A Seven-NF-κB-Related Gene Signature May Distinguish Patients with Ulcerative Colitis-Associated Colorectal Carcinoma

Chao-Yi Ge1,2,*
Li-Yuan Wei1,2,*
Yu Tian1
Hua-Hong Wang1

1Department of Gastroenterology, Peking University First Hospital, Beijing, People’s Republic of China; 2Department of Breast Surgery, Shanxi Bethune Hospital, Taiyuan, People’s Republic of China

*These authors contributed equally to this work

Purpose: Ulcerative colitis (UC) patients have an increased risk of colorectal cancer (CRC), and compared with sporadic CRC, ulcerative colitis-associated colorectal cancer (CAC) is more aggressive with a worse prognosis. This study aimed to identify a gene signature to predict the risk of CAC for patients with UC in remission.

Patients and Methods: Series of quiescent UC-related transcriptome data obtained from the Gene Expression Omnibus (GEO) data set were divided into a training set and a validation set. Gene Set Variation Analysis (GSVA), Gene Set Enrichment Analysis (GSEA), and Weighted Correlation Network Analysis (WGCNA) combined with protein–protein interaction (PPI) analysis were used to identify the pathways and gene signatures related to tumorigenesis among quiescent UC patients. A generalized linear model (GLM) of Poisson regression based on the training set was applied to estimate the diagnostic power of the gene signature in our validation set.

Results: The tumor necrosis factor (TNF) signaling via NF-κB pathway was significantly augmented with the highest normalized enrichment score (NES). The genes in the brown module from WGCNA have shown a significant correlation with CAC (Pearson coefficient = 0.83, p = 6e-06). A subset of NF-κB related genes (FOS, CCL4, CXCL1, MYC, CEBPB, ATF3, and JUNB) were identified with a relatively higher expression level in CAC samples. The diagnostic value of this 7-gene biomarker was estimated by the receiver operating characteristic (ROC) curve with an area under the ROC curve (AUC) at 0.82 (p<0.0001, 95% CI: 0.7098–0.9400) in the validation cohort.

Conclusion: In summary, the increased expression of this seven-NF-κB-related gene signature may act as a powerful index for tumorigenesis prediction among patients with UC in remission.

Keywords: ulcerative colitis; UC, remission, colitis-associated colorectal cancer; CAC, diagnostic model

Introduction
Ulcercative colitis (UC) is a kind of chronic idiopathic intestinal inflammatory disease with poorly understood etiology, mainly affecting the epithelial mucosa from the anus to the ileocecal area.1 It has been demonstrated that the extent of colonic involvement, severity of the inflammation, and duration of UC could increase the risk of colorectal cancer (CRC).2,3 Patients with inflammatory bowel disease (IBD) often suffer from CRC within the first 7 years after initial diagnosis.4 Moreover, compared with sporadic CRC, ulcerative colitis-associated colorectal...
cancer (CAC) is more aggressive and has a worse prognosis as evidenced by the multiple involved sites, and hence surveillance of CAC by colonoscopy is recommended among the UC patients during the remission state. Moreover, a proportion of dysplastic lesions cannot be detected exactly by surveillance colonoscopy. The researches on the assessment of the risk of CRC occurring between each two scheduled surveillance procedures are still limited. Therefore, this calls for further development of methods to predict and detect UC patients with high a risk of CAC at an earlier stage to improve the clinical outcomes during the scheduled surveillance procedures.

The progression of CAC from UC including the development from low-grade dysplasia, high-grade dysplasia, to invasive adenocarcinoma, is reported to have a dynamic, heterogeneous, and complex communication between the immune system and cytokines. Various immunological and inflammatory pathways including PI3K-Akt signaling, tumor necrosis factor (TNF) signaling, cytokine–cytokine receptor interaction, and extracellular matrix (ECM)–receptor interaction pathways have been confirmed to orchestrate the fate of tumorigenesis and progression. Moreover, some of the candidate biomarkers were demonstrated to be able to promote colonic tumorigenesis by regulating the immune system, such as CXCL10 and CXCL9, IDO1, CCR7, VCAM1 as well as ICAM1.

