Journal of Translational Medicine

RESEARCH

Open Access

Integrating single-cell RNA-Seq and machine learning to dissect tryptophan metabolism in ulcerative colitis

Guorong Chen[1,2†], Hongying Qi[3†], Li Jiang[4†], Shijie Sun[1], Junhai Zhang[1], Jiali Yu[1], Fang Liu[1], Yanli Zhang[1*] and Shiyu Du[1*]

Abstract

Background Ulcerative colitis (UC) is a persistent inflammatory bowels disease (IBD) characterized by immune response dysregulation and metabolic disruptions. Tryptophan metabolism has been believed as a significant factor in UC pathogenesis, with specific metabolites influencing immune modulation and gut microbiota interactions. However, the precise regulatory mechanisms and key genes involved remain unclear.

Methods AUCell, Ucell, and other functional enrichment algorithms were utilized to determine the activation patterns of tryptophan metabolism at the UC cell level. Differential analysis identified key genes associated with tryptophan metabolism. Five machine learning algorithms, including Random Forest, Boruta algorithm, LASSO, SVM-RFE, and GBM were integrated to identify and categorize disease-specific characteristic genes.

Results We observed significant heterogeneity in tryptophan metabolism activity across cell types in UC, with the highest activity levels in macrophages and fibroblasts. Among the key tryptophan metabolism-related genes, CTSS, S100A11, and TUBB were predominantly expressed in macrophages and significantly upregulated in UC, highlighting their involvement in immune dysregulation and inflammation. Cross-analysis with bulk RNA data confirmed the consistent upregulation of these genes in UC samples, highly indicating their relevance in UC pathology and potential as targets for therapeutic intervention.

Conclusions This study is the first to reveal the heterogeneity of tryptophan metabolism at the single-cell level in UC, with macrophages emerging as key contributors to inflammatory processes. The identification of CTSS, S100A11, and TUBB as key regulators of tryptophan metabolism in UC underscores their potential as biomarkers and therapeutic targets.

Keywords Tryptophan metabolism, Ulcerative colitis, Macrophage

†Guorong Chen, Hongying Qi and Li Jiang contributed equally to this work.

*Correspondence:
Yanli Zhang
13521234067@163.com
Shiyu Du
dushiyu1975@126.com

[1]Department of Gastroenterology, China-Japan Friendship Hospital (Institute of Clinical Medical Sciences), Beijing 100029, China
[2]Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 100730, China
[3]Department of Spleen and Stomach Diseases of Traditional Chinese Medicine, China-Japan Friendship Hospital (Institute of Clinical Medical Sciences), Beijing 100029, China
[4]Department of Endocrinology, Aviation General Hospital, Beijing 100025, China

Introduction

Ulcerative colitis (UC) is a chronic relapsing form of inflammatory bowel disease (IBD), commonly triggered by environmental exposures in genetically susceptible individuals. Dysfunctions in the epithelial barrier, immune response imbalances, and ecological disturbances participate in the inflammatory processes [1–3]. Recent researches highlight various metabolic disorders in UC, where small molecule metabolites engage in signaling, immune modulation, and interactions with the gut microbiota, significantly contributing to the onset and progression of UC [4]. Among these, tryptophan metabolism plays a crucial role, with its metabolites shown to affect immune balance and microbiota composition. However, the specific cellular mechanisms and genetic factors that regulate tryptophan metabolism in UC remain largely unexplored [5, 6].

Tryptophan, an essential amino acid, plays a key role in the synthesis of various biologically active compounds and must be obtained from the diet [7]. Its metabolism involves three primary pathways: the kynurenine (Kyn), 5-hydroxytryptamine (5-HT), and indole pathways, with the majority of tryptophan being metabolized through the Kyn pathway. These pathways allow Trp to be converted into various metabolically active intermediates, which possess immunomodulatory properties essential for maintaining immune homeostasis [8]. Dysregulated tryptophan and its metabolites contribute to immune imbalance and intestinal inflammation in UC, leading to epithelial barrier disruption and activation of inflammatory pathways. This metabolic imbalance not only aggravates chronic inflammation but also promotes UC progression by facilitating immune cell infiltration and cytokine release [5, 9, 10].

While current research highlights the significant role of tryptophan metabolites in UC onset and progression, identifying and validating the key regulatory genes involved remains challenging. Traditional experimental methods often lack the sensitivity to detect cell-type-specific gene expression and metabolic activity, which are crucial for understanding the regulatory heterogeneity of tryptophan metabolism in UC. Single-cell RNA sequencing (scRNA-seq) is an emerging tool that enables genomic examination, cell heterogeneity analysis, differential gene expression, and identification of specific cell types within individual cells of a tissue [11, 12]. This approach offers notable advantages over traditional bulk RNA sequencing by capturing cellular diversity within complex tissues. When combined with bioinformatics methods, scRNA-seq is widely used in disease diagnosis, prognosis, and investigating the role of metabolites in various diseases [13–15]. Machine learning, a branch of artificial intelligence, enables the analysis of vast data sets, facilitating tasks such as disease diagnosis and the identification of key regulatory genes and pathways. Metaheuristic optimization algorithms further enhance these capabilities by optimizing complex data analysis [16, 17]. Integrating machine learning with bioinformatics methods significantly boosts the advantages of scRNA-seq, improving data analysis accuracy and uncovering hidden relationships within the data [18–20].

In this study, we were the first to find the heterogeneity of tryptophan metabolism at the single cell level of UC, with significant differences among cell types. Through functional enrichment algorithms and machine learning, we recognized key genes linked to the upregulation of tryptophan metabolism, and contributed to the pathology and development of UC, suggesting that they might be new targets and providing critical insights for future research and potential therapeutic interventions in UC.

Method

Data acquisition and processing

In this study, scRNA-seq data for UC were obtained from the GSE214695 and GSE125527 database (https://www.ncbi.nlm.nih.gov/geo/), comprising 13 UC and 14 health control (HC) samples (Supplementary Table S1) [21, 22]. A total of 48 tryptophan metabolism-related genes (TrMGs) information were selected based on KEGG, GO, REACTOME databases and previous study (Supplementary Table S2) [23]. For the processing of scRNA-seq data, we preserved high-quality cells that had fewer than 20% mitochondrial genes and expressed more than 200 genes. We also focused on genes that were expressed at levels between 200 and 7000 and were active in at least three cells. A total of 39,080 eligible cells were kept for further exploration. After that, the integration workflow conducted by Seurat pipeline [24]. The remaining cells were further scaled and normalized using a linear regression model with the "Log-normalization" method and the top 3000 variable genes were detected by the "FindVariablFeatures" function. Subsequently, the dimensionality of the scRNA-seq data was diminished through Principal Component Analysis (PCA). Uniform manifold approximation and projection (UMAP) dimensional reduction, dataset integration, and cell types were annotated with the aid of the R package "single R" [25]. To remove the batch effects among the samples, soft k-means clustering was executed using the "Harmony" package [26]. The cell clustering was conducted using the "FindClusters" function, with the resolution parameter set at 0.8. The methodology for annotating cell clusters involved focusing on genes with elevated expression levels, genes exhibiting unique expression patterns, and documented canonical cellular markers. For Bulk RNAseq data, the GSE887466 cohort was employed as training set, and GSE53306 was validation set (http://www.ncbi.nlm.nih.gov/geo).

Gene set scoring algorithm in scRNA-seq

We used five different algorithms to score gene sets in scRNA-seq datasets: AUCell [27], UCell [28], singscore [29], ssGSEA [30], and AddModuleScore [31]. AUCell and UCell were specifically selected for their unique strengths in quantifying gene set activity at the single-cell level, which is essential for accurately identifying activation patterns in UC cells. AUCell calculated gene set activity in each cell by computing the area under the cumulative distribution curve (AUC) of gene expression ranks. UCell assessed gene set activity by calculating and normalizing rank scores within single-cell gene expression rankings. Singscore ranked genes within each cell for a given gene set and calculated the average rank score, providing a score based on the difference between average ranks of positive and negative genes. ssGSEA determined gene set enrichment by comparing expression values of gene sets to other genes, computing relative enrichment scores. AddModuleScore calculated weighted average expression of gene sets in each cell, normalizing these values to obtain the final score.

