Original Article

Landscape of cell heterogeneity and evolutionary trajectory in ulcerative colitis-associated colon cancer revealed by single-cell RNA sequencing

Quan Wang[1,2*], Zhu Wang[3*], Zhen Zhang[1,2], Wei Zhang[1,2], Mengmeng Zhang[1,2], Zhanlong Shen[1,2], Yingjiang Ye[1,2], Kewei Jiang[1,2], Shan Wang[1,2]

[1]Department of Gastroenterological Surgery, Peking University People's Hospital, Beijing 100044, China; [2]Laboratory of Surgical Oncology, Beijing Key Laboratory of Colorectal Cancer Diagnosis and Treatment Research, Peking University People's Hospital, Beijing 100044, China; [3]Department of Gastrointestinal Surgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan 250021, China

*These authors contributed equally to this work.

Correspondence to: Shan Wang. Department of Gastroenterological Surgery, Peking University People's Hospital, Beijing 100044, China. Email: shanwang60@sina.com; Kewei Jiang. Department of Gastroenterological Surgery, Peking University People's Hospital, Beijing 100044, China. Email: jiangkewei@pkuph.edu.cn.

Abstract

Objective: The goal of this study was to get preliminary insight on the intra-tumor heterogeneity in colitis-associated cancer (CAC) and to reveal a potential evolutionary trajectory from ulcerative colitis (UC) to CAC at the single-cell level.

Methods: Fresh samples of tumor tissues and adjacent UC tissues from a CAC patient with pT3N1M0 stage cancer were examined by single-cell RNA sequencing (scRNA-seq). Data from The Cancer Genome Atlas (TCGA) and The Human Protein Atlas were used to confirm the different expression levels in normal and tumor tissues and to determine their relationships with patient prognosis.

Results: Ultimately, 4,777 single-cell transcriptomes (1,220 genes per cell) were examined, of which 2,250 (47%) and 2,527 (53%) originated from tumor and adjacent UC tissues, respectively. We defined the composition of cancer-associated stromal cells and identified six cell clusters, including myeloid, T and B cells, fibroblasts, endothelial and epithelial cells. Notable pathways and transcription factors involved in these cell clusters were analyzed and described. Moreover, the precise cellular composition and developmental trajectory from UC to UC-associated colon cancer were graphed, and it was predicted that *CD74*, *CLCA1*, and *DPEP1* played a potential role in disease progression.

Conclusions: scRNA-seq technology revealed intra-tumor cell heterogeneity in UC-associated colon cancer, and might provide a promising direction to identify novel potential therapeutic targets in the evolution from UC to CAC.

Keywords: Ulcerative colitis-associated colon cancer; single-cell RNA sequencing; cell heterogeneity; evolutionary trajectory

Submitted Nov 18, 2020. Accepted for publication Mar 02, 2021.
doi: 10.21147/j.issn.1000-9604.2021.02.13
View this article at: https://doi.org/10.21147/j.issn.1000-9604.2021.02.13

Introduction

Colorectal cancer (CRC) is the third most common type of cancer worldwide, and approximately 147,950 new cases and 53,200 deaths were estimated in 2020 in the United States. CRC has been the second most common cause of cancer death, even including 17,930 cases and 3,640 deaths in patients under 50 years of age (1). In China, the incidence and mortality of CRC have been on the rise (2) and CRC incurs a heavy economic burden on the society and individuals (3). Moreover, patients with colitis-associated cancer (CAC), a particular type of CRC that develops from inflammatory bowel diseases (IBDs), have an earlier morbidity and a poorer prognosis (4). CAC is often thought to arise from flat dysplasia with indistinct margins, in a field of concomitant inflammation, scarring, and pseudopolyposis, rather than development from a polypoid adenoma, which is the major cause of sporadic CRC (5). Furthermore, at the molecular level, the sequence of events leading to CAC is distinct from that of sporadic CRC. A distinct set of genes in sporadic CRC, including *TP53* (6), *APC* (7), and *KRAS* (8) contains more mutations than genes in CACs. However, the molecular process underlying colorectal carcinogenesis in IBDs is still poorly understood. Ulcerative colitis (UC), the most common form of IBD, has become increasingly prevalent worldwide (9). In a previously performed meta-analysis, quantitative estimates of CAC risk in UC patients have been reported to be 2% after 10 years, 8% after 20 years, and 18% after 30 years of disease (10), thereby indicating the importance of intensive studies.

In the past, research on tumor origin only targeted the genetic and epigenetic changes of tumor cells. However, over the last 20 years, the tumor micro-environment (TME) has been shown to play an equally important role in cancer development. Intra-tumoral heterogeneity among malignant and non-malignant cells, and their interactions within the TME are critical to tumor initiation, progression, metastasis, and many other diverse aspects of tumor biology (11). Accurate TME information not only helps to gain a better understanding of the tumor origin and development, but also contributes to the development of novel therapeutic targets.

In previous studies, genomic and transcriptomic studies have revealed driver mutations, aberrant regulatory programs, and disease subtypes for major human tumors (12). However, these studies relied on profiling technologies that measure tumors in bulk, and resulted in data that represent an "average" of all cells present, thereby limiting their ability to capture intra-tumoral heterogeneity. Single-cell sequencing provides an avenue to explore genetic and functional heterogeneity at cellular resolution (13). Single-cell RNA sequencing (scRNA-seq) combined with computational methods for functional clustering of cell types provides a less biased approach to the understanding of cellular heterogeneity. ScRNA-seq has been used in many studies involving human tumors (14), circulating tumor cells (15), and patient-derived xenografts (16), and has exhibited its unique predominance in studies of tumor composition, genomic evolution, cancer stem cells, tumor metastasis, and drug resistance. Here, we used scRNA-seq to generate phylogenetic trees and determined the evolutionary process of UC-associated colon cancer. To our knowledge, this is the first study depicting the cellular landscape of TME in UC-associated colon cancer at the single-cell transcriptome level.

Materials and methods

Human specimen collection

Fresh tumor tissues and non-malignant tissues (adjacent UC tissues) were taken from a 43-year-old female patient with UC-related colon adenocarcinoma who had a history of UC for eight years. The postoperative pathology was classified as pT3N1M0 (IIIB stage), which was defined as median differentiation ulcerative adenocarcinoma and microsatellite stability. Written informed consent was provided by the patient. This study was approved by the Research and Ethical Committee of Peking University People's Hospital and complied with all relevant ethical regulations. Following surgical resection, a tumor tissue sample and a non-malignant colon tissue sample, which was at least 5 cm away from the neoplastic foci were obtained (*Supplementary Figure S1*).

Protocols of scRNA-seq and data quality control

Protocols for the preparation of single-cell suspensions, droplet-based scRNA-seq, and other methods related to single-cell analysis, are described in *Supplementary Materials*. Single cells were filtered for further analysis based upon the criteria, cells that had either fewer than 201 unique molecular identifiers (UMIs), over 6,000 or under 101 expressed genes, or over 10% UMIs derived from the mitochondrial genome were excluded. Gene expression (in UMI) was normalized for scale and transformed in log2 (UMI+1).

Principle component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE)

PCA was used to summarize the resulting variably expressed genes to reduce the dimensionality of this data set. tSNE was applied to recalculate the sample distance by the conditional probability of random neighbor fitting based on Student's *t*-distribution in high dimensional space, which was further conducted for the above principle components dimensionality reduction via the default settings of the Run tSNE function, in order that sample presents a clearly separated cluster in a low dimensional space.

Pathway and functional annotation analysis

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource for understanding high-level functions and effects of the biological system (http://www.genome.jp/kegg/), and performed via DAVID (https://david.ncifcrf.gov/). Enriched pathways with a Q value ≤0.05 were considered significantly different, like the functional annotation through the Gene Ontology database, for which the Fisher's exact test was used to select only significant categories including biological process, cellular components, and molecular function classifications. Q value ≤0.05 was considered significantly different.

Gene prognostic performance in The Cancer Genome Atlas (TCGA) data

To evaluate the role of cell type in a larger compendium of tumors, we assessed their expression in bulk RNA-seq data from the TCGA (http://www.cbioportal.org/). Specifically, we downloaded preprocessed gene expression data as well as clinical data for primary solid tumors and normal solid tissues for colon and rectal adenocarcinoma using the Bioconductor TCGA biolinks package (http://www.bioconductor.org/). A total of 598 CRC samples were included and divided into high- and low-expression groups based on median gene expression level. Z scores from the TCGA and the validation cohort were combined using the weighted Z method. For Kaplan-Meier analysis, marker gene expression categorization was optimized.

Immunohistochemical (IHC) staining in The Human Protein Atlas

To confirm the expression change in these genes along with disease progression, IHC staining was used for assessment of the different expression level of a specific gene between normal and tumor tissues from The Human Protein Atlas (https://www.proteinatlas.org/).

Results

scRNA-seq and cell typing of non-malignant and CAC tissues

Fresh tumor and non-malignant tissues were taken from a 43-year-old Chinese female patient with colon adenocarcinoma classified as pT3N1M0, and this patient had been diagnosed with UC for 8 years. Once non-malignant and CAC tissues were obtained (*Supplementary Figure S1*), it was rapidly digested to a single-cell suspension and analyzed using scRNA-seq involving a single-tube protocol with unique transcript counting through barcoding with UMIs (*Figure 1A*). To obtain detailed cellular genetic information on this tumor, over 1.6 billion post-normalization reads were performed for subsequent analysis, which were obtained from 4,777 cells; a median of 1,220 genes per cell were expressed. Of the sample cells, 2,250 (47%) originated from tumor tissues and 2,527 (53%) originated from non-malignant tissues (*Figure 1B*). Following gene expression normalization for read depth and mitochondrial read count, PCA was applied to genes that were variably expressed across all 4,777 cells (n=1,220 genes). Subsequently, cells were classified by cell type using graph-based clustering on the informative principle components (n=12). This approach identified cell clusters that, through marker genes, could be readily assigned to known cell lineages. In addition to cancer cells, myeloid cells, T cells, B cells, fibroblasts, endothelial cells and epithelial cells were identified (*Figure 1B*,*C*, *Supplementary Table S1*). The transcript analysis showed that these cells differed considerably in transcriptional activity, either between different cell types or between regions of the same type. This approach also distinguished the sample origin and numbers between diverse subgroups (*Figure 1D*).

Different angiogenesis pathways in tumor and non-malignant endothelial cells

A total of 228 endothelial cells were detected and four clusters were revealed (*Figure 2A*). We next aimed to identify marker genes for each of these clusters and assign them to known endothelial cell types (*Figure 2B*, *Supplementary Table S1*). This revealed three sets of vascular endothelial cells: two were mostly tumor-derived (clusters 1 and 3; *ACKR1*+ and *CA4*+, respectively) and one was mostly non-malignant tissue-derived (cluster 2; *CYR61*+). Another set of 25 lymphatic endothelial cells was