Considering the comprehensive biological process, we have compared the differentially expressed genes in intestinal epithelium biopsy tissues among patients with UC in remission or UC with remote neoplasm and normal individuals from the Gene Expression Omnibus (GEO) database. A group of 7 nuclear factor-kappa B (NF-kB) related genes were generated with up-regulated expression among UC with remote neoplasm tissues and then was validated to be an effective signature for discrimination of UC patients with a high risk of CAC.

**Patients and Methods**

**Patients and Samples**

In our study, both the discovery cohort and the validation cohort were based on patients with UC in remission exclusively. The patients with active UC and Crohn’s disease were not included. The discovery cohort consisted of 20 patients from Chicago including 5 normal controls, 4 UC patients in remission, and 11 patients with remote neoplasm. The total RNA of the 20 samples extracted from the colonic mucosae was analyzed via the Affymetrix Human U133p2 platform (GLP570) and the normalized microarray data was obtained from GEO (https://www.ncbi.nlm.nih.gov/geo) repository via the accession number of GSE37283. The validation cohort used in our study was composed of 41 normal controls, 26 UC patients in remission, as well as 15 patients with CRC derived from the other 3 GEO data sets (GSE13367, GSE38713, and GSE4183), in which the normalized microarray data generated from colonic mucosae of these 82 samples was also analyzed through the GLP570 platform. All of the samples in both discovery cohort and validation cohort were obtained with the ethical approval from their original institutions.

**Profiling of RNA Differential Expression**

Both the normalized microarray data and the corresponding clinical features were downloaded from the GEO database and statistical analyses were performed in the R platform (version 3.6.2). Differentially expressed genes were identified by the Limma package from the comparison among the normal control, UC in remission, CAC, and CRC groups. To merge the microarray data from these 3 different GEO data sets for our validation cohort, the existed bias and variation due to the batch effect of the different high-throughput data were removed by the Combat function of the sva package in the R platform.

**Gene Set Enrichment Analyses**

To investigate the variation of the Hallmark gene set enrichment among each sample in our discovery cohort (downloaded from Molecular Signatures Database (MSigDB); https://www.gsea-msigdb.org/gsea/msigdb/index.jsp), Gene Set variation Analysis (GSEA) was conducted by the GSEA package in R platform to calculate the sample-gene set enrichment scores, which were then visualized in a heatmap by the heatmap package. Furthermore, Gene Set Enrichment Analysis (GSEA) was carried out by the ClusterProfiler package to identify the core genes in some significantly enriched pathways based on the Hallmark gene set associated with CAC. Significantly enriched biological pathways were identified if their normalized enrichment scores (NES) were ≥2 or ≤-2 as well as false discovery rates (FDR) <0.05 after 1000-time permutations.

**Immune Cell Infiltration**

The assessment of immune infiltration of 28 immune cell types among the CAC, UC in remission, and normal control in our discovery cohort was implemented via single-sample gene set enrichment analysis (ssGSEA)
by using the GSVA package in R.23 The identified genes matched for each immune cell type were downloaded from the recent published work,25 and the ssGSEA scores for each immune cell type were then visualized in a heatmap. Considering the potential effect of these 28 immune cells on tumorigenesis, they were divided into 2 groups: an anti-tumor group and a pro-tumor group, by suppressing the immune system in the microenvironment.26

Weighted Correlation Network Analysis for the Key Module

The undirected, weighted gene co-expression network was constructed to detect the most correlated cluster of genes with CAC by using the Weighted Correlation Network Analysis (WGCNA) in our discovery cohort. The top 5,000 genes according to their median absolute deviation were extracted from the normalized microarray expression data and used for WGCNA. The selection of soft threshold, construction of correlation networks based on gene expression, detection and selection of highly correlated hub genes in the significantly related modules to CAC, calculation of softConnectivity, intramodularConnectivity, and topological overlap measure (TOM) similarity, as well as visualization of the module structure and network connections were all implemented in the WGCNA package.27

Identification of Diagnostic Gene Signature for CAC Patients

The overlapped genes including those enriched by GSEA and involved in the most significant module of WGCNA were illustrated in a Venn diagram from an online tool (http://bioinformatics.psb.ugent.be/webtools/Venn). The protein–protein interaction (PPI) information of these common genes was then integrated based on the STRING (11.0) online database (https://string-db.org/)28 with the confidence score > 0.7 between each 2 nodes, and 3 clusters were identified by K-means clustering algorithm. After removing the isolated and lower connected nodes, the achieved PPI network was further analyzed for hub genes by CytoHubba plugin29 and visualized in the Cytoscape software.30 The normalized expression of the selected hub genes in both training set and validation set was illustrated in boxplots by ggplot2 package in R platform.