Differential gene expression and functional enrichment analysis

Differential expression analysis was performed to identify differentially expressed genes (DEGs) between the high and low TrMG groups using the 'FindMarkers' function, with criteria set at |log2 fold change| > 0.25 and adjusted p value < 0.05 for further investigation. Additionally, correlation analysis was conducted to pinpoint genes most strongly linked with TrMGs expression, with the top 100 most correlated genes being included for future study. These DEGs and genes discovered through correlation analysis were the ones that had the greatest influence on up-regulated TrMGs expression. Subsequently, Gene Ontology (GO) enrichment analysis was carried out using 'clusterProfiler' package in R software, aiming to elucidate the potential mechanistic underpinnings governed by these DEGs.

Screening of optimal TrMGs

A total of five independent machine learning algorithms were utilized to screen out the optimal key genes in UC, including the Boruta algorithm [32], LASSO [33], SVM-RFE [34], the GBM [35], and random forest [36]. These algorithms were selected for their complementary strengths in feature selection, model optimization and avoiding single algorithm offset, which collectively enhance the robustness and accuracy of identifying UC-specific characteristic genes. The Boruta algorithm was chosen for its ability to rigorously identify all relevant features associated with the target variable, ensuring that only the most important genes are retained. This algorithm iteratively compares the importance of original features against that of randomly permuted shadow attributes, enabling us to retain only the most relevant features linked to UC while eliminating noise, thus providing a high-confidence selection of characteristic genes. LASSO was applied using the "glmnet" package, which introduces a regularization term to shrink coefficients, enabling the selection of the most predictive features while eliminating irrelevant or redundant genes. SVM-RFE algorithm, a feature elimination method based on support vector machines, was used to iteratively remove the least important features, thereby refining the set of key genes that contribute most to classification accuracy. GBM and Random Forest were included to further improve prediction reliability. GBM enhances model performance by sequentially building trees that correct errors from previous ones, which optimizes the overall predictive accuracy. Meanwhile, Random Forest builds multiple decision trees and averages their outcomes to rank feature importance, allowing us to select the top 20 diagnostic gene candidates. Finally, the overlapping genes based on the five above mentioned machine learning algorithms were selected as the hub TrMGs in UC and visualized by a venn plot.

GSVA enrichment analysis

Gene Set Variation Analysis (GSVA) enrichment analysis was conducted to investigate differential biological mechanisms between distinct risk groups using the "GSVA" R package. Gene sets from the c2.cp.kegg.v7.4.symbols.gmt and h.all.v2022.1.Hs.symbols.gmt collections were sourced from the MSigDB database (https://www.gsea-msigdb.org/gsea/msigdb). False discovery rate (FDR) adjustments were implemented via the Benjamini and Hochberg (BH) method, with significance defined as FDR < 0.05.

Cell communication

CellChat (20) was used to analyze gene expression data and explore variations in potential cell-cell communication networks. Employing the conventional CellChat pipeline, we relied on the default CellChatDB for ligand-receptor interactions. By identifying overexpressed ligands or receptors within specific cell groups, we inferred cell type-specific interactions. Furthermore, we identified heightened ligand-receptor interactions associated with overexpressed ligands or receptors. Additionally, we leveraged the R package Scenic to infer the activity of gene regulatory networks.

Pseudo-time analysis

The Monocle package was used to conduct reverse chronological analysis, aimed at reconstructing the developmental trajectory of cells based on single-cell gene expression data. This intricate process entailed

constructing a single-cell expression matrix, categorizing cells into distinct developmental states, and delineating cell developmental trajectories by discerning gene expression patterns. We also evaluated cell maturity or developmental status utilizing the Cytotrace method. This meticulous analysis quantified the developmental status of each cell by scrutinizing changes in gene expression within scRNA-seq data.

Statistical analysis

All data processing, statistical analysis, and visualization were conducted using R 4.1.3 software. Subtype-specific overall survival (OS) was estimated and compared employing the Kaplan-Meier method and log-rank test. Differences in continuous variables between groups were evaluated utilizing either the Wilcoxon test or t-test. Categorical variables were analyzed using the chi-squared test or Fisher's exact test. The FDR method was applied to adjust p-values. Pearson correlation analysis was employed to examine associations between variables. Two-tailed tests were utilized for all p-values, with statistical significance defined as $p<0.05$.

Results

The scRNA-seq profiling of UC

Before further analysis, quality control was performed on all the included samples (Fig. 1a), and ultimately, 27 samples (14 from the HC group and 13 from the UC group) were selected for analysis. Batch effect correction was applied across all samples (Fig. 1b), showing that the overall distribution was relatively stable, and sensitivity to batch effects was minimal. Following the Seurat pipeline analysis, all cells were grouped into 20 subpopulations with detailed clustering shown (Fig. 1c). The expression patterns of characteristic marker genes related to 11 cell subsets were illustrated (Fig. 1d), and cell types were identified based on using specific marker genes such as NKG7 and CD3D (Fig. 1e). The UMAP plot showed the presence of these 11 cell types, including macrophages, B cells, T cells, epithelial cells, etc. (Fig. 1f). To better understand the differences between the UC and HC groups, we compared the proportion of different single cell types between the two groups (Fig. 1g). The proportion of cell types involved in immune response, such as B cells, T cells and macrophages, increased significantly in the UC group, indicating that immune factors were pivotal in the UC pathogenesis mechanism.

Tryptophan metabolism in scRNA-seq data

The tryptophan metabolic pathway played a significant role in the progression of IBD, and its metabolic level was dysregulated in IBD patients. We employed the AUCell, UCell, singscore, ssGSEA, and Add algorithms to evaluate tryptophan metabolism at the scRNA-seq level, assessing the expression of TrMGs across different cell types by averaging the scores from above algorithms (Fig. 2a). We found that the activity of TrMGs in different cell types showed great heterogeneity (Fig. 2b). Specifically, these genes were more active in macrophages and intestinal epithelial cells but relatively suppressed in neutrophils, T cells, and NK cells (Fig. 2c). A comparative analysis of the HC and UC groups indicated that TrMGs were up-regulated in macrophages and fibroblasts and down-regulated in NK cells, plasma cells, and B cells in the UC group (Fig. 2d). The UMAP plot further demonstrated that these genes were predominantly expressed in macrophages (Fig. 2e). In summary, TrMGs were significantly upregulated in the macrophages of UC group compared to the HC group. By using the average expression scores, 39,080 eligible cells were classified into high-expression (scores above the mean) and low-expression (scores below the mean) groups. The high-expression group was primarily comprised of macrophages, fibroblasts, and epithelial cells (Fig. 2f). A Wilcoxon rank-sum test identified 78 differentially upregulated genes between the high and low expression groups (Fig. 2g). To identify the genes most closely associated with tryptophan metabolic activity, we conducted a correlation analysis and identified 385 genes significantly related to tryptophan expression ($r>0.1$, $p<0.01$) (Fig. 2h). The intersection of these and the differentially expressed genes yielded 22 upregulated genes highly associated with tryptophan metabolism (Fig. 2i).

Cross analysis of the overlapping genes based on bulk data

To verify the reliability of the 22 selected genes, we analyzed them within the bulk data. By cross-referencing these key genes identified from the scRNA-seq data with those in the bulk data, we identified 21 overlapping genes. Results showed that 10 of these genes, including ENO1, SOD2, CTSS, S100A11, and TUBB, were up-regulated in the UC group (Fig. 3a). Among these up-regulated genes, SOD2, TUBB, CTSS, HLA-DPA1, ENO1, and ANXA2 exhibited high DC scores and were closely connected with other genes (Fig. 3b). This was further corroborated by the heatmap plot (Fig. 3c). Moreover, a GO enrichment analysis across biological process (BP), cellular component (CC), and molecular function (MF) levels revealed that these 21 genes were strongly associated with biological processes, particularly immune and inflammatory responses (Fig. 3d).