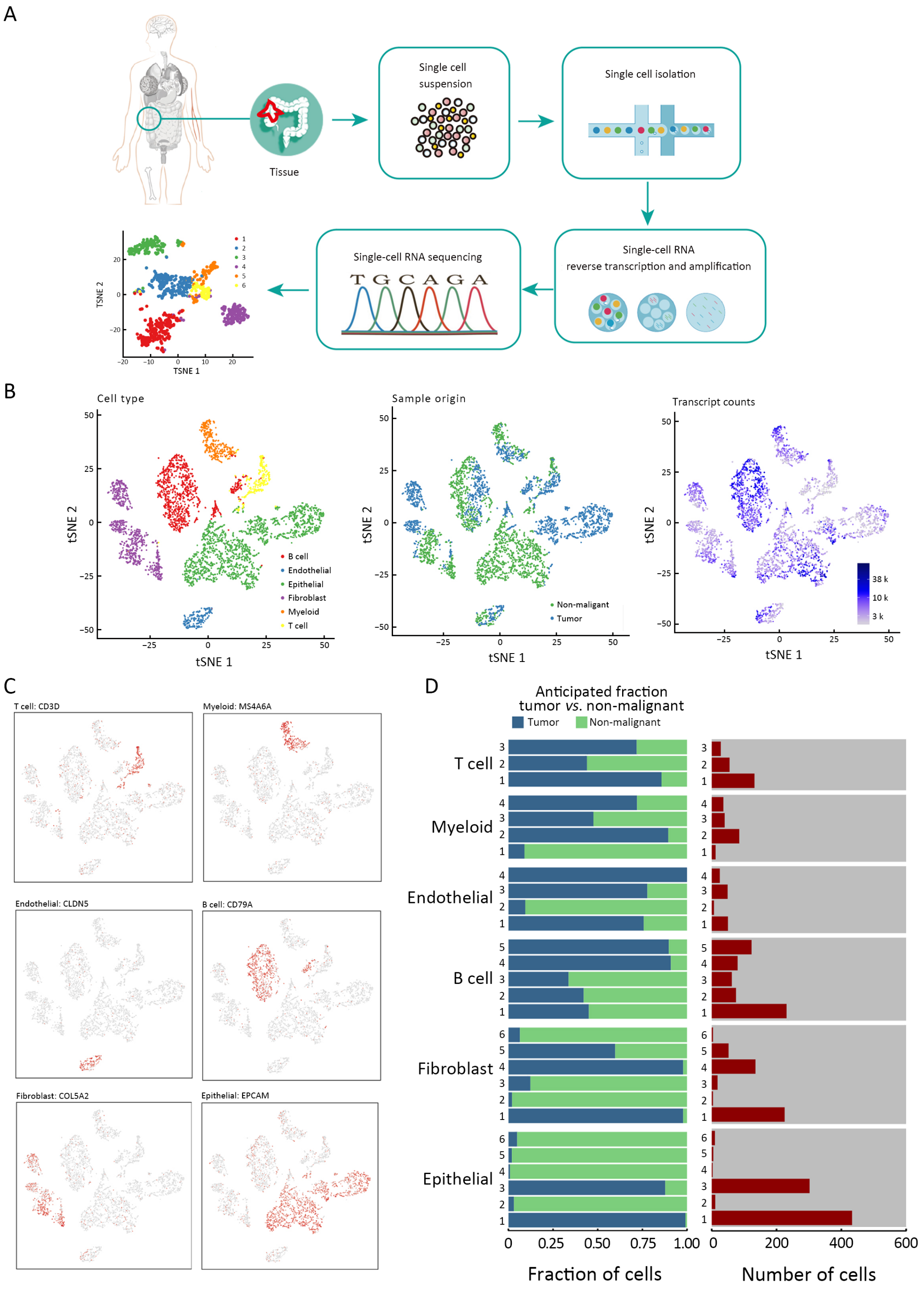

Figure 1 scRNA-seq and cell typing of non-malignant and CAC tissues. (A) Overview of the study design; (B) tSNE of the 4,777 cells profiled here, color-coded by (left to right) cell type, sample origin (tumor or non-malignant tissue) and transcripts counts detected in that cell (log scale as defined in the inset); (C) Expression of marker genes for cell types defined above each panel. In addition to cancer cells, we identified myeloid cells, T cells, B cells, fibroblasts, endothelial cells and epithelial cells; (D) For each of the cell subclusters (left to right): fractions of original cells, and number of cells. scRNA-seq, single-cell RNA sequencing; CAC, colitis-associated cancer; tSNE, t-distributed stochastic neighbor embedding; UMI, unique molecular identifier.

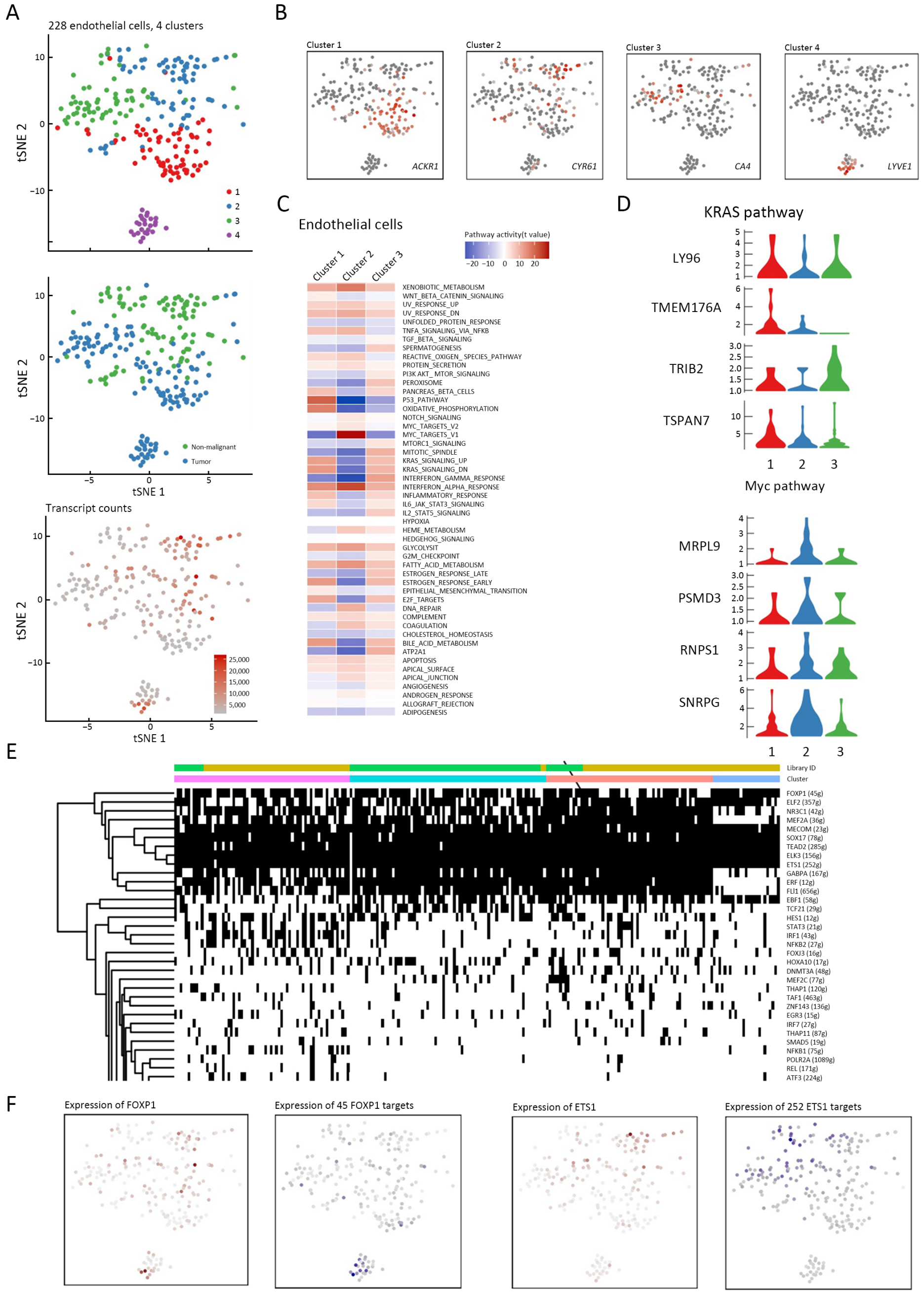

Figure 2 Different angiogenesis pathways in tumor and non-malignant endothelial cells. (A) tSNE plot of 228 endothelial cells (top to bottom), color-coded by their associated cluster, the sample type of origin, and the number of transcripts detected in each cell; (B) tSNE plot color-coded for expression (gray to red) of marker genes, (cluster 1, *ACKR1*+; cluster 2, *CYR61*+; cluster 3, *CA4*+ and cluster 4, *LYVE1*+); (C) Differences in pathway activities scored per cell by GSVA among the vascular clusters. The KRAS signaling pathway was significantly down-regulated in cluster 2, but up-regulated in clusters 1 and 3, while the Myc target pathway showed contrasting results; (D) Violin plots show the expression distribution of selected genes involved in the KRAS and Myc pathways; (E) SCENIC analysis of the involved transcription factors involved among the clusters. Many transcription factors took part in attending angiogenesis, including *FOXP1* and *ETS1*; (F) Exhibition of the involved pathways (KRAS, Myc) and transcription factors (*FOXP1*, *ETS1*) and their target genes, corresponding to the degree of expression. tSNE, t-distributed stochastic neighbor embedding; GSVA, gene set variation analysis; SCENIC, single-cell regulatory network inference and clustering.

found in the tumor sample (cluster 4; marker genes *LYVE1*+ and *CCL21*+), however, interestingly, no distinct cluster of lymphatic endothelial cells from the non-malignant sample was identified.

Except for lymphocyte-composed cluster 4, we used hallmark pathway gene signature analysis to identify different characteristics among the other three clusters, which were all from vascular endothelial cells (*Figure 2C*). Interestingly, the KRAS signaling pathway was significantly down-regulated in cluster 2, but up-regulated in clusters 1 and 3, whereas the Myc target pathway showed contrasting results (*Figure 2D*). Detailed analysis has shown that the two pathways were angiogenesis-related, however, different tissue-derived clusters seemed to have diverse mechanisms. The role of *KRAS* oncogenes in promoting cellular transformation is well described, and KRAS modulates tumor-stroma interactions and supports cancer invasiveness by influencing the expression of metalloproteinases and cytokines involved in angiogenesis (17). In our study, the high enrichment of KRAS pathway observed in tumor tissues was in accordance with the hyper-vascular nature of tumors.

Next, single-cell regulatory network inference and clustering (SCENIC) analysis (18) was applied, in which differentially expressed genes were scanned for over-expressed transcription factor binding sites, the co-expression of transcription factors and their putative target genes were analyzed (*Figure 2E*,*F*). Many transcription factors play a role in angiogenesis, including *FOXP1* and *ETS1*. Of note, *FOXP1* can stimulate angiogenesis by repressing semaphorin 5B in endothelial cells, and regulate angiogenesis through the circ-SHKBP1/miR-544a/FOXP1 pathway (19). Furthermore, *ETS1* enables angiogenesis in several ways (20).

Cancer associated fibroblasts (CAFs) play various roles in tumorigenesis

Fibroblasts have long been suggested to be a heterogeneous population, but the extent of heterogeneity has yet to be explored. Fibroblast phenotypes are considered highly context-dependent and unstable in culture. In our samples, 857 fibroblasts were detected (*Figure 3A*). Sub-clustering revealed six distinct subtypes: clusters 1 and 4 (cluster 1, *PCOLCE2*+ and cluster 4, *CXCL6*+) were strongly enriched in tumor tissues, while cluster 2 (*PLAT*+) was fully enriched in non-malignant tissues. In addition, other clusters (cluster 3, *ARHGDIB*+; cluster 5, *MYH11*+; cluster 6, *STMN2*+), however, were derived from a mixture of tumor and non-malignant tissues, but were mostly enriched in non-malignant tissues (*Figure 3B*). Remarkably, fibroblasts (*CD34*+ and *KLF4*+) were generally enriched in tumor tissues. The marker genes of CAFs, *PCOLCE2* and *CXCL6* were significantly up-regulated in tumor tissues by the bulk RNA-seq data from TCGA data (*Figure 3C*).

The prime role of CAFs is promoting the proliferation of cancer cells. Clusters 1 and 4 were enriched in Wnt and KRAS signaling, and are closely related with tumor proliferation (21) (*Figure 3D*,*E*). SCENIC analysis showed that *KLF12*, which promotes CRC growth, was highly expressed in cluster 1 (22) (*Figure 3G*, *Supplementary Figure S2*). Furthermore, cluster 4 showed high expression in TGFβ and Wnt signaling, which are also related to cancer invasion and metastasis (23). Another remarkable characteristic of CAFs is extracellular matrix (ECM) remodeling, and collagens, important ECM components, participate in tumor progression (24). We found that various collagens were highly expressed in fibroblasts, and different clusters seemed to have different expression inclinations (*Figure 3F*). In addition to collagens, many other ECM components, such as fibronectin, periostin, hyaluronan, and proteoglycans (marker genes *PRG4*, *POSTN*, *HAS2* and *FN1*, respectively) were also up-regulated by CAFs (25,26). The corresponding genes were highly expressed in CAFs (*Figure 3G*). CAFs also influenced the drug resistance of tumors, and CXCR4 expression predicted patient outcome and recurrence patterns after hepatic resection for colorectal cancer with liver metastases. *CXCL12* can mediate drug resistance by combining with CXCR4 that is expressed in cancer cells (27). In addition, *FAP*+ CAF can mediate immune suppression by *CXCL12* (28).

