, et al., A multicenter experience with infliximab for ulcerative colitis: outcomes and predictors of response, optimization, colectomy, and hospitalization, Am. J. Gastroenterol. 105 (12) (2010) 2617–2625.
[29] M. Ferrante, et al., Predictors of early response to infliximab in patients with ulcerative colitis, Inflamm. Bowel Dis. 13 (2) (2007) 123–128.
[30] T. Tanaka, M. Narazaki, T. Kishimoto, IL-6 in inflammation, immunity, and disease, Cold Spring Harbor Perspect. Biol. 6 (10) (2014), a016295 a016295.
[31] S.A. O'Brien, M. Zhu, W. Zhang, The importance of IL-6 in the development of LAT-mediated autoimmunity, J. Immunol. 195 (2) (2015) 695–705.
[32] K. Mitsuyama, M. Sata, S. Rose-John, Interleukin-6 trans-signaling in inflammatory bowel disease, Cytokine Growth Factor Rev. 17 (6) (2006) 451–461.
[33] D. Bernardo, et al., IL-6 promotes immune responses in human ulcerative colitis and induces a skin-homing phenotype in the dendritic cells and Tcells they stimulate, Eur. J. Immunol. 42 (5) (2012) 1337–1353.
[34] Z. He, J. Liang, B. Wang, Inhibin, beta A regulates the transforming growth factor-beta pathway to promote malignant biological behaviour in colorectal cancer, Cell Biochem. Funct. 39 (2) (2021) 258–266.
[35] M.O. Li, S. Sanjabi, R.A. Flavell, Transforming growth factor-beta controls development, homeostasis, and tolerance of T cells by regulatory T cell-dependent and -independent mechanisms, Immunity 25 (3) (2006) 455–471.
[36] J.C. Marie, D. Liggitt, A.Y. Rudensky, Cellular mechanisms of fatal early-onset autoimmunity in mice with the T cell-specific targeting of transforming growth factor-beta receptor, Immunity 25 (3) (2006) 441–454.

[37] S. Sanjabi, et al., Anti-inflammatory and pro-inflammatory roles of TGF-beta, IL-10, and IL-22 in immunity and autoimmunity, Curr. Opin. Pharmacol. 9 (4) (2009) 447–453.
[38] T. Korn, et al., IL-17 and Th17 cells, Annu. Rev. Immunol. 27 (2009) 485–517.
[39] S. Salvador-Martín, et al., Gene signatures of early response to anti-TNF drugs in pediatric inflammatory bowel disease, Int. J. Mol. Sci. 21 (9) (2020).
[40] Y. Liu, Y. Duan, Y. Li, Integrated gene expression profiling analysis reveals probable molecular mechanism and candidate biomarker in anti-TNFα non-response IBD patients, J. Inflamm. Res. 13 (2020) 81–95.
[41] I. Arijs, et al., Predictive value of epithelial gene expression profiles for response to infliximab in Crohn's disease, Inflamm. Bowel Dis. 16 (12) (2010) 2090–2098.
[42] G. Privitera, et al., Predictors and early markers of response to biological therapies in inflammatory bowel diseases, J. Clin. Med. 10 (4) (2021).
[43] E. Ricciotti, G.A. FitzGerald, Prostaglandins and inflammation, Arterioscler. Thromb. Vasc. Biol. 31 (5) (2011) 986–1000.
[44] D.G. Cox, et al., Polymorphisms in prostaglandin synthase 2/cyclooxygenase 2 (PTGS2/COX2) and risk of colorectal cancer, Br. J. Cancer 91 (2) (2004) 339–343.
[45] F. Sanchez-Munoz, et al., Transcript levels of Toll-like receptors 5, 8 and 9 correlate with inflammatory activity in ulcerative colitis, BMC Gastroenterol. 11 (2011) 138.
[46] M. Saruta, et al., High-frequency haplotypes in the X chromosome locus TLR8 are associated with both CD and UC in females, Inflamm. Bowel Dis. 15 (3) (2009) 321–327.
[47] Y. Lu, et al., Toll-like receptors and inflammatory bowel disease, Front. Immunol. 9 (2018) 72.
[48] H. Dooms, Interleukin-7: fuel for the autoimmune attack, J. Autoimmun. 45 (2013) 40–48.
[49] R. Baccala, A.N. Theofilopoulos, The new paradigm of T-cell homeostatic proliferation-induced autoimmunity, Trends Immunol. 26 (1) (2005) 5–8.
[50] Z. Tatari-Calderone, et al., Age-related accumulation of T cells with markers of relatively stronger autoreactivity leads to functional erosion of T cells, BMC Immunol. 13 (1) (2012) 8.
[51] N.Y. Han, et al., Label-free quantification for discovering novel biomarkers in the diagnosis and assessment of disease activity in inflammatory bowel disease, J Dig Dis 14 (4) (2013) 166–174.
[52] T.H. Kim, et al., Gene expression profile predicting the response to anti-TNF treatment in patients with rheumatoid arthritis; analysis of GEO datasets, Joint Bone Spine 81 (4) (2014) 325–330.
[53] U. Mahat, et al., Lymphocyte cytosolic protein 1 (L-plastin) I232F mutation impairs granulocytic proliferation and causes neutropenia, Blood Advances 6 (8) (2022) 2581–2594.
[54] Y. Yang, et al., Aureusidin derivative CNQX inhibits chronic colitis inflammation and mucosal barrier damage by targeting myeloid differentiation 2 protein, J. Cell Mol. Med. 25 (15) (2021) 7257–7269.
[55] V. Rajamanickam, et al., Selective targeting of the TLR4 co-receptor, MD2, prevents colon cancer growth and lung metastasis, Int. J. Biol. Sci. 16 (8) (2020) 1288–1302.