Identification of the optimal genes by machine learning

Five machine learning algorithms, including the LASSO algorithm (Fig. 4a), the Boruta algorithm (Fig. 4b), the GBM algorithm (Fig. 4c), the SVM algorithm (Fig. 4d), and the random forest model (Fig. 4e), were employed to screen for the most relevant candidate feature genes

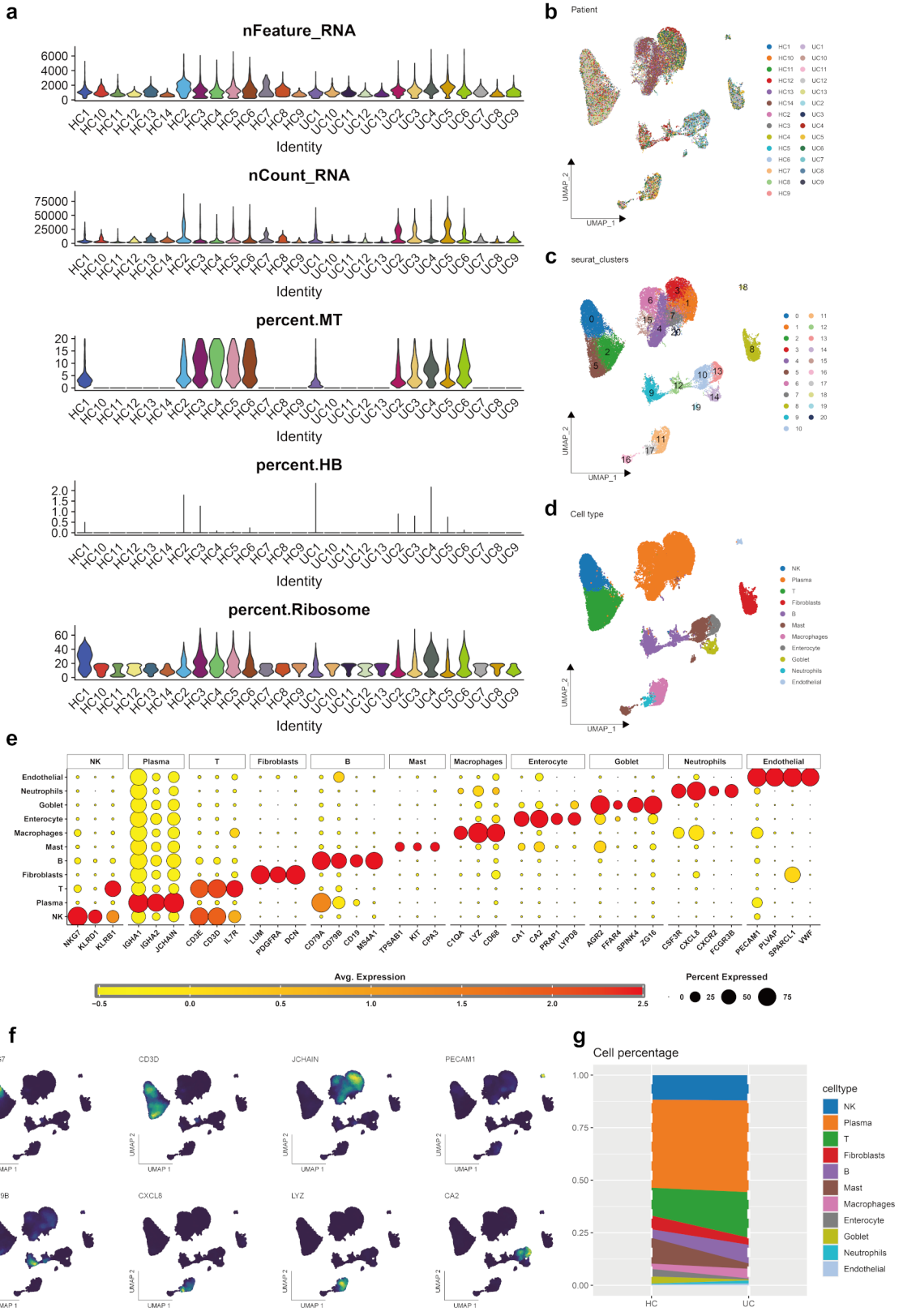

Fig. 1 Explanation of cellular subpopulations. (**a**) Quality control for inclusive data. (**b**) Excluding batch effects between samples. (**c**) Seurat clusters of eligible cells in umap plot. (**d**) Cellular annotations unveil 11 distinct cell phenotypes. (**e**) Bubble plot of relative expression of marker genes for each cell type. (**f**) UMAP plot reveals marker gene expression levels across diverse cell types. (**g**) The proportion of each single cells in HC and UC groups