CRC-related pathways are enriched in tumor-derived B cells

In our study, we detected 1,100 B lymphocyte cells. B cells are the most prevalent type of stromal cells (*Figure 4A*). Clustering revealed five clusters, including two clusters (clusters 4 and 5; *REG3A*+ and *MS4A1*+, respectively) that were mostly tumor enriched, whereas the other three clusters (cluster 1, *IGHM*+; cluster 2, *IGLL5*+ and cluster 3, *IGHGP*+, respectively) were composed of both tumor and non-malignant cells (*Figure 4B*). Moreover, clusters 1, 2 and 3 showed plasma properties and were not grouped into distinct clusters. Although all cells expressed immunoglobulin A, tumor-derived cells showed higher

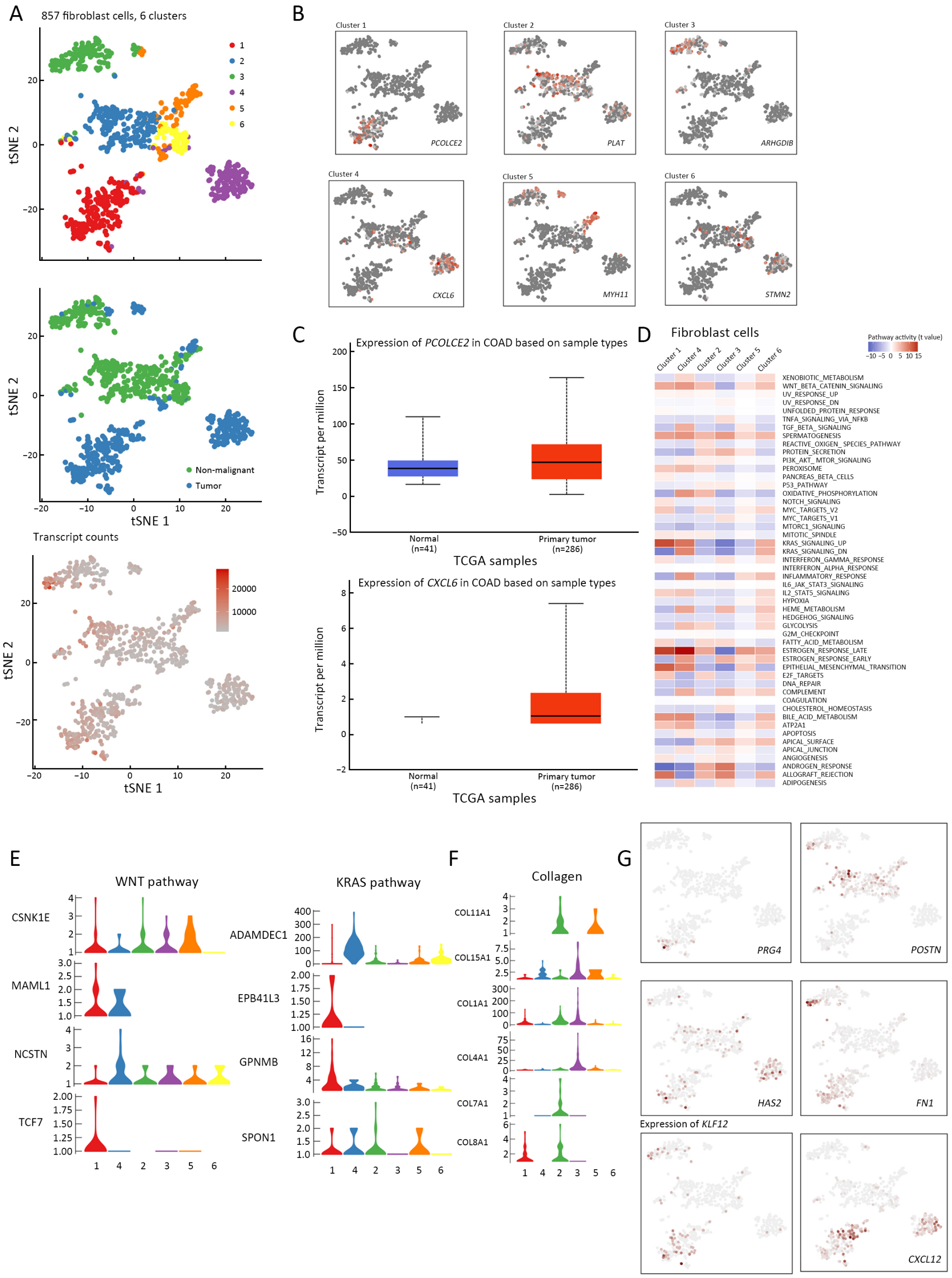

Figure 3 CAF plays various roles in tumorigenesis. (A) tSNE plot of 857 fibroblast cells (top to bottom), color-coded by their associated cluster, the sample type of origin, and the number of transcripts detected in each cell; (B) tSNE plot color-coded for expression (gray to red) of marker genes (cluster 1, *PCOLCE2*+; cluster 2, *PLAT*+; cluster 3, *ARHGDIB*+; cluster 4, *CXCL6*+; cluster 5, *MYH11*+; cluster 6; *STMN2*+); (C) CAFs marker genes, *PCOLCE2* and *CXCL6* were confirmed to be significantly up-regulated in tumor tissues, as confirmed in bulk RNA-seq data from TCGA; (D) Differences in pathway activities scored per cell by GSVA among the clusters. Clusters 1 and 4 were enriched in Wnt and KRAS signaling, which have a close relationship with tumor proliferation; (E) Violin plots show the expression distribution of selected genes involved in Wnt and KRAS pathways; (F) Different fibroblast clusters expressed different kinds of collagens; (G) The involved marker genes (*FN1*, *HAS2*, *CXCL12*, *POSTN*, *PRG4*), and transcription factor *KLF12* and its target genes, corresponding to the degree of expression. tSNE, t-distributed stochastic neighbor embedding; CAF, cancer associated fibroblast; TCGA, The Cancer Genome Atlas; GSVA, gene set variation analysis; COAD, colon adenocarcinoma.

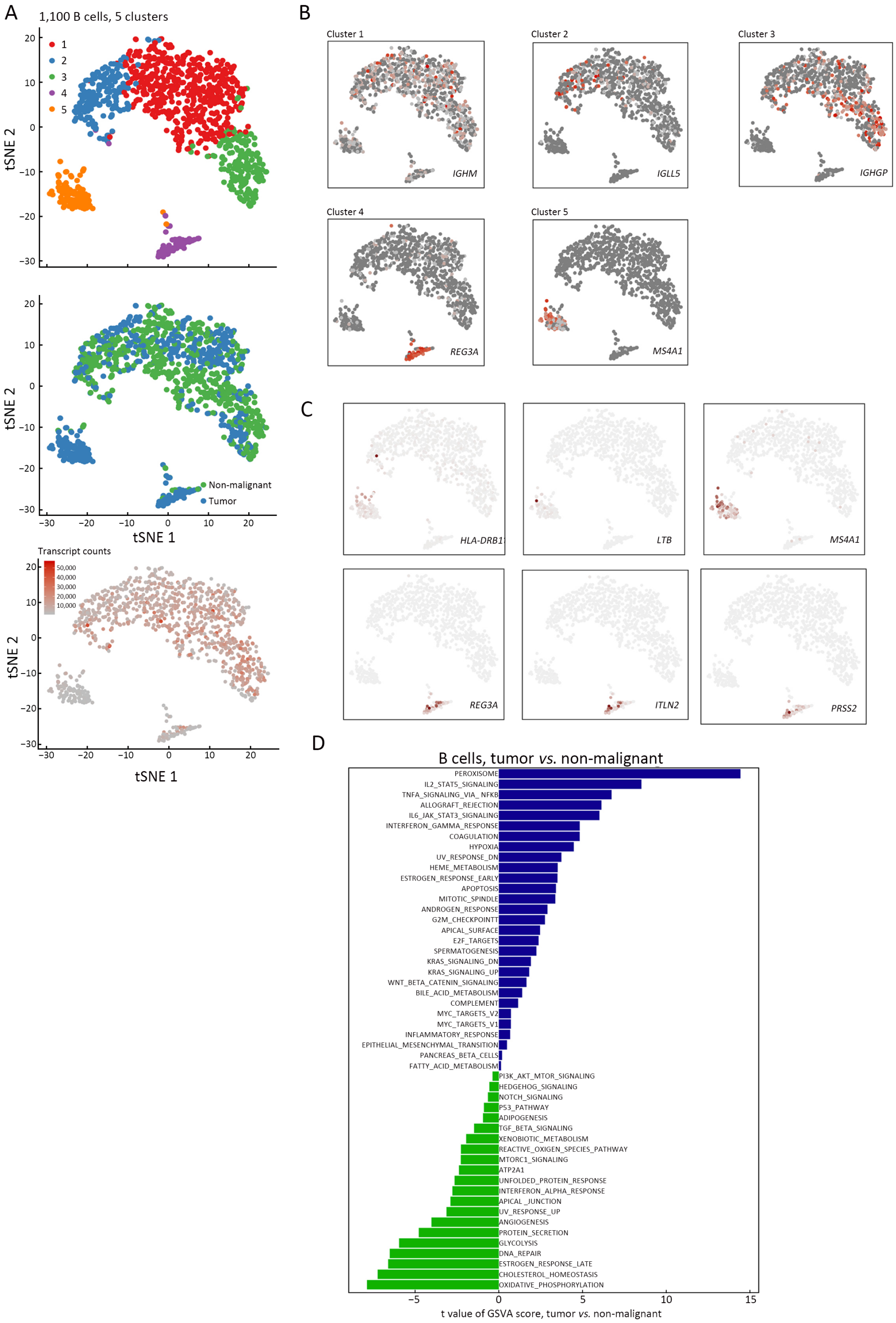

Figure 4 CRC-related pathways are enriched in tumor-derived B cells. (A) tSNE plot of 1,100 B cells (top to bottom), color-coded by their associated cluster, the sample type of origin, and the number of transcripts detected in each cell; (B) tSNE plot color-coded for expression (gray to red) of marker genes (cluster 1, *IGHM*+; cluster 2, *IGLL5*+; cluster 3, *IGHGP*+; cluster 4, *REG3A*+ and cluster 5, *MS4A1*+); (C) Exhibition of the marker genes (*REG3A*, *PRSS2*, *ITLN2*, *HLA-DRB1*, *MS4A1*, *LTB*) corresponding to the degree of expression; (D) Differences in pathway activities scored per cell by GSVA among the clusters. Pathway analyses showed that peroxisome signaling was highly expressed in tumor cells. CRC, colorectal cancer; tSNE, t-distributed stochastic neighbor embedding; GSVA, gene set variation analysis.

levels of IgM expression, while non-malignant-derived cells showed a higher IgA expression. Cluster 4 significantly expressed many innate immunity-related genes, such as *REG3A*, *PRSS2*, *ITLN2* and *LYZ*, but had a negative CD5 expression. Therefore, these cells were identified as B1 cell-like cells. Cluster 5 showed a high expression of *MS4A1*, *LTB* and *HLA-DRB1*, which were characteristics of follicular B cells (*Figure 4C*).

Pathway analyses showed that peroxisome signaling was highly expressed in tumor cells (*Figure 4D*). Moreover, peroxisome signaling was closely related to CRC risk (29) and was positively correlated with lymph node metastasis and poor prognosis of CRC (30). The results also showed that the oxidative phosphorylation pathway was enriched in non-malignant-derived cells, which might be attributed to the impact of colitis. SCENIC analysis failed to identify differences between non-malignant tissue-derived and tumor-derived plasma cells (*Supplementary Figure S3*). In addition, no differences were observed in transcript numbers in tumor-associated *vs.* non-malignant tissue-associated plasma cells, however, the plasma cells showed a higher transcription trend than clusters 4 and 5 (*Figure 4A*).

Different derived myeloid cells show diverse expressing properties

The 362 myeloid cells clustered into four subsets, which were not completely separated (*Figure 5A*). One cluster corresponded to macrophages (cluster 1, *CAPG*+), another cluster to monocytes (cluster 2, *CXCL2*+). There was also a dendritic cell cluster (cluster 3, *IDO1*+) and a granulocyte cluster (cluster 4, *CCL20*+) (*Figure 5B*). Dendritic cells and granulocytes were typically more abundant in tumors than in non-malignant tissues. In contrast, macrophages were more abundant in non-malignant tissues. In both tissues, monocytes were detected at similar numbers.

The cell numbers of macrophages and monocytes displayed extensive heterogeneity in tumors compared with those in non-malignant tissues, however, SCENIC analysis did not show significance among different derived cells (*Supplementary Figure S4*). Furthermore, the pathways of the differently derived cells were analyzed and revealed a tumor-associated increase in tumorigenesis, cell proliferation, and low-oxygen metabolism pathways (that is, pathways associated with tumor necrosis factor α signaling, KRAS signaling and hypoxia). Non-malignant tissue-derived preferred oxidative phosphorylation and biomass production pathways (pathways associated with phosphorylation and protein secretion) (*Figure 5C*).