Validation of the Diagnostic Signature for UC Patients with a High Risk of Colorectal Carcinoma

A generalized linear model (GLM) with Poisson regression based on our training set was employed to evaluate the diagnostic power of the identified signature. The predicted diagnostic scores were generated from the weighted linear diagnostic model combined with the gene expression in our validation cohort by the predict function in R software. The predicted risk score = expression of gene1 *β1+expression of gene2 *β2+expression of gene_n *βn. The receiver operating characteristic (ROC) curve plotted by ROCR package31 in R platform combined with the corresponding area under the ROC curve (AUC) was applied to quantify the accuracy of the gene signature in the predicted model. A two-tailed P value < 0.05 with 95% confidence intervals (CI) was supposed to be significant.

Results

Profiling of Microarray Data for CAC in the Training Set

Samples of UC patients in remission (quiescent UC), CAC patients, and normal controls were collected from GSE37283 as our discovery data set. Considering the fact that the extent, duration, as well as the severity of inflammation on the intestinal epithelial cells may increase the risk for CAC,17 we first investigated the inherent heterogeneity of the microarray-based transcriptomic profiles of these samples based on 50 pathways of the Hallmark gene set by GSVA. The hierarchical clustering diagram of the 50 gene sets variation analysis showed that the pathways on the regulation of immune system such as interferon alpha/gamma response, allograft rejection, IL6-JAK-STAT3 signaling, IL2-STAT5 signaling, TNF signaling via NF-κB displayed a strong similarity with higher expression in CAC patients compared with those in other 2 groups (Figure 1A). Then, the ssGSEA was carried out to quantify the relative abundance of immune cell infiltration among these patients. Visualized in the heatmap, we observed that UC patients with remote neoplasia could be distinguished from the other 2 groups of patients remarkably (Figure 1B). According to the potential function on tumorigenesis, these 28 immune cell populations were divided into 2 clusters: anti-tumor immunity (activated CD4 T cell (ActCD4), ActCD8, central memory CD4 T cell (TcmCD4), TcmCD8, effector memory CD4
T cell (TemCD4), TemCD8, Th1, Th17, activated dendritic cell (ActDC), CD56 bright natural kill cell (CD56brNK), natural kill T cell (NKT), NK) and pro-tumor reactivity (Treg, Th2, CD56dimNK, immature dendritic cell (imDC), tumor-associated macrophage (TAM), myeloid-derived suppressor cells (MDSCs), Neutrophil, and plasmacytoid dendritic cell (pDC)) by suppressing the immune system. 26 As depicted in the correlation diagram (Figure 1C), the relative abundance of anti-tumor immune cell types of these samples was positively related to the pro-tumor ones, especially for CAC samples (R = 0.9676, p = 3.2966e-12). This highly positive correlation may suggest that the enhancement of immune suppression could be fed back by the facilitation of anti-tumor immunity. 26 The presence of inflammation caused by UC may have an effect on tumorigenesis for CAC.

Functional Pathways Enrichment Analysis
To further explore the special mechanism of tumorigenesis among the quiescent UC patients, GSEA was used to compare the difference between CAC and UC in remission samples, between CAC and normal control samples, and between UC in remission and normal control samples based on the Hallmark gene set. Consistent with the result of GSVA, the TNF signaling via NF-kB pathway was significantly augmented with the highest NES in both CAC patients compared with UC (NES: 2.64, p = 0.0016, p.adjust = 0.0064) as well as normal controls (NES:2.47, p = 0.001.
p.adjust = 0.0027) and quiescent UC patients compared with normal controls (NES:2.17, \( p = 0.001, \) p.adjust = 0.0018) (Figure 2 and Supplementary Tables 1–3).