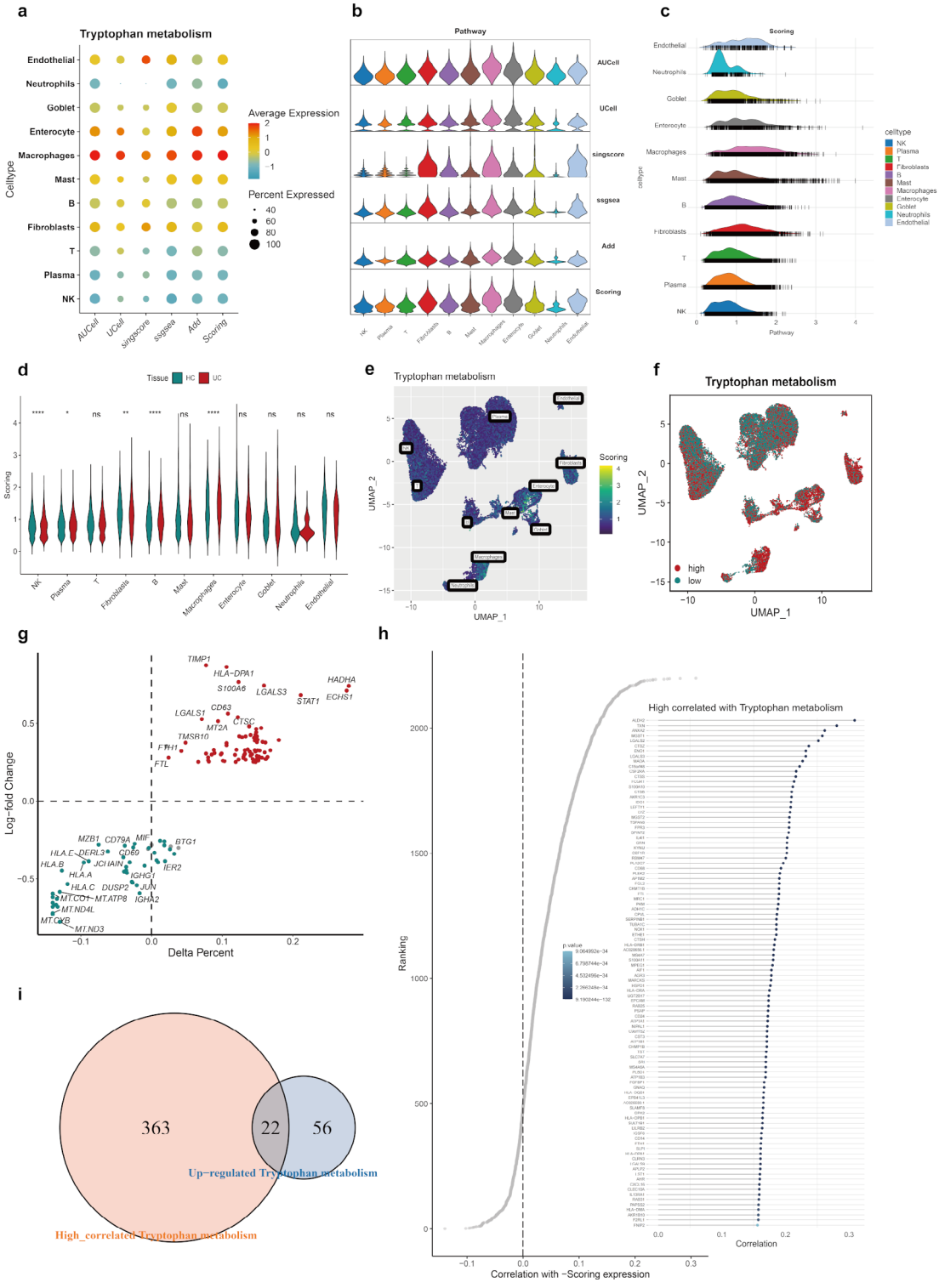

Fig. 2 Heterogeneity among the expression of TrMGs. (**a**, **b**, **c**) Bubble plot (**a**), Violin plot (**b**), and Density map (**c**) showed expression scores of TrMGs for each cell type using AUCell, UCell, singscore, ssGSEA, and Add algorithms. (**d**) Violin plot showed the difference in TrMGs score of the HC and UC groups. (**e**) UMAP plot showed the activity of TrMGs. (**f**) Scoring groups reflecting TrMGs activity in each cell was projected, with red denoting the high group and blue indicating the low group. (**g**) Percentage difference (Delta means percent of cells) and log-fold change based on the Wilcoxon rank-sum test results for differential expressed genes between high and low expression group. (**h**) Correlation analysis between Scoring expression and TrMGs. (**i**) Venn plots identified the up-regulated genes most associated tryptophan metabolism

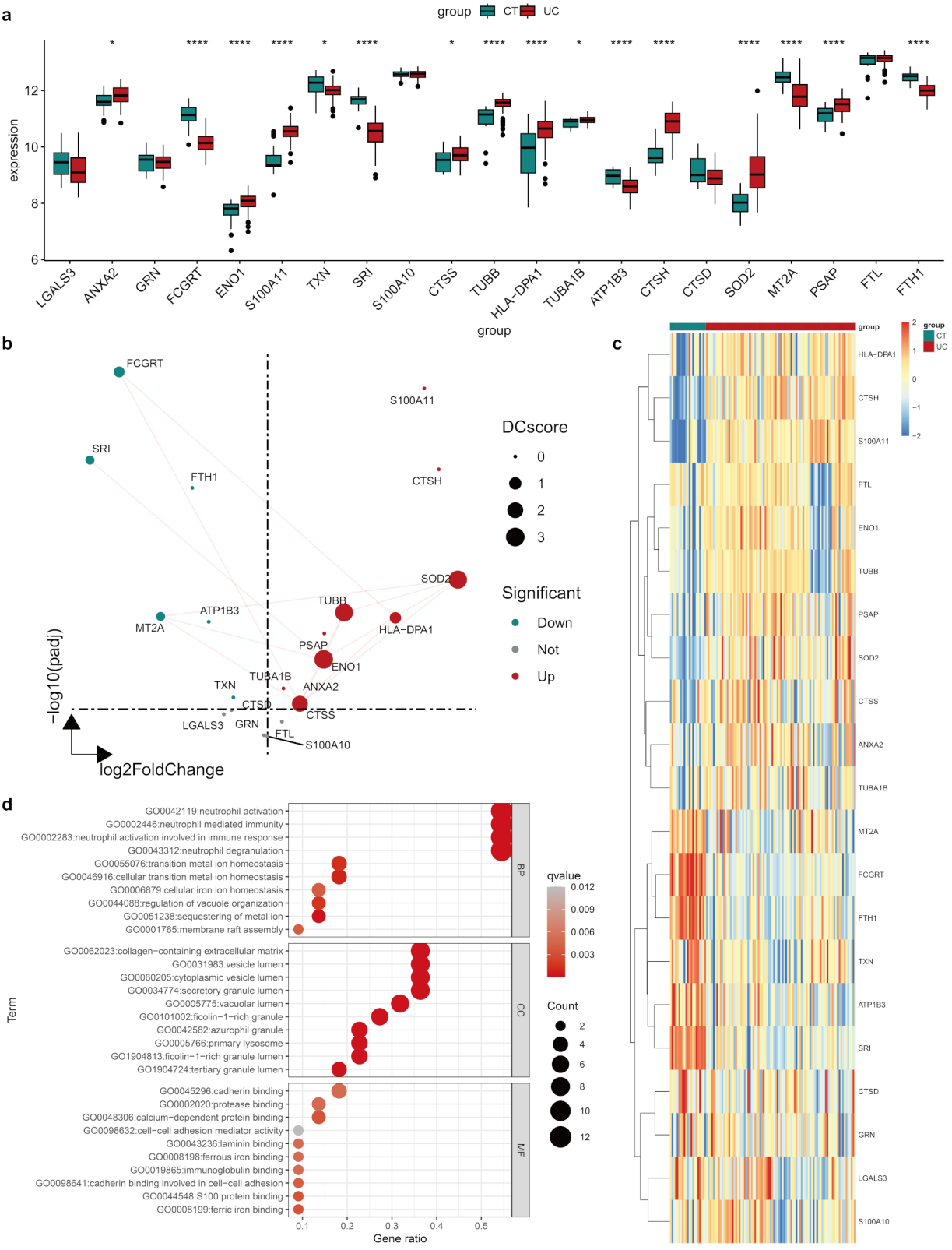

Fig. 3 Cross analysis of the key genes based on bulk data. (**a**) The expression of overlapping genes in bulk data. (**b**) The volcano plot showed the significances and links in overlapping genes. (**c**) The heatmap of overlapping genes expression: blue means low expression; red means high expression. (**d**) Gene Ontology (GO) enrichment analysis of the overlapping genes