Several typical species of T cells identified in multiple analysis

With 318 cells detected, T cells were mainly divided into three clusters, and were designated as naive T cells (cluster 1, *YPEL5*+), cytotoxic T cells (cluster 2, *GZMA*+), and natural killer T cells (cluster 3, *PIGR*+) (*Figure 6A*,*B*). In cluster 1, the significantly expressed genes, including *YPEL5* and *GPR18*, were closely related to proliferation and cell differentiation (31). Pathway analysis showed that many proliferation- and differentiation-related pathways, such as Myc targets, G2M checkpoints and E2F targets, were highly expressed in cluster 1 (*Figure 6C*). SCENIC analysis also showed that T cell-specific differentiation-related transform factor ELF-1 was significantly up-regulated (*Figure 6D*,*E*). In cluster 2, we found that cytotoxic T cell-specific genes, such as *GZMA* and *GNLY*, were highly expressed. Additionally, the glycolysis pathway was most highly expressed in cluster 2 among the three clusters, and a related gene, *PKM*, was also highly expressed (32). Furthermore, in cluster 2, a small number of populations of cells were detected expressing higher levels of the immune checkpoint molecule HAVCR, which acts in the tolerance and exhaustion of T cells (33) and is currently targeted in clinical trials of immunotherapy for cancers (34). At the meantime, this subtype of cells showed high expression of MKI67, which encoded proliferation-related proteins (35), and the notable transcription numbers also reflected their high proliferative activity (*Figure 6A*,*B*).

Heterogeneity of epithelial cells was demonstrated

In total, 1,912 epithelial cells were characterized and divided into six clusters; two tumor-derived clusters (cluster 1, *ENPEP*+; cluster 3, *OLFM4*+, respectively), and four clusters were almost exclusively from non-malignant tissues (cluster 2, *PI3*+; cluster 4, *MUC1*+; cluster 5 *CA4*+; cluster 6, *HMGB2*+, respectively) (*Figure 7A*,*B*). Pathway analysis showed significant differences between tumor- and non-malignant-derived tissues (*Figure 7C*). In particularly, cluster 1 typically exhibited malignant properties. High proliferation- and embryonic developmental process-related pathways were highly expressed (Wnt signaling, Myc targets and EMT signaling) (36), and lesions repair-associated pathways, including the DNA repair pathway, were down-regulated (37) (*Figure 7D*).

Through SCENIC analysis, we demonstrated that the

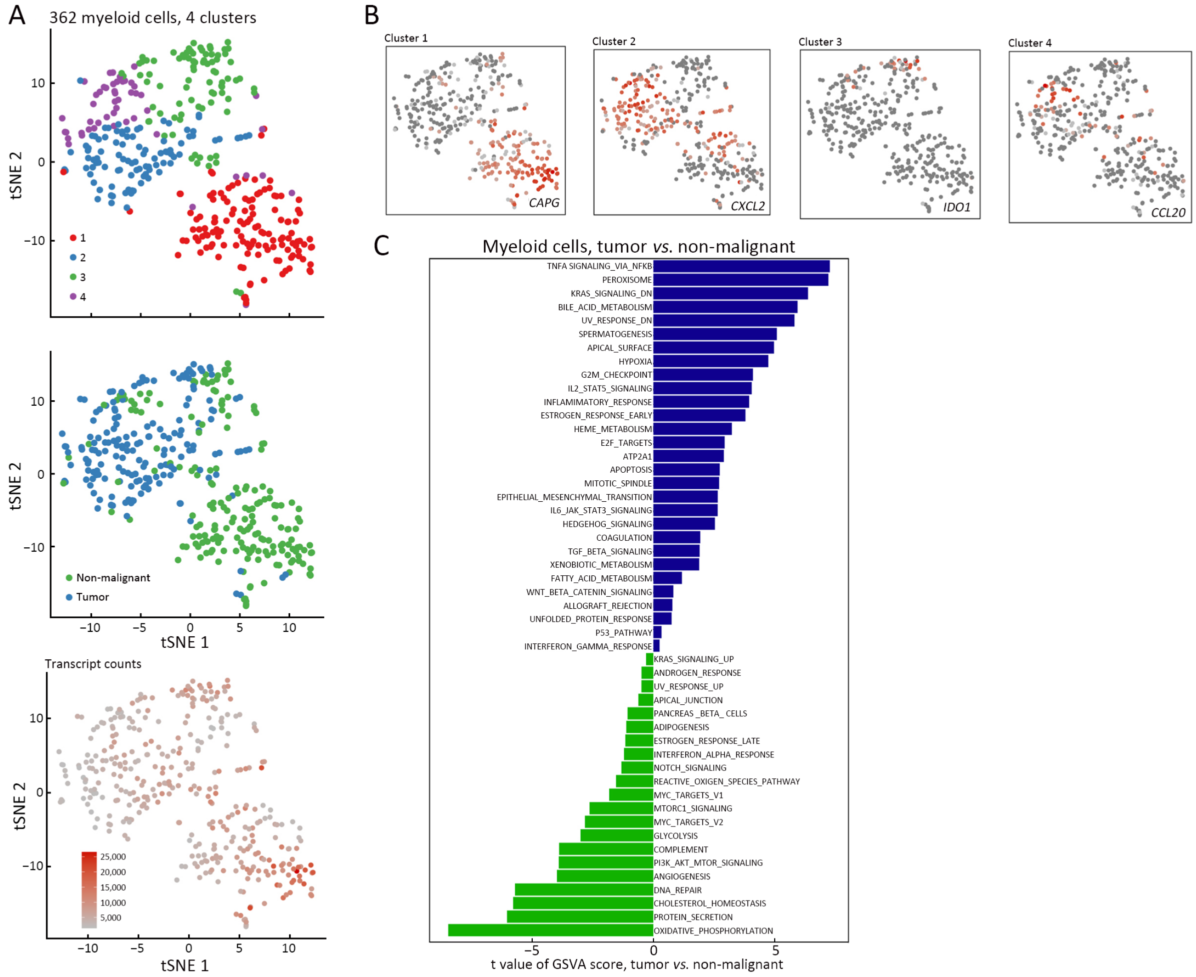

Figure 5 Different derived myeloid cells showed diverse expression properties. (A) tSNE plot of 362 myeloid cells (top to bottom), color-coded by their associated cluster, the sample type of origin, and the number of transcripts detected in each cell; (B) tSNE plot color-coded for expression (gray to red) of marker genes (cluster 1, *CAPG*+; cluster 2, *CXCL2*+; cluster 3, *IDO1*+ and cluster 4, *CCL20*+); (C) Differences in pathway activities scored per cell by GSVA among the clusters. Tumor-associated cells increased in tumorigenesis, cell proliferation and low-oxygen metabolism pathway (that is, pathways associated with TNFα signaling, KRAS signaling and hypoxia), while the non-malignant tissue-derived cells preferred the oxidative phosphorylation and biomass production pathways (pathways associated with phosphorylation and protein secretion). tSNE, t-distributed stochastic neighbor embedding; GSVA, gene set variation analysis; TNF, tumor necrosis factor.

transcription factors CDX2 and STAT3 were significantly up-regulated in cluster 1, and showed almost no expression in other clusters except for cluster 3 (*Supplementary Figure S5*). CDX2 inhibits aggressive phenotypes of colon cancer cells both *in vitro* and *in vivo* (38) and might be a prognostic marker related to the benefit of adjuvant chemotherapy (39). STAT3 is essential for the transduction of tumor-promoting signals of the IL-6/STAT3 pathway, which is highly activated in CAC (40). Collectively, different clusters exhibited obviously diverse properties even in the same tissues (*Figure 7E*).

Evolutionary trajectory of disease development and internal variation in crucial genes

Complete transcriptome data for many epithelial cells allowed us to gain insights into the functional states of and relationship among these cells. Carcinogenesis follows the principles of Darwinian evolution, whereby somatic cells acquire genomic alterations that provide them with a survival and/or growth advantage (41). Therefore, the dynamic information of gene expressions could be used to track the progress of the disease. Transcriptional similarity-

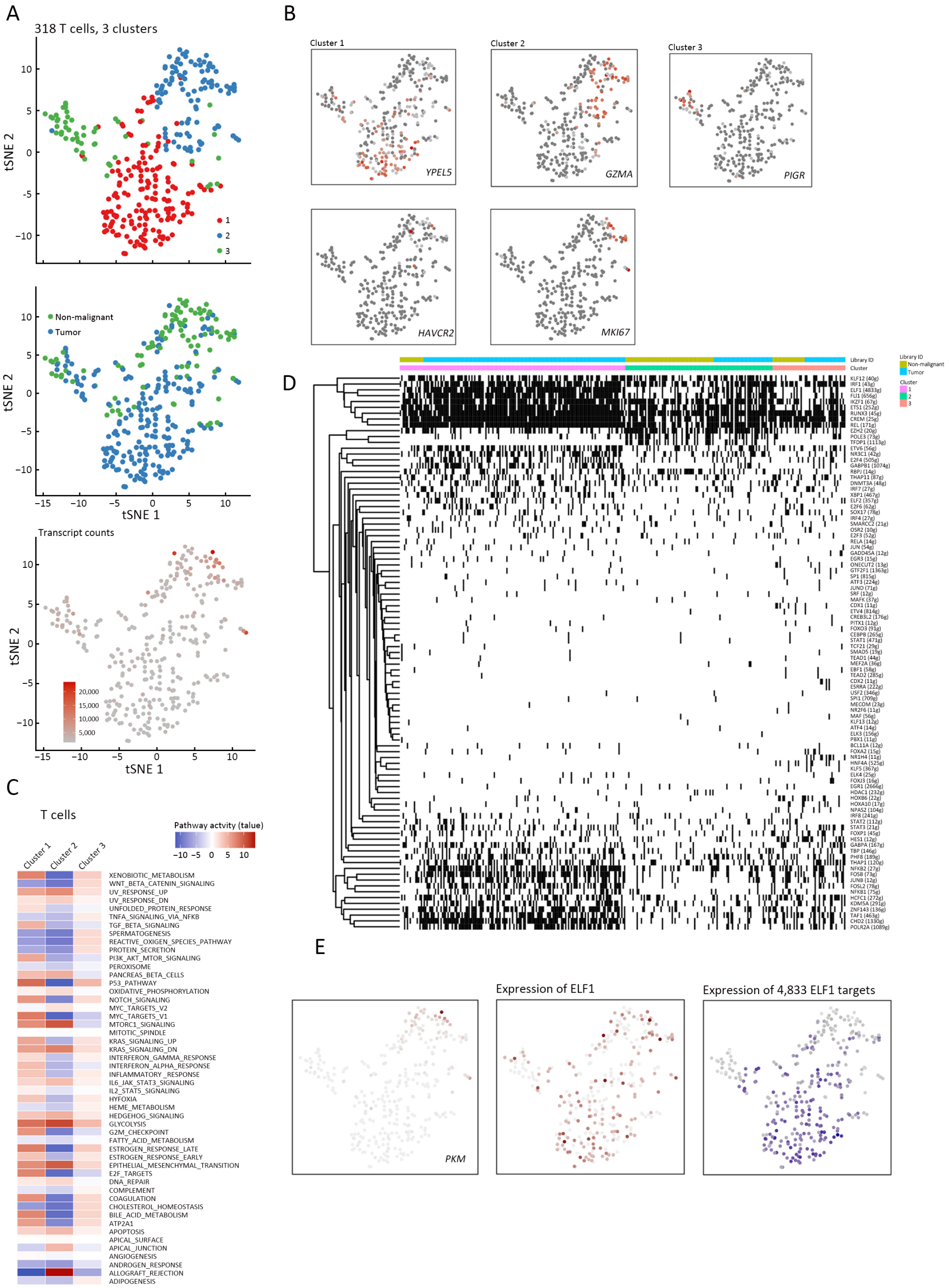

Figure 6 Several typical species of T cells are identified in multiple analysis. (A) tSNE plot of 318 T cells (top to bottom), color-coded by their associated cluster, the sample type of origin, and the number of transcripts detected in each cell; (B) tSNE plot color-coded for expression (gray to red) of marker genes, naive T cells (cluster 1, *YPEL5*+), cytotoxic T cells (cluster 2, *GZMA*+) and natural killer and natural killer T cells (cluster 3, *PIGR*+); (C) Differences in pathway activities scored per cell by GSVA among the clusters. Pathway analysis showed that many proliferation- and differentiation-related pathways, such as Myc targets, G2M checkpoints and E2F targets, were highly expressed in cluster 1; (D,E) SCENIC analysis of transcription factors involved among the clusters, and SCENIC analysis showed that the T cell-specific differentiation-related transform factor ELF-1 was significantly up-regulated. tSNE, t-distributed stochastic neighbor embedding.