[56] F. Nimmerjahn, J.V. Ravetch, Fc-receptors as regulators of immunity, Adv. Immunol. 96 (2007) 179–204.
[57] T. Castro-Dopico, M.R. Clatworthy, IgG and Fcγ receptors in intestinal immunity and inflammation, Front. Immunol. 10 (2019) 805.
[58] D.K. Simanshu, D.V. Nissley, F. McCormick, RAS proteins and their regulators in human disease, Cell 170 (1) (2017) 17–33.
[59] A.E. Karnoub, R.A. Weinberg, Ras oncogenes: split personalities, Nat. Rev. Mol. Cell Biol. 9 (7) (2008) 517–531.
[60] S. Vasaikar, et al., Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities, Cell 177 (4) (2019) 1035–1049.e19.
[61] D. Wang, et al., A deep proteome and transcriptome abundance atlas of 29 healthy human tissues, Mol. Syst. Biol. 15 (2) (2019) e8503.
[62] C. Ternet, C. Kiel, Signaling pathways in intestinal homeostasis and colorectal cancer: KRAS at centre stage, Cell Commun. Signal. 19 (1) (2021) 31.
[63] H. Ji, et al., K-ras activation generates an inflammatory response in lung tumors, Oncogene 25 (14) (2006) 2105–2112.
[64] C.A. Ellis, G. Clark, The importance of being K-Ras, Cell. Signal. 12 (7) (2000) 425–434.
[65] J.H. van Krieken, et al., KRAS mutation testing for predicting response to anti-EGFR therapy for colorectal carcinoma: proposal for an European quality assurance program, Virchows Arch. 453 (5) (2008) 417–431.
[66] R. Torres, R.L. Judson-Torres, Research Techniques made simple: feature Selection for Biomarker discovery, J. Invest. Dermatol. 139 (10) (2019) 2068–2074, e1.
[67] J. Noorbakhsh, et al., Machine learning in biology and medicine, Advances in Molecular Pathology 2 (1) (2019) 143–152.
[68] J.-P. Vert, et al., An accurate and interpretable model for siRNA efficacy prediction, BMC Bioinf. 7 (1) (2006) 520.
[69] B. Verstockt, et al., Mucosal IL13RA2 expression predicts nonresponse to anti-TNF therapy in Crohn's disease, Aliment Pharmacol. Therapeut. 49 (5) (2019) 572–581.
[70] E.P. Karmele, et al., Anti-IL-13Rα2 therapy promotes recovery in a murine model of inflammatory bowel disease, Mucosal Immunol. 12 (5) (2019) 1174–1186.
[71] J. Cosín-Roger, et al., The activation of Wnt signaling by a STAT6-dependent macrophage phenotype promotes mucosal repair in murine IBD, Mucosal Immunol. 9 (4) (2016) 986–998.
[72] A.M. Manicone, J.K. McGuire, Matrix metalloproteinases as modulators of inflammation, Semin. Cell Dev. Biol. 19 (1) (2008) 34–41.
[73] R.J. Brezski, R.E. Jordan, Cleavage of IgGs by proteases associated with invasive diseases: an evasion tactic against host immunity? mAbs 2 (3) (2010) 212–220.
[74] P. Biancheri, et al., Proteolytic cleavage and Loss of function of biologic agents that neutralize tumor Necrosis factor in the mucosa of patients with inflammatory bowel disease, Gastroenterology 149 (6) (2015) 1564–1574.e3.
[75] B. Barberio, et al., Matrix metalloproteinase 3 predicts therapeutic response in inflammatory bowel disease patients treated with infliximab, Inflamm. Bowel Dis. 26 (5) (2020) 756–763.
[76] K. Ahmed, et al., Deorphanization of GPR109B as a receptor for the beta-oxidation intermediate 3-OH-octanoic acid and its role in the regulation of lipolysis, J. Biol. Chem. 284 (33) (2009) 21928–21933.
[77] A. Peters, et al., Metabolites of lactic acid bacteria present in fermented foods are highly potent agonists of human hydroxycarboxylic acid receptor 3, PLoS Genet. 15 (5) (2019), e1008145.
[78] A. Peters, et al., Hydroxycarboxylic acid receptor 3 and GPR84 - two metabolite-sensing G protein-coupled receptors with opposing functions in innate immune cells, Pharmacol. Res. 176 (2022), 106047.
[79] G. Xu, et al., Bioinformatics analysis of key candidate genes and pathways in ulcerative colitis, Biol. Pharm. Bull. 43 (11) (2020) 1760–1766.
[80] Z.A. Chen, et al., Integrated analysis of multiple microarray studies to identify novel gene signatures in ulcerative colitis, Front. Genet. 12 (2021), 697514.
[81] L. Shi, et al., Identification of differentially expressed genes in ulcerative colitis and verification in a colitis mouse model by bioinformatics analyses, World J. Gastroenterol. 26 (39) (2020) 5983–5996.
[82] J.A. Hamilton, Colony-stimulating factors in inflammation and autoimmunity, Nat. Rev. Immunol. 8 (7) (2008) 533–544.
[83] B.J. van Raam, et al., Granulocyte colony-stimulating factor delays neutrophil apoptosis by inhibition of calpains upstream of caspase-3, Blood 112 (5) (2008) 2046–2054.
[84] P. Yang, et al., A review of ensemble methods in bioinformatics, Curr. Bioinf. 5 (4) (2010) 296–308.
[85] A. Osareh, B. Shadgar, An efficient ensemble learning method for gene microarray classification, BioMed Res. Int. 2013 (2013), 478410.