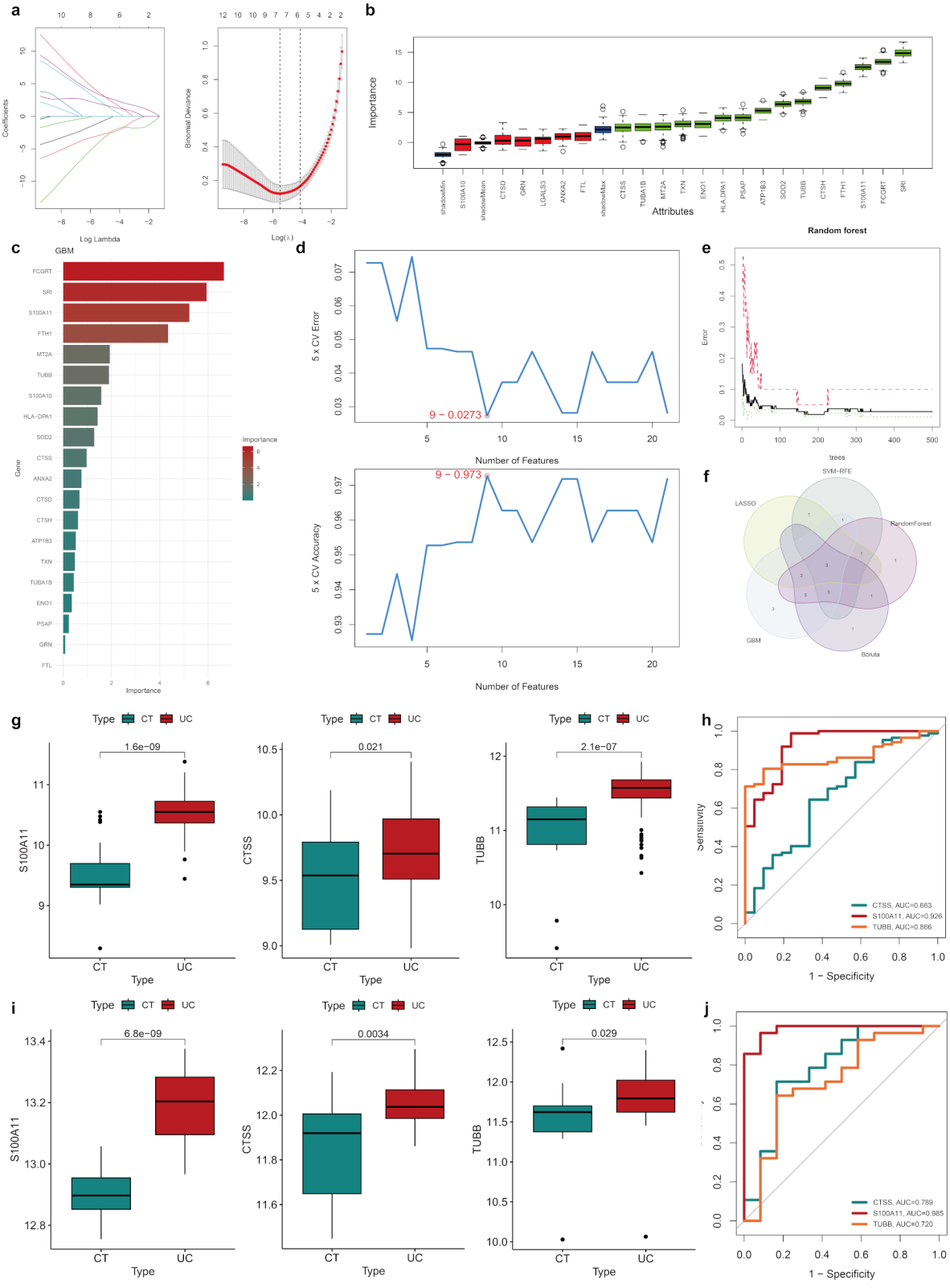

Fig. 4 Identification of the marker genes by using machine learning. (**a-e**) The LASSO algorithm (**a**), Boruta algorithm (**b**), GBM algorithm (**c**), SVM algorithm (**d**) and random forest (**e**) determined the candidate optimal feature genes. (**f**) Venn diagram displayed the three optimal key genes overlapped by the five above-mentioned machine learning outputs. (**g**) The expression levels of the three genes in the training set. (**h**) ROC curves in the training set. (**i**) The expression levels of the three genes in the validation set. (**j**) ROC curves in the validation set

in the training set. Through intersection analysis of the marker genes co-selected by these algorithms, three feature genes were identified: S100A11, CTSS, and TUBB (Fig. 4f). We further clarified the distribution and prediction efficiency of the three genes in the training set. The results showed that S100A11, CTSS, and TUBB were significantly up-regulated in the UC group (Fig. 4g). According to receiver operator characteristic (ROC) curve analysis, all three genes had good diagnostic performance, with S100A11(AUC=0.926) having the strongest distinguishing ability, followed by TUBB (AUC=0.866), and CTSS (AUC=0.663) (Fig. 4h). The reliability of these core genes was then re-examined in an external validation set. The results showed the expression of these genes was significantly elevated in the UC group (Fig. 4i), and also showed good diagnostic ability (Fig. 4j).

Validation CTSS gene in scRNA-seq data

We identified core genes in sc-RNA seq that up-regulated tryptophan metabolism in macrophages. TUBB was predominantly expressed in B cells and fibroblasts, CTSS showing significant expression in macrophages, and S100A11 was highly expressed in macrophages, fibroblasts, and neutrophils (Fig. 5a and b). A comparative analysis of these genes revealed that CTSS had the most concentrated expression and the highest average expression level in macrophages (Fig. 5c). In the high-expression and low-expression groups previously defined (Fig. 2f), all three genes were significantly enriched in the high-expression group ($p<0.01$) (Fig. 5d). Furthermore, we conducted a correlation analysis using the bulk data, comparing the expression of these three genes against 48 TrMGs. The analysis demonstrated that TUBB, CTSS, and S100A11 all showed significant positive correlations with these genes, with CTSS exhibiting the highest correlation (Pearson $r=0.62$, $p=6.18e-13$) (Fig. 5e and f). Based on the expression levels of the CTSS gene in macrophages, we divided the samples into high and low CTSS expression groups and performed a GSEA enrichment analysis using the KEGG database. The results indicated that the tryptophan metabolism pathway was significantly up-regulated in the high CTSS expression group (NSE=1.2918) (Fig. 5g).

Cellular communication and trajectory analysis in CTSS + macrophages

To elucidate the biological function of the CTSS gene in macrophages, we classified macrophages from UC samples into two groups based on CTSS expression: CTSS+ (1213 cells) and CTSS- (230 cells). We analyzed the quantity and intensity of cellular communication between the two groups and other cell types. The analysis revealed that CTSS+macrophages significantly engaged in more interactions with other cells, particularly with neutrophils, endothelial cells, and fibroblasts, suggesting potential synergistic relationships (Fig. 6a). Additionally, the GALECTIN and VEGF pathways were more active in CTSS+than in CTSS- macrophages. Among the incoming signals, IFN-II, IL16, and CSF were predominantly expressed in CTSS+macrophages (Fig. 6b). Furthermore, CTSS+macrophages exhibited a higher total volume of intercellular communication compared to CTSS- macrophages (Fig. 6c). Figure 6d further investigated the ligand-receptor interactions between various cell types and both CTSS+and CTSS- macrophages within UC intestinal mucosal tissues. It was found that CTSS+macrophages communicated with endothelial through NAMPT - (ITGA5+ITGB1) and NAMPT-INSR signaling pathways. In terms of signal reception, endothelial, neutrophils, and NK cells communicated more frequently with CTSS+macrophages via ANXA1-FPR1, NAMPT - (ITGA5+ITGB1), and ANXA1-FPR1 ligand-receptors. By using Monocle 2 for trajectory analysis of macrophages, we observed that as pseudo-time progressed, macrophages differentiated from left to right, with the proportion of CTSS+cells initially increasing in the early stages of development before declining in the later stages (Fig. 6e). CytoTRACE was also employed to further define the developmental order and starting point of macrophages, revealing that CTSS+macrophages primarily resided in the early stages of development (Fig. 6f). Examination across different groups showed that CTSS expression in macrophages gradually increased over time, aligning closely with the distribution in the UC group (Fig. 6g).

Discussion

The dysregulation of tryptophan metabolism has been increasingly recognized as a pivotal factor in the pathophysiology of UC. Prior researches have established that tryptophan metabolism through the kynurenine pathway plays a crucial role in modulating and activating immune responses in various inflammatory conditions, including UC [6, 9]. However, the regulation mechanism and the key genes still unclear. In this study, we utilized scRNA-seq integrated with machine learning algorithms to dissect the complex interplay of tryptophan metabolism in UC. Our innovative approach allowed for the precise identification of cellular subtypes that exhibit distinct metabolic signatures linked to UC pathogenesis.

Notably, we identified three genes—CTSS, S100A11 and TUBB—that are significantly upregulated and strongly associated with the dysregulated tryptophan metabolic pathways in affected UC. These findings have never been reported before, and are supported by our differential gene expression analyses and the functional enrichment analyses, which further validated the significant upregulation of these genes in UC compared to