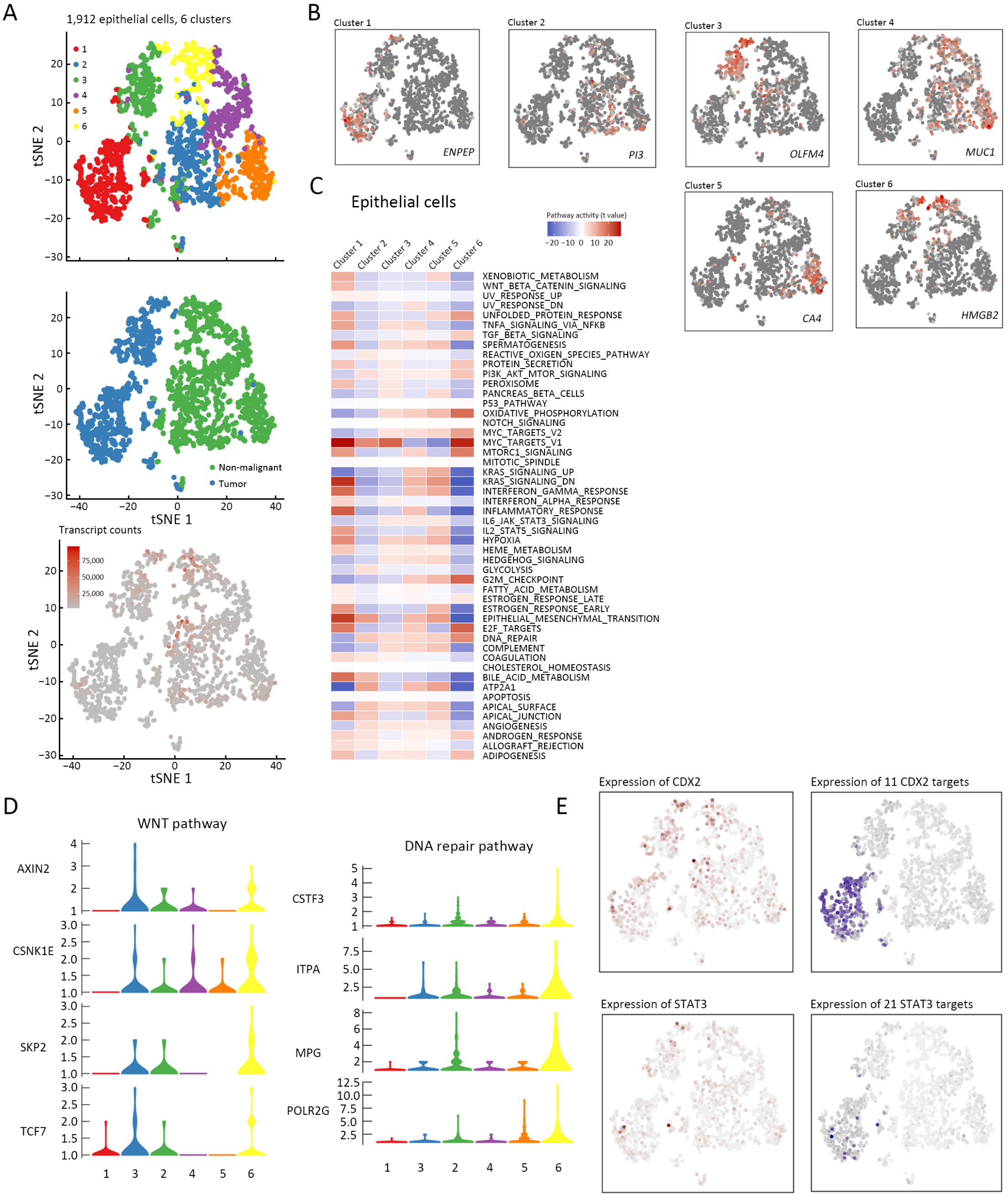

Figure 7 Heterogeneity of epithelial cells was demonstrated. (A) tSNE plot of 1,912 epithelial cells (top to bottom), color-coded by their associated cluster, the sample type of origin, a+nd the number of transcripts detected in each cell; (B) tSNE plot color-coded for expression (gray to red) of marker genes (cluster 1, *ENPEP*+; cluster 2, *PI3*+; cluster 3, *OLFM4*+; cluster 4, *MUC1*+; cluster 5, *CA4*+; cluster 6, *HMGB2*+); (C) Differences in pathway activities scored per cell by GSVA among the clusters. Pathway analysis showed significant differences between the tumor- and non-malignant-derived tissues. In particularly, cluster 1 and 3 existed typically exhibited malignant properties; (D) Violin plots show the expression distribution of selected genes involved in the Wnt and DNA repair pathways. In cluster 1 and 3, high proliferation- and embryonic developmental process-related pathways were highly expressed (Wnt signaling, Myc targets and EMT signaling), and lesions repair-associated pathways, such as the DNA repair pathway, were down-regulated; (E) Exhibition of the involved transcription factors (*CDX2*, *STAT3*) and their target genes, corresponding to the degree of expression. tSNE, t-distributed stochastic neighbor embedding; EMT, epithelial-mesenchymal transition.

based pseudotime analysis was applied (42) to order epithelial cells and indicated their developmental trajectories (*Figure 8A*). All cells from each cluster aggregated into nine states based on expression similarities, and different states formed a trajectory by pseudotime analysis that began with states 4, 7 and 8 (non-malignant-derived cells), followed by states 2, 3 and 5 (mixed derived cells), and ended with states 1, 6 and 9 (tumor-derived cells) (*Figure 8A*,*B*). Following this trajectory, the differentially expressed genes were identified, and these genes might play crucial roles in the evolution from colitis to cancer. Therefore, in the subsequent analysis, we focused on the top three differentially expressed genes, *CD74*, *CLCA1* and *DPEP1* (*Figure 8C*). Different degrees of IHC staining between normal and tumor tissues for CD74, CLCA1 and DPEP1 from the public data website of The Human Protein Atlas confirmed the change in expression in these genes along with disease progression (*Figure 8D*).

CD74, cluster of differentiation 74, also known as HLA-DR antigen-associated invariant chain and encoded by the *CD74* gene, is a polypeptide that is involved in the formation and transport of MHC class II protein, which is found in several types of cancer cells (43,44). In CRC, stimulation of CD74 by MIF induces a signaling cascade, leading to up-regulation of Bcl-2, thereby resulting in a significantly increased survival of patients with colon cancer. The MIF/CD74 axis is a target for novel therapies (45). Taken together, our data demonstrated that CD74 was down-regulated as tumorigenesis progressed, which was consistent with the findings presented in previous reports. TCGA data showed that CD74 expression was significantly high in normal tissues (P<0.05) (*Figure 8D*,*E*), and patients with high CD74 expression had a better survival (P<0.01) (*Figure 8F*).

CLCA1, calcium-activated chloride channel regulator 1, is a protein that is encoded by the *CLCA1* gene in humans and plays many roles, including the regulation of mucus production and secretion in goblet cells (46). *CLCA*1 regulates tissue inflammation in the innate immune response (47) and tumor suppression in CRC (48), and can suppress CRC aggressiveness via inhibition of the Wnt/β-catenin signaling pathway. Low expression of CLCA1 predicts a poor prognosis in CRC (49). Thus, our data demonstrated that CLCA1 was first up-regulated, then down-regulated during the late period of the pseudotime analysis. TCGA data indicated no significance for the expression of CLCA1 between tumor and normal tissues (P=0.07) (*Figure 8D*,*E*), and no relevance between CLCA1 expression and patient prognosis (P=0.11) (*Figure 8F*).

DPEP1, dipeptidase 1, encoded by the *DPEP1* gene, hydrolyzes a wide range of dipeptides and plays a role in many biological processes, including the metabolism of glutathione and β-lactam hydrolysis in the kidney (50). In CRC, the expression of DPEP1 has been shown to affect cancer cell invasiveness in early stage cases, and can act as a candidate tumor-specific molecular marker for the detection of rare disseminated colorectal tumor cells in peripheral venous blood and intraperitoneal saline lavage samples (51). Our data demonstrated that *DPEP1* was up-regulated as tumorigenesis progressed, which was consistent with the data presented in previous reports. TCGA data showed that the expression of DPEP1 was significantly higher in tumor tissues (P<0.05) (*Figure 8D*,*E*), however, no relevance was observed between DPEP1 expression and patient prognosis (P=0.09) (*Figure 8F*).

Discussion

In previous studies, cell identity was defined by various methods, such as morphological appearance, tissue context, and marker gene expression. As mRNA encodes cellular function and phenotype, single-cell transcriptomics could precisely refine the cellular identity based on comprehensive and quantitative readout of mRNA (52). Thus, scRNA-seq technology has attracted significant attention since its inception, and a large number of applied study results have been published recently (53). In the context of human cancer, scRNA-seq was used to reveal the intra-tumor heterogeneity and transcriptional trajectories of malignant transformation (54).

In this study, 4,777 single-cell transcriptomes of human colon tumorous tissues and non-malignant tissues from UC-associated colon cancer were analyzed. Furthermore, the composition of cancer-associated stromal cells was defined, the different subgroups of tumor cells were analyzed, and the notable pathways and transcription factors involved in the disease were described. Many of the tumor-derived cell types identified by a scRNA-seq approach, including B cells, T cells, endothelial cells, myeloid cells, fibroblasts, and epithelial cells (summarized in *Figure 9*), presented an altered gene expression profile with pro-tumoral properties compared with non-malignant cells. Of note, the evolutionary trajectory of tumor development was graphed, and pseudotime analysis revealed the cellular composition of CAC and its developmental trajectory. The TME might play a crucial