**Construction of Co-Expression Networks by WGCNA**

Co-expression networks among these 20 samples were established to figure out those genes with a high relationship to CAC. The top 5,000 genes of these samples were clustered into 8 modules (MEblack, MEblue, MEgreen, MEred, MEyellow, MEbrown, MEturquoise, and MEgrey), of which genes in the MEbrown module were most positively related to CAC with significance (Figure 3A and B, Pearson coefficient = 0.83, \( p = 6e^{-06} \)). There were a total of 1,234 genes involved in the brown module (Supplementary Table 4), and the correlational analysis of these genes between gene significance to CAC and the module membership was

---

**Figure 2.** GSEA plots depicting the enrichment of signal pathways based on the Hallmark gene set. Ridge plots (left) of the most enriched 20 pathways of CAC patients relative to those with UC in remission (A) or normal patients (B), and patients with UC in remission compared with normal controls (C). The TNF signaling via NF-κB pathway is positively enriched with the highest NES (right).

**Abbreviations:** GSEA, gene set enrichment analysis; CAC, ulcerative colitis-associated colorectal cancer; UC, ulcerative colitis; NES, normalized enrichment score.
conducted. As shown in Figure 3C, these 1,234 genes presented a significant contribution to both CAC and the brown module membership (correlation coefficient = 0.85, p<1e-200). Besides, the hierarchical clustering analysis indicated that these genes in the CAC group also expressed differently from the other 2 groups (Figure 3D). Therefore, these genes co-expressed with CAC should be paid attention to.

Identification of Hub Genes Involved in CAC in the Discovery Cohort

Having found that the TNF signaling via NF-κB pathway was positively enriched both CAC and UC patients, 81 NF-κB related genes were commonly upregulated in CAC patients compared with quiescent UC patients and normal controls (Figure 4A, the common part of the red and green channels). Additionally, considering the NF-κB related
genes also enriched in quiescent UC patients compared with normal controls, there were 56 NF-κB related genes commonly increased in both the CAC group and the quiescent UC group (Figure 4A). To further discriminate the NF-κB related genes with a closer correlation with CAC, genes from WGCNA were taken into account and a total of 32 NF-κB related hub genes were found (Figure 4A). The identified genes were then submitted to the STRING online database to map a PPI network and divided into 3 clusters by K-means clustering algorithm (Figure 4B). 23 nodes with high interaction score (confidence > 0.7) were discriminated and further analyzed by 12 topological algorithms including Degree, Edge Percolated Component (EPC), Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC), Maximal Clique Centrality (MCC) and centralities based on shortest paths, such as Bottleneck (BN), EcCentricity, Closeness, Radiality, Betweenness, and Stress in CytoHubba plugin29 of Cytoscape software (Table 1). Finally, 7 NF-κB related hub genes were identified as follows with a relatively higher expression level in CAC patients: Fos proto-oncogene (FOS), C-C motif chemokine ligand 4 (CCL4), C-X-C motif chemokine ligand 1 (CXCL1), MYC proto-oncogene (MYC), CCAAT enhancer binding protein beta (CEBPB), activating transcription factor 3 (ATF3), and JunB proto-oncogene (JUNB) (Figure 4C and Figure 5A).

**Performance of the Diagnostic Seven-Gene Signature in the Validation Cohort**

In order to assess the strength of the seven-NF-κB-related gene signature in predicting tumorigenesis among
quiescent UC patients, the transcriptomic data of 82 patients were integrated as our validation cohort, including 41 normal controls (20 cases in GSE13367; 13 cases in GSE38713; 8 cases in GSE4183), 15 patients with CRC (GSE4183), and 26 quiescent UC patients (18 patients in GSE13367; 8 patients in GSE38713). Consistent with the expression levels observed in our training set, these 7 NF-κB related genes were also expressed at a higher level in CRC samples, suggestive of the potential role in tumorigenesis of colorectal carcinoma (Figure 5B). Moreover, gene functional analysis also indicated that the TNF signaling via NF-κB pathway was highly augmented in CRC samples compared with quiescent UC patients or normal intestinal mucosa significantly (Figure 6A; Supplementary Table 5 and 6).