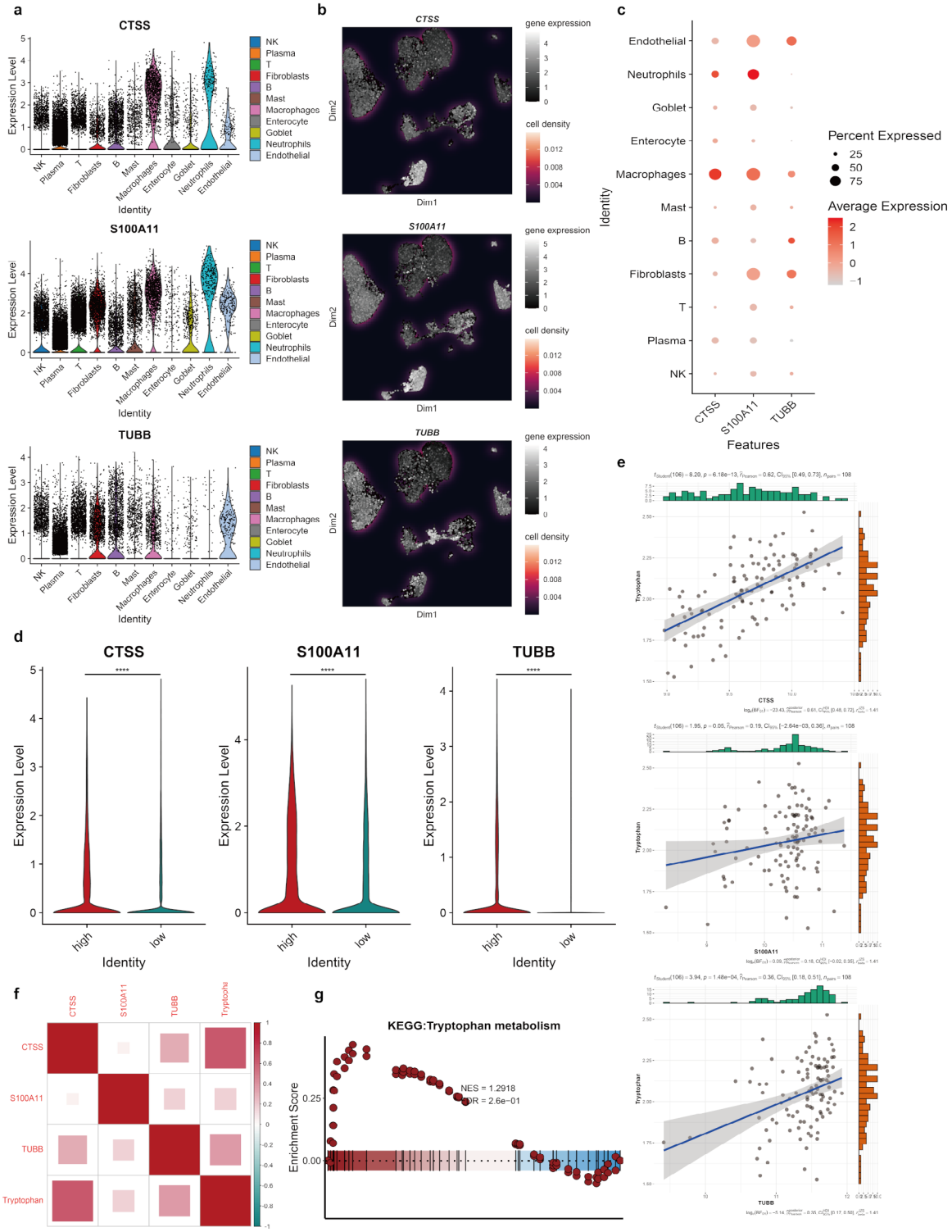

Fig. 5 Validation CTSS gene in scRNA-seq data. (**a-b**) The distribution of TUBB, CTSS, and S100A11 expression across different cells types. (**c**) Bubble map illustrated the percentage of cells expressing each gene in different cell types. (**d**) Violin plots showed the distribution of expression levels for CTSS, S100A11, and TUBB in high and low tryptophan metabolism expression group with $p < 0.01$. (**e**) Correlation analysis between CTSS, S100A11, and TUBB against 48 TrMGs. (**f**) The heatmap showed the correlation between CTSS, S100A11, TUBB and tryptophan metabolism. (**g**) Gene set enrichment analysis (GSEA) enrichment analysis focused on CTSS expression in macrophages with tryptophan metabolism

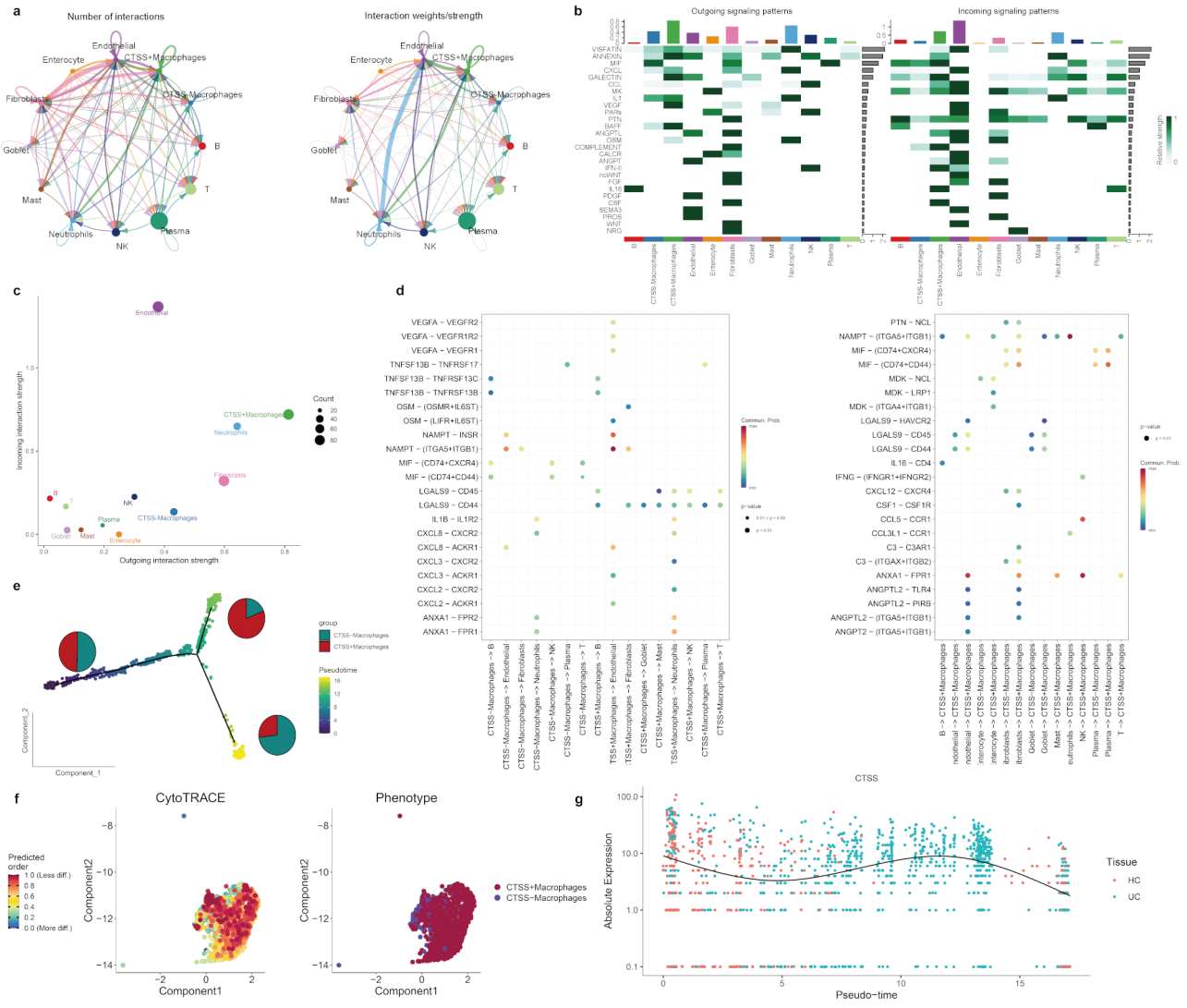

Fig. 6 Cellular communication and trajectory analysis in CTSS + macrophages. (**a**) Number and strength of cellular communications between CTSS + macrophages and other type cells. (**b**) Signaling patterns between macrophages and other cell types. (**c**) The interaction dynamics among different cell types. (**d**) Ligand-receptor bubble diagram of different types of cells acting on CTSS + and CTSS- macrophages. (**e**) Macrophages' differentiation trajectories, pseudotime distribution, and cell clusters on pseudotime. (**f**) CytoTRACE and phenotype of CTSS + macrophages. (**g**) Absolute expression of CTSS + macrophages in pseudo-time

healthy controls. Moreover, the application of machine learning algorithms enhanced the robustness of our gene selection process, ensuring that the genes identified are not only statistically significant but also biologically relevant to UC pathogenesis. The identification of these genes provides new insights into the molecular mechanisms underpinning UC and highlights potential targets for therapeutic intervention.