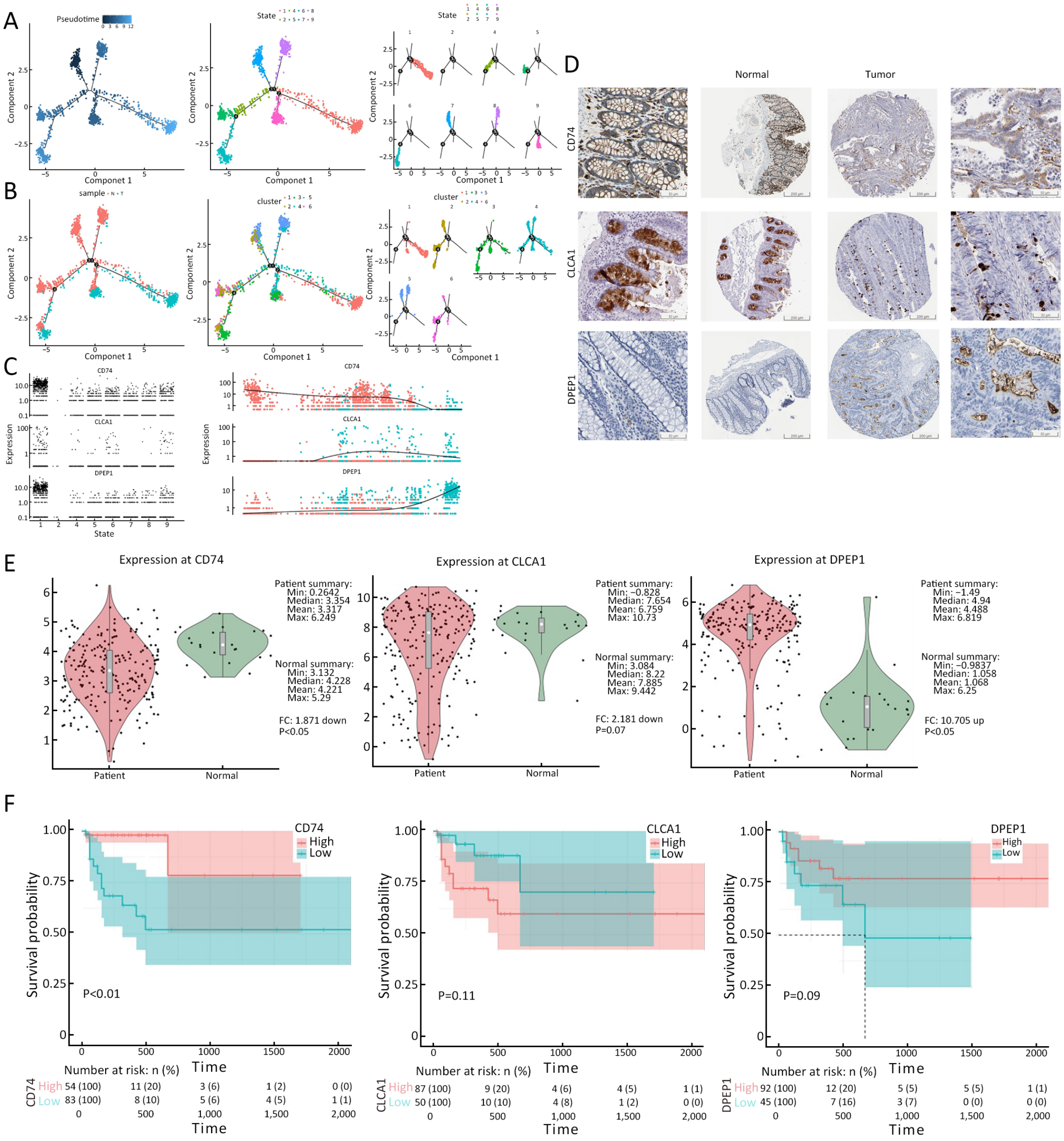

Figure 8 Evolutionary trajectory of disease development and internal variation in crucial genes. (A) Transcriptional similarity-based pseudotime analysis was used to order epithelial cells and indicated their developmental trajectories. All the cells from each cluster aggregated into nine states based on pseudotime analysis; (B) Distribution of the cells originally from the tumor and non-malignant tissues into six clusters (left to right); (C) Following this trajectory, the differentially expressed genes were identified, and these genes might play crucial roles in the evolution from colitis to cancer. Top 3 differential expressed genes (*CD74*, *CLCA1* and *DPEP1*) were showed along with evolutionary trajectory; (D) Different degrees of immunohistochemical staining between normal and tumor tissues for CD74, CLCA1 and DPEP1 from the public data website of The Human Protein Atlas; (E) Different expression of genes (*CD74*, *CLCA1* and *DPEP1*) in tumor and normal tissues from TCGA. TCGA data showed that the expression of *CD74* was significantly high in normal tissues (P<0.05), the expression of *DPEP1* was significantly high in tumor tissues (P<0.05), but there was no significance for the expression of *CLCA1* between tumor and normal tissues (P=0.07); (F) Relationships between gene expression levels and survival time. Patients with high *CD74* expression had better survival (P<0.01), but there were no relevance between *CLCA1* or *DPEP1* expression and the prognosis of patients (*CLCA1*, P=0.11; *DPEP1*, P=0.09).

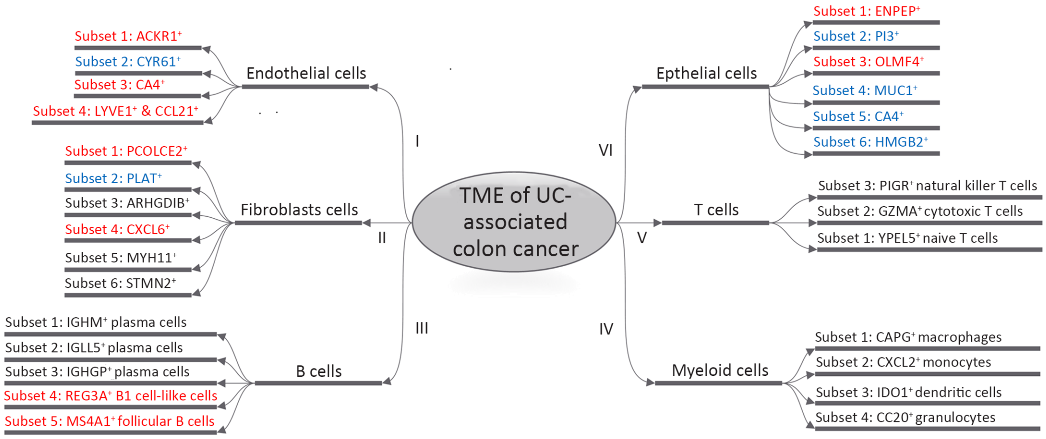
TME of UC-associated colon cancer
I
Endothelial cells
Subset 1: ACKR1+
Subset 2: CYR61+
Subset 3: CA4+
Subset 4: LYVE1+ & CCL21+
II
Fibroblasts cells
Subset 1: PCOLCE2+
Subset 2: PLAT+
Subset 3: ARHGDIB+
Subset 4: CXCL6+
Subset 5: MYH11+
Subset 6: STMN2+
III
B cells
Subset 1: IGHM+ plasma cells
Subset 2: IGLL5+ plasma cells
Subset 3: IGHGP+ plasma cells
Subset 4: REG3A+ B1 cell-lilke cells
Subset 5: MS4A1+ follicular B cells
VI
Epthelial cells
Subset 1: ENPEP+
Subset 2: PI3+
Subset 3: OLMF4+
Subset 4: MUC1+
Subset 5: CA4+
Subset 6: HMGB2+
V
T cells
Subset 3: PIGR+ natural killer T cells
Subset 2: GZMA+ cytotoxic T cells
Subset 1: YPEL5+ naive T cells
IV
Myeloid cells
Subset 1: CAPG+ macrophages
Subset 2: CXCL2+ monocytes
Subset 3: IDO1+ dendritic cells
Subset 4: CC20+ granulocytes

Figure 9 Intra-tumor landscape and statement of TME of UC-associated colon cancer. (Red words mean the cluster of tumor cells, while blue words represent the cluster of non-malignant tissues and others belong to the mixed both tumorous- and non-malignant tissue cells). TME, tumor micro-environment; UC, ulcerative colitis.

role in the evolutionary process from a UC to a CAC. Moreover, we identified the top three differentially expressed genes that played a role in disease progression, *CD74*, *CLCA1* and *DPEP1*. These results may represent a promising strategy that identifies novel potential therapeutic targets in the evolution from UC to CAC and may prevent the development of CRC.

Changes in gene expression in the tumorigenesis trajectory suggest directions for the design of therapies (55). For instance, CD74 is abundant in non-malignant cells, but is down-regulated in tumor endothelial cells. In addition, tumor-derived epithelial cells up-regulate Wnt signaling, but down-regulate DNA repair pathways. Likewise, SCENIC analysis in fibroblasts predicts transcription factors responsible for the transformation toward CAFs, and patients might acquire therapeutic benefits when these conversion processes are blocked. Distinctive features of tumor cells may represent vulnerabilities and provide potential entry points for the design of novel therapies.

To our knowledge, a study which focused on the spatiotemporal evolution from UC to CAC was published (56). In this study, the authors dissected the evolutionary history of CAC using multi-region exome sequencing, but did not perform scRNA-seq and the landscape of cell heterogeneity and evolutionary trajectory in UC-associated CRC revealed by scRNA-seq was absent. It is acknowledged that this is the first study depicting the cell landscape of UC-associated colon cancer at the single-cell transcriptome level, which describes the intra-tumoral heterogeneity mainly from three different viewpoints: gene expression, pathway enrichment, and transcription factor analysis. This provides a more accurate perspective for analyzing the evolutionary progress of UC-associated colon cancer when compared to an average calculation (57).

However, there are also some limitations. Firstly, the results of the study have been determined based on the evolutionary process from UC to CAC in a single patient, which obviously lacks more patients with CAC to compare the obtained results. Secondly, the methods used to classify the cell types are based on distinct and highly-expressed genes, coupled with previous reports. Therefore, an authorized or unified standard should be established. Thirdly, the patient was diagnosed with UC-associated colon cancer and samples from an UC-associated tumor and adjacent inflamed tissues were collected to show the intra-tumor landscape via scRNA-seq. However, the cohort of the TCGA database has sporadic CRC patients, which might cause bias for UC-associated colon cancer. Lastly, three genes (*CD74*, *CLCA1* and *DPEP1*) were found to be a potential role in colon cancer disease progression. However, it is obviously limited and biased since there is only one case and no experiments were done to demonstrate their function in the evolutionary process. To show reliable gene panels, the results should be further validated in larger cohorts in the future, which will be helpful to yield more accurate and convincing results.

Conclusions

This study primarily elucidates the composition of TME and developmental trajectory of UC-associated colon cancer. Furthermore, these results may represent a promising strategy that identifies novel potential therapeutic targets in the evolution from UC to CAC. In the future, the researchers should confirm whether the transcriptome of CAC is associated with consensus molecular subtypes of CRC, and determine whether the transcriptome of the tumor cells could be a signature related to UC among more patients.

Acknowledgements

This study was supported by National Key Research and Development Program of China (No. 2017YFC1308800) and Industry-University-Research Innovation Fund in Ministry of Education of the People's Republic of China (No. 2018A01013).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin 2020;70:145-64.
2. National Health Commission of the People's Republic of China. National guidelines for diagnosis and treatment of colorectal cancer 2020 in China (English version). Chin J Cancer Res 2020;32:415-45.
3. Yang Y, Han Z, Li X, et al. Epidemiology and risk factors of colorectal cancer in China. Chin J Cancer Res 2020;32:729-41.
4. Eaden JA, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut 2001;48:526-35.
5. Ullman T, Odze R, Farraye FA. Diagnosis and management of dysplasia in patients with ulcerative colitis and Crohn's disease of the colon. Inflamm Bowel Dis 2009;15:630-8.
6. Hao XP, Frayling IM, Sgouros JG, et al. The spectrum of p53 mutations in colorectal adenomas differs from that in colorectal carcinomas. Gut 2002;50:834-9.
7. Tarmin L, Yin J, Harpaz N, et al. Adenomatous polyposis coli gene mutations in ulcerative colitis-associated dysplasias and cancers versus sporadic colon neoplasms. Cancer Res 1995;55:2035-8.
8. Burmer GC, Levine DS, Kulander BG, et al. c-Ki-ras mutations in chronic ulcerative colitis and sporadic colon carcinoma. Gastroenterology 1990;99:416-20.
9. Ungaro R, Mehandru S, Allen PB, et al. Ulcerative colitis. Lancet 2017;389:1756-70.
10. Castaño-Milla C, Chaparro M, Gisbert JP. Systematic review with meta-analysis: the declining risk of colorectal cancer in ulcerative colitis. Aliment Pharmacol Ther 2014;39:645-59.
11. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature 2013;501:328-37.
12. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719-24.
13. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature 2017;541:331-8.
14. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352:189-96.
15. Ni X, Zhuo M, Su Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci U S A 2013;110:21083-8.
16. Kim KT, Lee HW, Lee HO, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol 2015;16:127.
17. Sparmann A, Bar-Sagi D. Ras-induced interleukin-8 expression plays a critical role in tumor growth and angiogenesis. Cancer cell 2004;6:447-58.
18. Aibar S, González-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017;14:1083-6.
19. He Q, Zhao L, Liu Y, et al. circ-SHKBP1 regulates the angiogenesis of U87 glioma-exposed endothelial cells through miR-544a/FOXP1 and miR-379/FOXP2 pathways. Mol Ther Nucleic Acids 2018;10:331-48.
20. Chen J, Fu Y, Day DS, et al. VEGF amplifies transcription through ETS1 acetylation to enable angiogenesis. Nat Commun 2017;8:383.
21. Bahrami A, Amerizadeh F, ShahidSales S, et al. Therapeutic potential of targeting Wnt/β-Catenin pathway in treatment of colorectal cancer: Rational and progress. J Cell Biochem 2017;118:1979-83.