According to the clinical feature of our discovery cohort, a GLM with the Poisson regression was then applied to our validation cohort. The ROC curve with an AUC at 0.82 (p<0.0001, 95% CI: 0.7098–0.9400) demonstrated the potential ability of diagnosis for quiescent UC patients with a high risk of colorectal carcinoma (Figure 6B).

**Discussion**

The elucidation of pathogenesis of CAC remains an unmet need to develop new surveillance tools or therapeutic targets for patients with both UC and CAC. The consistent action of inflammatory on intestinal mucosa has been elucidated to predispose UC to CAC by multiple tumor-related genetic changes in intestinal epithelial cells. Watanabe et al generated 20 discriminator genes to predict the development of CAC among the UC patients, including the cancer-related genes CYP27B1, RUNX3, SAMSNI, EDIL3, NOL3, CXCL9, ITGB2, and LYN. Nevertheless, they did not distinguish the active UC patients from those in remission. Besides, the previous study based on the same population as our discovery cohort has identified nine highly expressed genes in UC patients with neoplasia compared to quiescent UC patients and healthy controls, which might regulate the immune function, inflammation, proliferation, and apoptosis of the intestinal mucosa cells and contribute to neoplastic transformation, but without validation in a large clinical cohort. Furthermore, Shi’s group integrated 8 gene profiles and reanalyzed the differentially expressed genes between patients with quiescent UC and active UC, UC and adenoma, as well as UC and CRC. A group of hub genes (CXCL10, VCAM1, CXCL9, MMP9, IDO1, and CCR7)
were found to be significantly associated with tumorigenic processes via pathways such as platelet activation, ligand-receptor interaction, immune dysregulation, and inflammation. Among the involved pathways, the aberrant stimulation of NF-κB have been detected in intestinal epithelial cells from both IBD and CRC, and can contribute to inflammation-associated tissue injury as well as tumor cell proliferation via the overexpression of angiogenesis-related genes. In our present study, we have identified a 7-NF-κB-related gene signature (FOS, CCL4, CCL1).

Figure 5 The box plots depicting the 7 hub genes differentially expressed in CAC, UC in remission, and normal control groups in our training data set (A) GSE37283) and validation data set (B) GSE4183 & GSE13367 & GSE38713). *P<0.1, **P<0.05; *** P<0.01; **** P < 0.001.

Abbreviations: UC, ulcerative colitis; CAC, ulcerative colitis-associated colorectal cancer.

Figure 6 (A) GSEA plots depicting the TNF signaling via NF-kB pathway which was positively enriched in CRC patients in our validation cohort compared with patients with UC in remission (top) and normal patients (bottom). (B) The ROC curve with AUC at 0.82 for the diagnosis value of the 7-NF-κB-signaling-gene signature in detecting UC patients with high risk for CRC (p<0.0001, 95% CI: 0.7098–0.9400).

Abbreviations: GSEA, gene set enrichment analysis; UC, ulcerative colitis; CRC, colorectal cancer; ROC, receiver operating characteristic; AUC, area under curve.
CXCL1, MYC, CEBPB, ATF3, and JUNB) which was potentially a tool for the selection of patients UC in remission stage with a high risk of cancerization during the scheduled surveillance procedures.

Considering the crucial role in UC and CAC, there is no doubt that the TNF signaling via NF-κB pathway could be highly enriched in CAC samples with significance in our study. The activation of the NF-κB pathway upregulates the expression of pro-inflammatory cytokines and adhesion factors, inhibits the production of apoptotic factors, and promotes tumor angiogenesis and metastasis. In azoxymethane (AOM) followed dextran sulfate sodium (DSS) induced CAC mice models, the deletion of IκB kinase β (IKKβ) led to inactivation of the NF-κB pathway and thereby reduced the incidence of tumors, which proved the stimulation on tumor progression. On the other hand, the activation of the NF-κB pathway promotes the production of TNF-α, which also plays an important role in tumorigenesis. The expression of TNF-α was observed to be significantly increased in AOM/DSS CAC models, and blocking TNF-α obviously reduced tumors amount and lesion as well as reduced colonic infiltration by neutrophils and macrophages. The higher expression level of these identified genes in the CAC group in our discovery and validation cohort might point out the potential importance in tumorigenesis during the development from quiescent UC to CAC. Indeed, interrogation of extra-data which was composed of the normal intestinal mucosa, UC in remission stage, and CRC also revealed an outstanding diagnostic power of this identified gene signature. Nevertheless, given the limited size of the training set and the non-coherent sample features of our validation set, the seven-NF-κB-related gene signature needs to be further validated in large scale cohorts of UC and CAC patients, and it may serve as a useful tool for predicting and identifying UC patients in remission stage with a high risk of CAC after the routine therapy.