CTSS (Cathepsin S) is a protease highly expressed in immune cells such as macrophages, where it plays a key role in protein degradation and antigen presentation. Increased activity of CTSS can contribute to the breakdown of the extracellular matrix and disruption of epithelial barrier function [37]. Notably, CTSS are found primarily in immune cells, including antigen-presenting cells, B cells, dendritic cells and macrophages, showing its special function in immune system. In UC, macrophages destroy extracellular matrix by secreting CTSS, which may aggravate colitis by promoting paracellular permeability and influx of inflammatory cells [38]. Moreover, CTSS is expressed strongly macrophages in colon tissue of UC, and are preferentially secreted into the colon lumen, amplifying the visceral motor response to rectal dilation and inducing overexcitation of colonic pain receptors [39]. CTSS expression has been associated with tryptophan metabolism-related pathways, either. For instance, in breast cancer, the expression pattern of CTSS correlates with Tryptophan hydroxylase 1 (TPH1)

and 5-Hydroxytryptamine receptor 7 (5-HT7), and knockdown of these genes downregulates CTSS expression, suggesting a strong link between CTSS and the tryptophan metabolic pathway [40]. Given these roles, CTSS may serve as a therapeutic target, where inhibiting CTSS activity could potentially reduce UC-related inflammation and alleviate symptoms.

S100A11, a member of the S100 protein family, is implicated in various inflammatory and metabolic diseases, where it influences cell proliferation, differentiation, and cytokine production [41–43]. In rheumatoid arthritis, S100A11 are increased in patients' synovial tissue and synovial fluid, stimulating the synthesis of the pro-inflammatory mediator IL-6, suggesting an association in inflammation and disease activity [44]. Additionally, S100A11 expression correlates with HbA1c levels, indicating a role in glucose metabolism and in the pathogenesis of type 2 diabetes (T2D), thus further supporting its association with metabolic and inflammatory pathways [45]. TUBB (tubulin beta class I), a structural component of microtubules, is essential for cell division, intracellular signaling, and transport. Upregulation of TUBB in immune cells is associated with increased cell migration and motility, which are critical for immune responses in inflammatory diseases like UC [46, 47]. The upregulation in TUBB has been considered with worse prognosis, metastasis, and tumor cell survival in breast cancer, lung adenocarcinoma and other malignant tumor [48–50]. The migration of immune cells to inflammatory tissue also needs tubulin proteins in cell division and motility. This migratory function in immune cells highlights TUBB as a possible intervention target in UC, where moderating immune cell mobility could help manage disease activity. These connections between our findings and the existing literatures underscore the importance of investigating the regulatory mechanisms underlying tryptophan metabolism in UC and highlight the potential of CTSS, S100A11, and TUBB as therapeutic targets in UC.

Another interesting finding is that the activity of TrMGs in different cell types showed great heterogeneity, and are more active and up-regulated in macrophages. This elevated activity in macrophages suggests their critical role in UC pathogenesis, as they are key players in the inflammatory response and tissue remodeling associated with the disease. Macrophages, as a type of innate immune cell, significantly increased during active phase of UC, indicating their potential involvement in the inflammatory response and tissue remodeling processes associated with UC [51]. Macrophages engage multiple tryptophan metabolism pathways, which allow them to influence immune responses by producing metabolites like kynurenine and indole derivatives with diverse biological activities. The kynurenine pathway, for instance, is known to inhibit T cell proliferation and activation and to induce the generation of regulatory T cells, thereby playing a role in maintaining immune tolerance [8]. This immunosuppressive effect is particularly relevant in the inflamed intestinal environment of UC, where immune balance is disrupted. Additionally, tryptophan metabolites interact with the aryl hydrocarbon receptor (AhR), which plays a key role in maintaining intestinal homeostasis by regulating interactions between epithelial cells and macrophages. This suggests that tryptophan metabolism could influence not only immune cell behavior but also intestinal barrier integrity, further contributing to UC pathology [52, 53]. The observed heterogeneity of TrMG activity across cell types provides new insights into the cell-specific roles of tryptophan metabolism in UC, highlighting macrophages as potential therapeutic targets due to their significant contribution to immune dysregulation and inflammation in UC.

Our study also has several limitations that must be acknowledged. Firstly, relying on public datasets to build this model may introduce biases related to dataset variability and patient selection, though we have used multiple algorithms and databases to minimize these biases. More diverse cohorts and in vivo or in vitro experiments could be conducted to verify the results. Additionally, while we employed multiple machine learning algorithms to improve robustness, algorithm performance could still be influenced by data biases and feature selection, potentially affecting the reliability of identified genes. Applying cross-validation with independent datasets would help to enhance the model's generalizability. Moreover, our study focuses primarily on transcriptomic alterations within the UC landscape, which does not account for post-transcriptional modifications, protein-level changes, or the metabolic environment, all of which play crucial roles in disease pathology. This highlights the need for integrative studies that combine transcriptomics with proteomics and metabolomics to provide a more comprehensive understanding of UC.

Conclusion

Our study identified three key genes—CTSS, S100A11, and TUBB—associated with dysregulated tryptophan metabolism, providing new insights into UC pathogenesis. Our findings emphasize the critical role of macrophages in the inflammatory response of UC, with significant heterogeneity observed in tryptophan metabolism activity across different cell types. These genes not only enhance our understanding of UC's metabolic alterations but also highlight potential therapeutic targets for future interventions.

Abbreviations

5-HT	5-hydroxytryptamine
5-HT7	5-Hydroxytryptamine receptor 7
AUCell	Area Under the Curve cell-level

AUC	Area under the cumulative distribution curve
BH	Benjamini and Hochberg
BP	Biological process
CC	Cellular component
DEGs	Differentially expressed genes
FDR	False discovery rate
GBM	The Gradient Boosting Machine
GO	Gene Ontology
IBD	Inflammatory bowels disease
IDO1	Indoleamine 2,3-dioxygenase 1
Kyn	Kynurenine
LASSO	Least absolute shrinkage and selection operator
MF	Molecular function
OS	Overall survival
PCA	Principal component analysis
scRNA-seq	Single-cell RNA sequencing
ssGSEA	Single-sample Gene Set Enrichment Analysis
T2D	Type 2 diabetes
TPH1	Tryptophan hydroxylase 1
TrMGs	Tryptophan metabolism-related genes
Trp	Tryptophan
TUBB	Tubulin beta class I
UC	Ulcerative colitis
UMAP	Uniform manifold approximation and projection

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-024-05934-w.

Supplementary Material 1

Acknowledgements
This work was supported by China Health and Medical Development Foundation (2022-HX-6).

Author contributions
All authors contributed to the study conception and design. Qi H and Jiang L performed material preparation and Chen G conducted data analysis. The first draft of the manuscript was written by Chen G, Qi H and Jiang L. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability
The data that support the findings of this study are available from Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/).

Declarations

Consent for publication
All authors have agreed to publish this manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 30 July 2024 / Accepted: 1 December 2024
Published online: 20 December 2024