22. Kim SH, Park YY, Cho SN, et al. Krüppel-like factor 12 promotes colorectal cancer growth through early growth response protein 1. PloS One 2016;11:e0159899.
23. O'Connell JT, Sugimoto H, Cooke VG, et al. VEGF-A and Tenascin-C produced by S100A4+ stromal cells are important for metastatic colonization. Proc Natl Acad Sci U S A 2011;108:16002-7.
24. Fang M, Yuan J, Peng C, et al. Collagen as a double-edged sword in tumor progression. Tumour Biol 2014;35:2871-82.
25. Cirri P, Chiarugi P. Cancer associated fibroblasts: the dark side of the coin. Am J Cancer Res 2011;1:482-97.
26. Silver DJ, Siebzehnrubl FA, Schildts MJ, et al. Chondroitin sulfate proteoglycans potently inhibit invasion and serve as a central organizer of the brain tumor microenvironment. J Neurosci 2013;33:15603-17.
27. Yopp AC, Shia J, Butte JM, et al. CXCR4 expression predicts patient outcome and recurrence patterns after hepatic resection for colorectal liver metastases. Ann Surg Oncol 2012;19(3 suppl):S339-46.
28. Fearon DT. The carcinoma-associated fibroblast expressing fibroblast activation protein and escape from immune surveillance. Cancer Immunol Res 2014;2:187-93.
29. Voutsadakis IA. Peroxisome proliferator-activated receptor gamma (PPARgamma) and colorectal carcinogenesis. J Cancer Res Clin Oncol 2007; 133:917-28.
30. Yun SH, Roh MS, Jeong JS, et al. Peroxisome proliferator-activated receptor γ coactivator-1α is a predictor of lymph node metastasis and poor prognosis in human colorectal cancer. Ann Diagn Pathol 2018;33:11-6.
31. Lampert F, Stafa D, Goga A, et al. The multi-subunit GID/CTLH E3 ubiquitin ligase promotes cell proliferation and targets the transcription factor Hbp1 for degradation. Elife 2018;7:e35528.
32. Kuranaga Y, Sugito N, Shinohara H, et al. SRSF3, a splicer of the PKM gene, regulates cell growth and maintenance of cancer-specific energy metabolism in colon cancer cells. Int J Mol Sci 2018;19:3012.
33. Huang YH, Zhu C, Kondo Y, et al. CEACAM1 regulates TIM-3-mediated tolerance and exhaustion. Nature 2015;517:386-90.
34. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer 2012;12:252-64.
35. Miele A, Medina R, van Wijnen AJ, et al. The interactome of the histone gene regulatory factor HiNF-P suggests novel cell cycle related roles in transcriptional control and RNA processing. J Cell Biochem 2007;102:136-48.
36. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J Clin Invest 2009;119:1420-8.
37. Hoeijmakers JH. Genome maintenance mechanisms for preventing cancer. Nature 2001;411:366-74.
38. Zheng J, He S, Qi J, et al. Targeted CDX2 expression inhibits aggressive phenotypes of colon cancer cells *in vitro* and *in vivo*. Int J Oncol 2017;51:478-88.
39. Dalerba P, Sahoo D, Paik S, et al. CDX2 as a prognostic biomarker in stage II and stage III colon cancer. N Engl J Med 2016;374:211-22.
40. Li Y, de Haar C, Chen M, et al. Disease-related expression of the IL6/STAT3/SOCS3 signalling pathway in ulcerative colitis and ulcerative colitis-related carcinogenesis. Gut 2010;59:227-35.
41. Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012;481:306-13.
42. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381-6.
43. Ghoochani A, Schwarz MA, Yakubov E, et al. MIF-CD74 signaling impedes microglial M1 polarization and facilitates brain tumorigenesis. Oncogene 2016;35:6246-61.
44. Murayama T, Nakaoku T, Enari M, et al. Oncogenic fusion gene CD74-NRG1 confers cancer stem cell-like properties in lung cancer through a IGF2 autocrine/paracrine circuit. Cancer Res 2016;76: 974-83.
45. Bozzi F, Mogavero A, Varinelli L, et al. MIF/CD74 axis is a target for novel therapies in colon carcinomatosis. J Exp Clin Cancer Res 2017;36:16.
46. Gruber AD, Elble RC, Ji HL, et al. Genomic cloning, molecular characterization, and functional analysis of human CLCA1, the first human member of the family of Ca2+-activated Cl- channel proteins. Genomics 1998;54:200-14.
47. Toda M, Tulic MK, Levitt RC, et al. A calcium-activated chloride channel (HCLCA1) is strongly related to IL-9 expression and mucus production in bronchial epithelium of patients with asthma. J

Allergy Clin Immunol 2002;109:246-50.

48. Bustin SA, Li SR, Dorudi S. Expression of the Ca2+-activated chloride channel genes CLCA1 and CLCA2 is downregulated in human colorectal cancer. DNA Cell Biol 2001;20:331-8.

49. Yang B, Cao L, Liu J, et al. Low expression of chloride channel accessory 1 predicts a poor prognosis in colorectal cancer. Cancer 2015;121:1570-80.

50. Nakagawa H, Inazawa J, Inoue K, et al. Assignment of the human renal dipeptidase gene (DPEP1) to band q24 of chromosome 16. Cytogenet Cell Genet 1992;59:258-60.

51. McIver CM, Lloyd JM, Hewett PJ, et al. Dipeptidase 1: a candidate tumor-specific molecular marker in colorectal carcinoma. Cancer Lett 2004;209:67-74.

52. Ziegenhain C, Vieth B, Parekh S, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell 2017;65:631-43.e4.

53. Haque A, Engel J, Teichmann SA, et al. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 2017;9:75.

54. Young MD, Mitchell TJ, Vieira Braga FA, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science 2018;361:594-9.

55. Song M, Willett WC, Hu FB, et al. Trajectory of body shape across the lifespan and cancer risk. Int J Cancer 2016;138:2383-95.

56. Baker AM, Cross W, Curtius K, et al. Evolutionary history of human colitis-associated colorectal cancer. Gut 2019;68:985-95.

57. Hirsch D, Wangsa D, Zhu YJ, et al. Dynamics of genome alterations in Crohn's disease-associated colorectal carcinogenesis. Clin Cancer Res 2018;24:4997-5011.

Cite this article as: Wang Q, Wang Z, Zhang Z, Zhang W, Zhang M, Shen Z, Ye Y, Jiang K, Wang S. Landscape of cell heterogeneity and evolutionary trajectory in ulcerative colitis-associated colon cancer revealed by single-cell RNA sequencing. Chin J Cancer Res 2021;33(2):271-288. doi: 10.21147/j.issn.1000-9604.2021.02.13

Supplementary materials

Methods and protocols of single cells

Preparation of single-cell suspensions

Following resection in the operating room, samples of the tumor and adjacent non-malignant colon tissues at maximal distance (>5 cm) (*Supplementary Figure S1*), which were isolated and saved in Dulbecco's Modified Eagle Medium (DMEM, Gibco) and transported rapidly to the research facility. On arrival, samples were cut into smaller pieces of less than 1 mm^3 and rinsed with Hanks' balanced salt solution (HBSS, ThermoFisher Scientific), and then transferred to 10 mL of digestion medium containing 0.2% collagenase IV (ThermoFisher Scientific) in DMEM. Samples were incubated for 75 min at 37 °C, with manual shaking every 5 min. Next, 30 mL of ice-cold HBSS was added and samples were filtered using a 40-μm nylon mesh (ThermoFisher Scientific). Following centrifugation at 120,000 r/min and 4 °C for 3 min, the supernatant was decanted and discarded, and the cell pellet was resuspended in 2 mL of red blood cell lysis buffer and transferred to a 2-mL DNA low-binding tube. Following a 5-min incubation at room temperature, samples were centrifuged (120,000 r/min, 4 °C, 5 min) using a swing-out rotor. And then resuspended in 2 mL of HBSS and centrifuged (120,000 r/min, 4 °C, 2 min) using a swing-out rotor. Subsequently, samples were resuspended in 2 mL of DMEM containing 10% fetal bovine serum (Gibco).

Droplet-based single-cell RNA sequencing (scRNA-seq)

Single-cell suspensions were converted to barcoded scRNA-seq libraries by using the Chromium Single Cell 3' Library, Gel Bead & Multiplex Kit and Chip Kit (10× Genomics), aiming for an estimated 5,000 cells per library and following the manufacturer's instructions. Samples are processed using kits pertaining to either the V1 or V2 barcoding chemistry from 10x Genomics. Single samples are always processed in a single well of a polymerase chain reaction (PCR) plate, allowing all cells from a sample to be treated with the same master mix and in the same reaction vessel. All samples (non-malignant and tumor) were processed in parallel on the same thermal cycler. Libraries were sequenced on an Illumina HiSeq4000 and mapped to the human genome (build hg19) using CellRanger (10× Genomics).

Single-cell gene expression quantification and determination of major cell types

Raw gene expression matrices generated per sample using CellRanger (Version 2.0.0; https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines) were combined in R (Version 3.3.2; https://www.r-project.org/), and converted to a Seurat object using the Seurat R package (Version 1.4.0.7; https://cran.r-project.org/web/packages/Seurat/index.html). From this, all cells were excluded which had either fewer than 201 unique molecular identifiers (UMIs), over

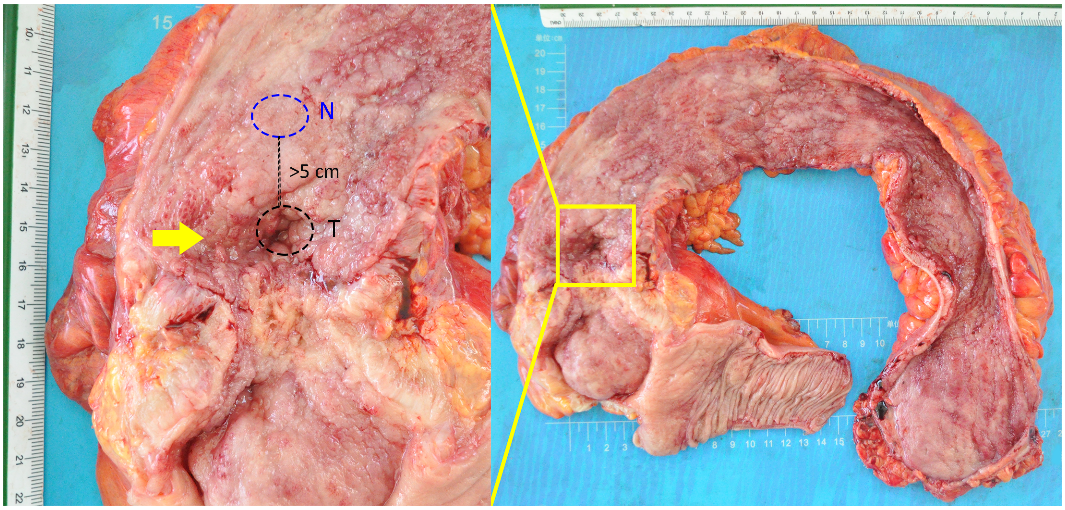

Figure S1 A tumor tissue sample and a non-malignant colon tissue sample (>5 cm away from the neoplastic foci) were obtained following surgical resection. N, non-malignant colon tissue; T, tumor tissue.

6,000 or under 101 expressed genes, or over 10% UMIs derived from the mitochondrial genome. From the remaining cells, gene expression matrices were normalized to total cellular read count and to mitochondrial read count using linear regression as implemented in Seurat's Regress Out function. As a result, none of the principle components subsequently identified were correlated with transcript count. From the remaining cells, variably expressed genes were selected as having a normalized expression between 0.125 and 3 and a quantile-normalized variance exceeding 0.5. To reduce the dimensionality of this data set, the resulting variably expressed genes were summarized by principle component analysis (PCA), and the first six principle components were further summarized using t-distributed stochastic neighbor embedding (tSNE) dimensionality reduction using the default settings of the Run tSNE function. Cell clusters in the resulting two-dimensional representation were annotated to known biological cell types using canonical marker genes.

Subclustering of major cell types

To identify subclusters within these six cell types, we reanalyzed cells belonging to each of these six cell types separately. Specifically, we applied dimensionality reduction using principal component analysis (PCA) in each cell type on variably expressed genes as described above. To identify which principle components were informative, we applied Horn's parallel analysis for PCA as implemented in the R paran package (Version 1.5.1; https://cran.r-project.org/web/packages/paran/index.html), selecting those principle components having eigenvalues that exceed the eigenvalues generated using ten random permutations by >50%. Using the graph-based clustering approach implemented in the Find Clusters function of the Seurat package, with a conservative resolution of 0.5 and otherwise default parameters, each cell type was reclustered by its principle components. Notably, subclustering was robust to alterations in the number of principle components, in the resolution or in the K parameter. Moreover, few of the subclusters identified contained many cells wherein less than 300 genes were detected, indicating that increasing the threshold of 100 genes would not affect our results. This yielded 28 subclusters in total, as listed in *Supplementary Table S1*. For visualization purposes, these informative principle components were converted into tSNE plots as above.

Identification of marker genes

To identify marker genes for each of these 28 subclusters within these six cell types, we contrasted cells from that subcluster to all other cells of that subcluster using the Seurat Find Markers function. Marker genes were required to have an average expression in that subcluster that was >2.5-fold higher than the average expression in the other subclusters from that cell type, and a detectable expression in >15% of all cells from that subcluster. Additionally, marker genes were required to have the highest mean expression in that subcluster.

Single-cell regulatory network inference and clustering (SCENIC) analysis

SCENIC analysis was run as previously described on the 4,777 cells that remained after filtering, using the 20-thousand motifs database for RcisTarget and GRNboost (SCENIC Version 0.1.5, which corresponds to RcisTarget 0.99.0 and AUCell 0.99.5; with RcisTarget.hg19.motifDatabases.20k).