Among the 7 NF-κB related genes, CCL4 and CXCL1 encode members of chemotactic cytokines families which are small secreted proteins and are known to control immune cell migration. Besides, they also participate in pathological processes such as tumor occurrence, development, and metastasis. MYC is a type of proto-oncogenes encoding nuclear transcriptions which regulates cell cycles. The inhibition of TNF α-induced c-Myc upregulation can be mediated by attenuating NF-κB signaling, which can then induce uncontrolled cell proliferation and inhibit apoptosis of colorectal carcinoma. ATF3 is a stress-responsive factor that plays a vital role in controlling the expression of cell-cycle regulators and tumor suppressor. Intriguingly, it has been reported that ATF3 played a protective role in regulating the gutfollicle helper T cells among IBD patients and inhibiting the invasion of CRC, which may need further investigation. CEBPB encodes a member of the CCAAT/enhancer-binding protein family, which can be activated by the NF-κB and STATS pathways in the inflammatory microenvironment to regulate gene transcription in response to IL-1 and IL-6. The deletion of CEBPB would impair the function of regular T cells and thus aggravate the T cell-mediated colitis. In contrast, the overexpression of CEBPB may promote tumor cell invasion in CRC individuals, and regulate the immunosuppressive environment via MDSCs in CAC. The JUNB and FOS encode proteins which belongs to the dimeric transcription factor activator protein 1(AP-1), play roles in regulating tumor cell proliferation, survival, differentiation, invasiveness, or angiogenesis. Regardless of the controversial perspective over these NF-κB related genes, our results from discovery and validation cohorts both indicated the potential effect on tumorigenesis from UC in the remission stage.

Nevertheless, there are also some limitations in our study. In terms of the patients in our discovery cohort, the detailed clinical features including the patients’ demographics, the treatment of these patients, and the grade of CAC were not assessable. Therefore, we cannot exclude the confounding factors and make a further investigation on the process of tumor transformation from quiescent UC to CAC. Moreover, unlike the discovery cohort, patients in our validation cohort were collected from 3 different institutions, which could lead to bias and variation due to the batch effect inevitably. Considering the heterogeneity of intestinal mucosa derived from these different institutions, further researches based on the comparison between quiescent UC patients harboring or not CAC after receiving the unified regimen will be more instrumental to delineate the tumorigenesis function of the seven-gene signature in NF-κB pathway.

**Conclusion**

In summary, we have identified a 7-NF-κB-related gene signature (CCL4, CXCL1, MYC, JUNB, FOS, ATF3, and CEBPB) with high expression in CAC, which may facilitate the tumorigenesis by regulating the immunological and inflammatory reactions via the NF-κB pathway. High levels of this signature may act as a powerful predictor for
CAC among quiescent UC patients during their scheduled surveillance.

**Abbreviations**

UC, Ulcerative colitis; CRC, colorectal cancer; CAC, ulcerative colitis-associated colorectal cancer; TNF, tumor necrosis factor; NF-kB, nuclear factor-kappa B; GSEA, gene set enrichment analysis; ssGSEA, single-sample gene set enrichment analysis; GSVA, gene set variation analysis; DC, dendritic cell; NK, natural kill cell; NKT, natural kill T cell; TAM, tumor-associated macrophage; MDSC, myeloid-derived suppressor cells; pDC, plasmacytoid dendritic cell; PPI, protein–protein interaction; WGCNA, weighted correlation network analysis; NES, normalized enrichment score; GEO, Gene Expression Omnibus; IBD, inflammatory bowel disease; GLM, generalized linear model; ROC, receiver operating characteristic curve; AUC, area under curve; CI, confidence intervals.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