References

1. Le Berre C, Honap S, Peyrin-Biroulet L. Ulcerative colitis. Lancet Lond Engl. 2023;402(10401):571–84.
2. da Silva BC, Lyra AC, Rocha R, Santana GO. Epidemiology, demographic characteristics and prognostic predictors of ulcerative colitis. World J Gastroenterol. 2014;20(28):9458–67.
3. Singh N, Bernstein CN. Environmental risk factors for inflammatory bowel disease. United Eur Gastroenterol J. 2022;10(10):1047–53.
4. Schirmer M, Stražar M, Avila-Pacheco J, et al. Linking microbial genes to plasma and stool metabolites uncovers host-microbial interactions underlying ulcerative colitis disease course. Cell Host Microbe. 2024;32(2):209–e2267.
5. Sofia MA, Ciorba MA, Meckel K, et al. Tryptophan Metabolism through the Kynurenine Pathway is Associated with endoscopic inflammation in Ulcerative Colitis. Inflamm Bowel Dis. 2018;24(7):1471–80.
6. Nikolaus S, Schulte B, Al-Massad N, et al. Increased tryptophan metabolism is Associated with Activity of Inflammatory Bowel diseases. Gastroenterology. 2017;153(6):1504–e15162.
7. Zhu H, Yang X, Zhao Y. Recent advances in current Uptake Situation, Metabolic and Nutritional Characteristics, Health, and Safety of Dietary Tryptophan. J Agric Food Chem. 2024;72(13):6787–802.
8. Xue C, Li G, Zheng Q, et al. Tryptophan metabolism in health and disease. Cell Metab. 2023;35(8):1304–26.
9. Shi Y, Luo S, Zhai J, Chen Y. A novel causative role of imbalanced kynurenine pathway in ulcerative colitis: Upregulation of KMO and KYNU promotes intestinal inflammation. Biochim Biophys Acta Mol Basis Dis. 2024;1870(2):166929.
10. Harris DMM, Szymczak S, Schuchardt S, et al. Tryptophan degradation as a systems phenomenon in inflammation - an analysis across 13 chronic inflammatory diseases. EBioMedicine. 2024;102:105056.
11. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: a brief overview. Clin Transl Med. 2022;12(3):e694.
12. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol. 2018;18(1):35–45.
13. Guan J, Xu X, Qiu G, et al. Cellular hierarchy framework based on single-cell/multi-patient sample sequencing reveals metabolic biomarker PYGL as a therapeutic target for HNSCC. J Exp Clin Cancer Res. 2023;42(1):162.
14. Lin Y, Jing X, Chen Z, et al. Histone deacetylase-mediated tumor microenvironment characteristics and synergistic immunotherapy in gastric cancer. Theranostics. 2023;13(13):4574–600.
15. Shen X, Mo S, Zeng X, et al. Identification of antigen-presentation related B cells as a key player in Crohn's disease using single-cell dissecting, hdWGCNA, and deep learning. Clin Exp Med. 2023;23(8):5255–67.
16. Cui EH, Zhang Z, Chen CJ, Wong WK. Applications of nature-inspired metaheuristic algorithms for tackling optimization problems across disciplines. Sci Rep. 2024;14(1):9403.
17. Duan X, Zhang C, Tan X, et al. Exploring optimization algorithms for establishing patient-based real-time quality control models. Clin Chim Acta. 2024;554:117774.
18. Deng Y, Feng Y, Lv Z, et al. Machine learning models identify ferroptosis-related genes as potential diagnostic biomarkers for Alzheimer's disease. Front Aging Neurosci. 2022;14:994130.
19. Li H, Sun X, Li Z, Zhao R, Li M, Hu T. Machine learning-based integration develops biomarkers initial the crosstalk between inflammation and immune in acute myocardial infarction patients. Front Cardiovasc Med. 2023;9:1059543.
20. Zhao X, Duan L, Cui D, Xie J. Exploration of biomarkers for systemic lupus erythematosus by machine-learning analysis. BMC Immunol. 2023;24(1):44.
21. Boland BS, He Z, Tsai MS, et al. Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. Sci Immunol. 2020;5(50):eabb4432.
22. Garrido-Trigo A, Corraliza AM, Veny M, et al. Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. Nat Commun. 2023;14(1):4506.
23. Luo P, Chen G, Shi Z, et al. Comprehensive multi-omics analysis of tryptophan metabolism-related gene expression signature to predict prognosis in gastric cancer. Front Pharmacol. 2023;14:1267186.
24. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
25. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72.
26. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96.
27. Liu Y, Li H, Zeng T, et al. Integrated bulk and single-cell transcriptomes reveal pyroptotic signature in prognosis and therapeutic options of hepatocellular carcinoma by combining deep learning. Brief Bioinform. 2023;25(1):bbad487.
28. Andreatta M, Carmona SJ, UCell. Robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;19:3796–8.
29. Mao Y, Gide TN, Adegoke NA, et al. Cross-platform comparison of immune signatures in immunotherapy-treated patients with advanced melanoma using a rank-based scoring approach. J Transl Med. 2023;21(1):257.

30. Jin Y, Wang Z, He D, et al. Identification of novel subtypes based on ssGSEA in immune-related prognostic signature for tongue squamous cell carcinoma. Cancer Med. 2021;10(23):8693–707.
31. Mei Y, Li M, Wen J, et al. Single-cell characteristics and malignancy regulation of alpha-fetoprotein-producing gastric cancer. Cancer Med. 2023;12(10):12018–33.
32. Foulquier N, Le Dantec C, Bettacchioli E, et al. Machine learning for the identification of a common signature for Anti-SSA/Ro 60 antibody expression across Autoimmune diseases. Arthritis Rheumatol Hoboken NJ. 2022;74(10):1706–19.
33. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
34. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics. 2018;19(1):432.
35. Zhang K, Ye B, Wu L, et al. Machine learning–based prediction of survival prognosis in esophageal squamous cell carcinoma. Sci Rep. 2023;13(1):13532.
36. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform. 2023;24(2):bbad002.
37. Smyth P, Sasiwachirangkul J, Williams R, Scott CJ. Cathepsin S (CTSS) activity in health and disease - A treasure trove of untapped clinical potential. Mol Aspects Med. 2022;88:101106.
38. Reddy VY, Zhang QY, Weiss SJ. Pericellular mobilization of the tissue-destructive cysteine proteinases, cathepsins B, L, and S, by human monocyte-derived macrophages. Proc Natl Acad Sci U S A. 1995;92(9):3849–53.
39. F C. V L, E J, Cathepsin S is activated during colitis and causes visceral hyperalgesia by a PAR2-dependent mechanism in mice. Gastroenterology. 2011;141(5).
40. Gautam J, Bae YK, Kim JA. Up-regulation of cathepsin S expression by HSP90 and 5-HT7 receptor-dependent serotonin signaling correlates with triple negativity of human breast cancer. Breast Cancer Res Treat. 2017;161(1):29–40.
41. Zeng X, Guo H, Liu Z, et al. S100A11 activates the pentose phosphate pathway to induce malignant biological behaviour of pancreatic ductal adenocarcinoma. Cell Death Dis. 2022;13(6):568.
42. Sobolewski C, Abegg D, Berthou F, et al. S100A11/ANXA2 belongs to a tumour suppressor/oncogene network deregulated early with steatosis and involved in inflammation and hepatocellular carcinoma development. Gut. 2020;69(10):1841–54.
43. Zhang L, Zhu T, Miao H, Liang B. The calcium binding protein S100A11 and its roles in diseases. Front Cell Dev Biol. 2021;9:693262.
44. L AC, B Š, K P, et al. Calgizzarin (S100A11): a novel inflammatory mediator associated with disease activity of rheumatoid arthritis. Arthritis Res Ther. 2017;19(1).
45. Eo JFPV. L, Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc Natl Acad Sci U S A. 2014;111(38).
46. Sferra A, Petrini S, Bellacchio E, et al. TUBB variants underlying different phenotypes result in altered vesicle trafficking and Microtubule dynamics. Int J Mol Sci. 2020;21(4):1385.
47. La MKVC. K, B G. The mechanism of Tubulin Assembly into microtubules: insights from Structural studies. iScience. 2020;23(9).
48. Alhammad R. Bioinformatics Identification of TUBB as potential prognostic biomarker for worse prognosis in ERα-Positive and better prognosis in ERα-Negative breast Cancer. Diagnostics. 2022;12(9).
49. E FB. J, H K, H F. Intrinsic and extrinsic factors affecting Microtubule dynamics in Normal and Cancer cells. Mol Basel Switz. 2020;25(16).
50. Jm XYYZBW. K, A P. The miR-195 Axis regulates Chemoresistance through TUBB and Lung Cancer Progression through BIRC5. Mol Ther Oncolytics. 2019;14.
51. Na YR, Stakenborg M, Seok SH, Matteoli G. Macrophages in intestinal inflammation and resolution: a potential therapeutic target in IBD. Nat Rev Gastroenterol Hepatol. 2019;16(9):531–43.
52. Hezaveh K, Shinde RS, Klötgen A, et al. Tryptophan-derived microbial metabolites activate the aryl hydrocarbon receptor in tumor-associated macrophages to suppress anti-tumor immunity. Immunity. 2022;55(2):324–e3408.
53. Yu K, Li Q, Sun X, et al. Bacterial indole-3-lactic acid affects epithelium-macrophage crosstalk to regulate intestinal homeostasis. Proc Natl Acad Sci U S A. 2023;120(45):e2309032120.

Publisher's note