Analysis of differential pathway or regulon activities

To assess the differential activities of pathways (GSVA) or regulons (SCENIC) between cell sets (for example, derived from tumor or normal samples, or belonging to different subclusters), we contrasted the activity scores for each cell using a generalized linear model. To avoid inflating signals because of inter individual differences (for example, in the relative frequencies of cells from different patients), we always included the patient of origin as a categorical variable. The results of these linear models were visualized using bar plots or heat maps. For the latter, pathways or regulons that did not show significant changes (Benjamini-Hochberg-corrected P>0.05) in any of the cell sets that were contradictory in one analysis were not visualized.

Statistics and reproducibility

No statistical method was used to predetermine sample sizes. For all experiments, samples from a single patient were processed in parallel, and cells for each sample of one patient were processed for scRNA-seq (10× Genomics) at the same time, but in separate lanes and vials. Box plots were generated using the R base package and default parameters. Hence, the boxes span the interquartile range (IQR; from the 25th to the 75th percentiles), with the centerline corresponding to the median. Lower whiskers represent data minimum or the 25th percentile minus 1.5×IQR, whichever is greater. Upper whiskers represent the data maximum or the 75th percentile plus 1.5×IQR (lower), whichever is lower. Violin plots were generated using the bean plot R package, and the data distribution band width was estimated by kernel density estimation, as per the built-in "nrd0" option. Bar plots indicate the mean ± standard error of the mean and include individual data points. Given the number of data points represented on box and violin plots, we opted not to display each data point, as this would obscure the overall distribution. Comparisons between two groups were performed using unpaired two-tailed *t*-tests. One-way analysis of variance (ANOVA) with Tukey's multiple comparisons tests were used for multiple group comparisons. Linear models were generated when multiple parameters were taken into account. Fitting of Cox proportional hazards regression models was performed using the Cox Proportional Hazards (CoxPH) function in the R survival package (https://cran.r-project.org/web/packages/survival/index.html), with tied death times handled using the Breslow method. All statistical analyses and presentation of data were performed using R (https://www.r-project.org/).

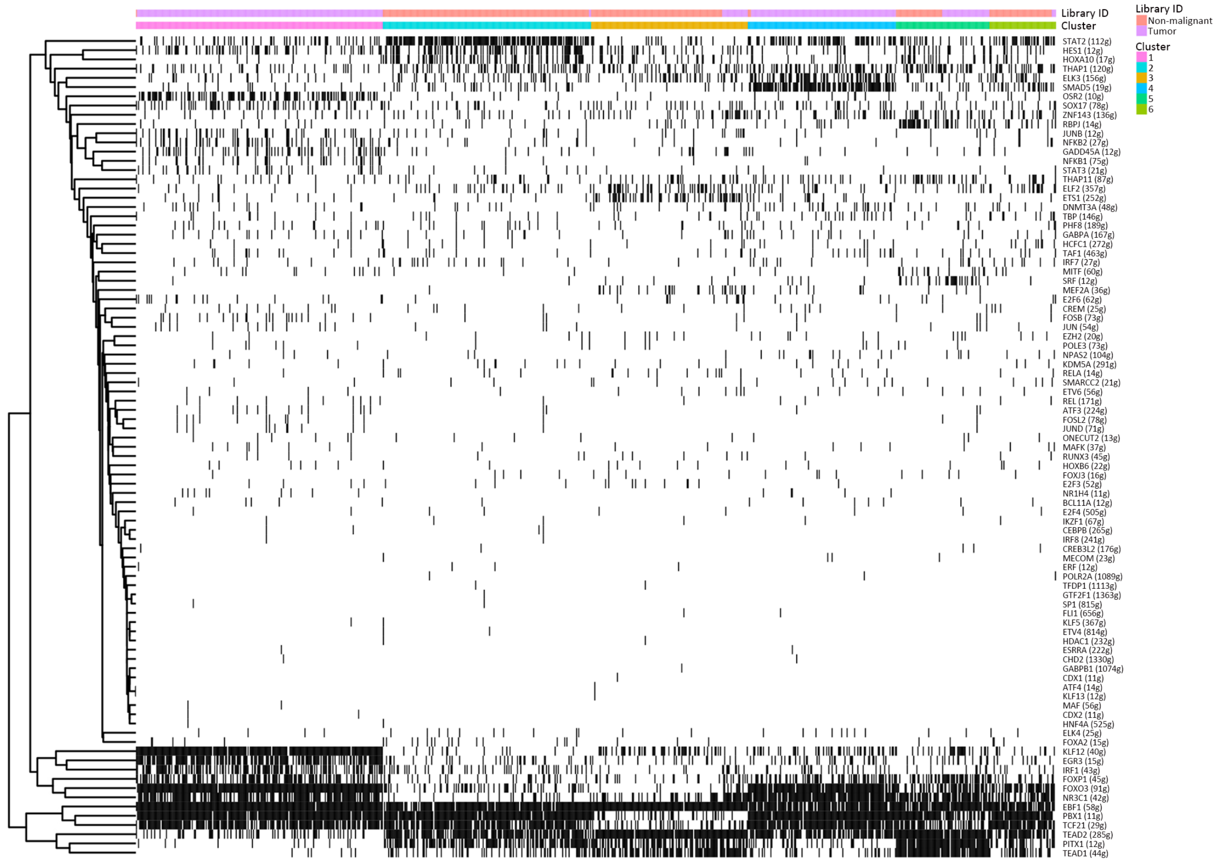

Figure S2 SCENIC analysis of transcription factors involved among the fibroblasts clusters, and showed that *KLF12*, which promotes CRC growth, was highly expressed in cluster 1. SCENIC, single-cell regulatory network inference and clustering; CRC, colorectal cancer.

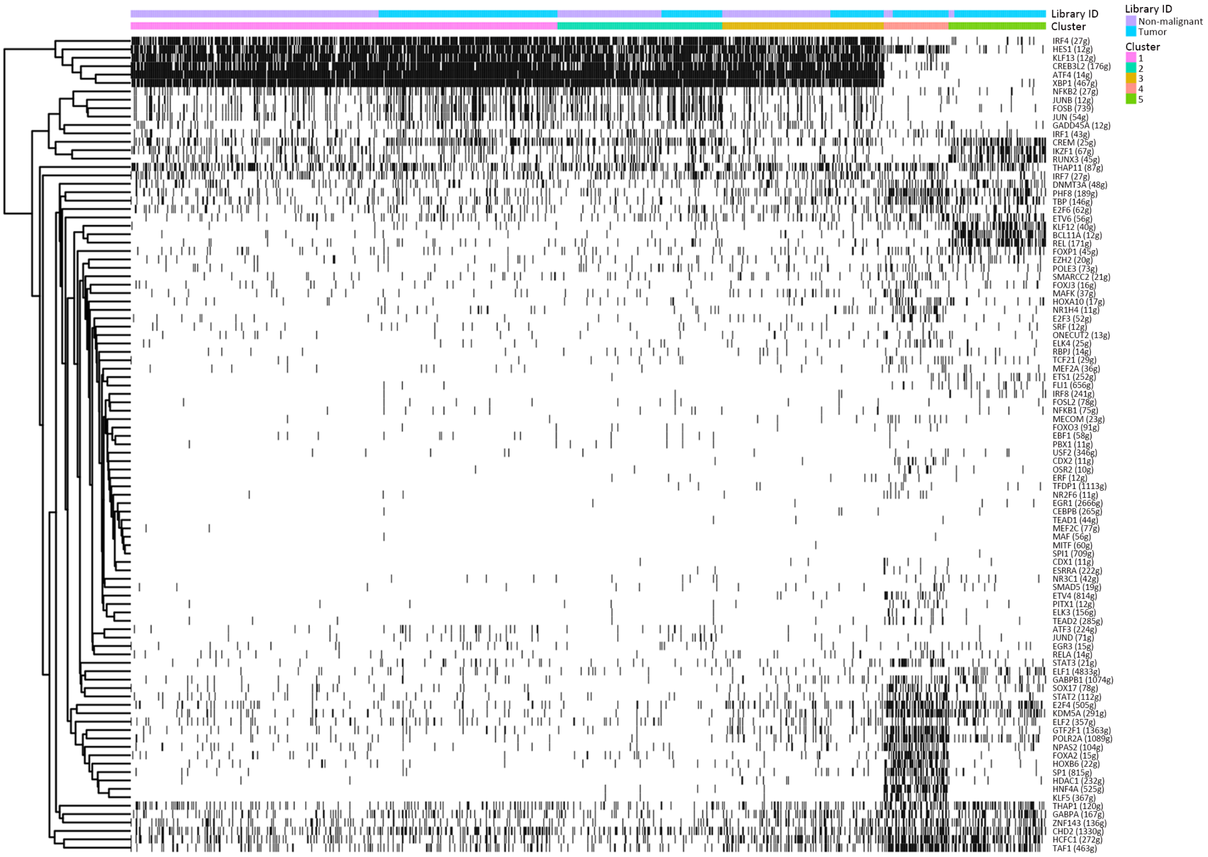

Figure S3 SCENIC analysis of transcription factors involved among B cell clusters, and SCENIC analysis failed to identify differences between non-malignant tissue-derived and tumor-derived plasma cells. SCENIC, single-cell regulatory network inference and clustering.

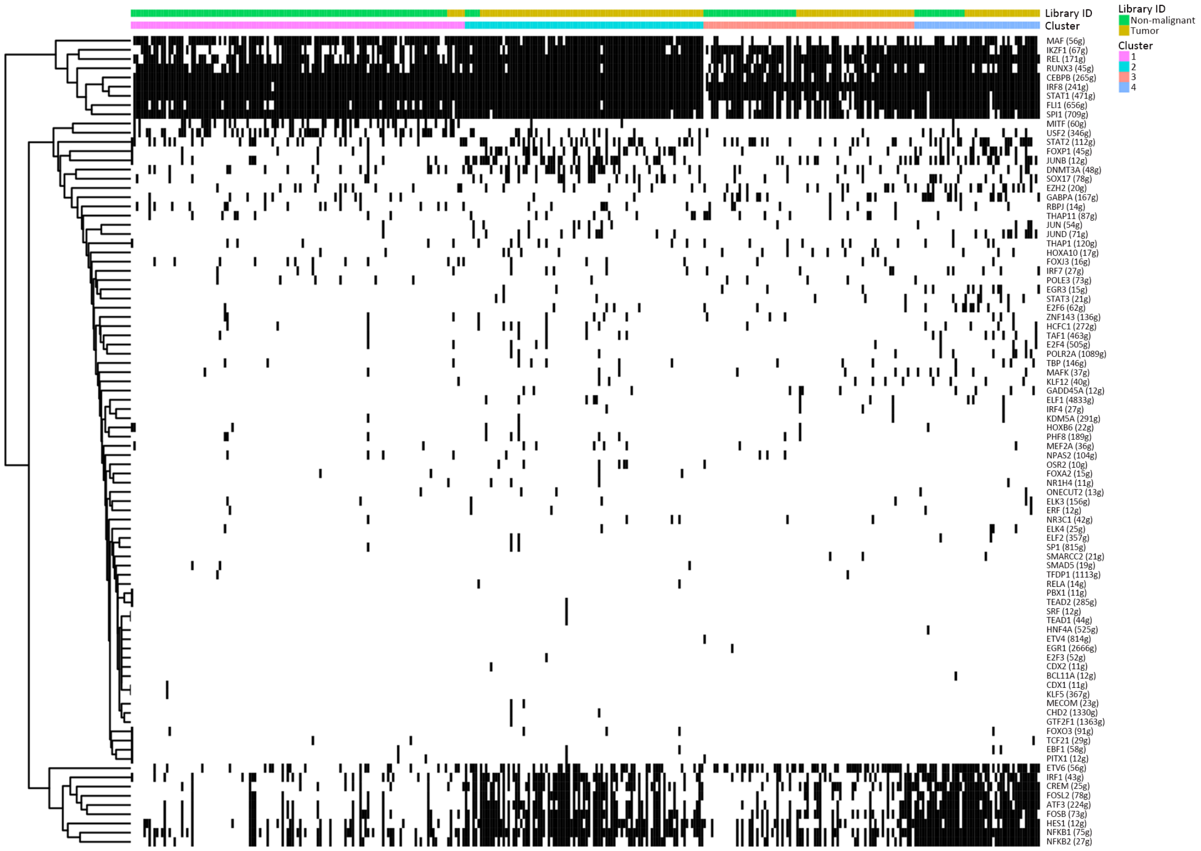

Figure S4 SCENIC analysis of transcription factors involved among myeloid cell clusters. The cell numbers of macrophages and monocytes displayed extensive heterogeneity in tumors compared with those in non-malignant tissues, but SCENIC analysis showed no significance among different derived cells. SCENIC, single-cell regulatory network inference and clustering.

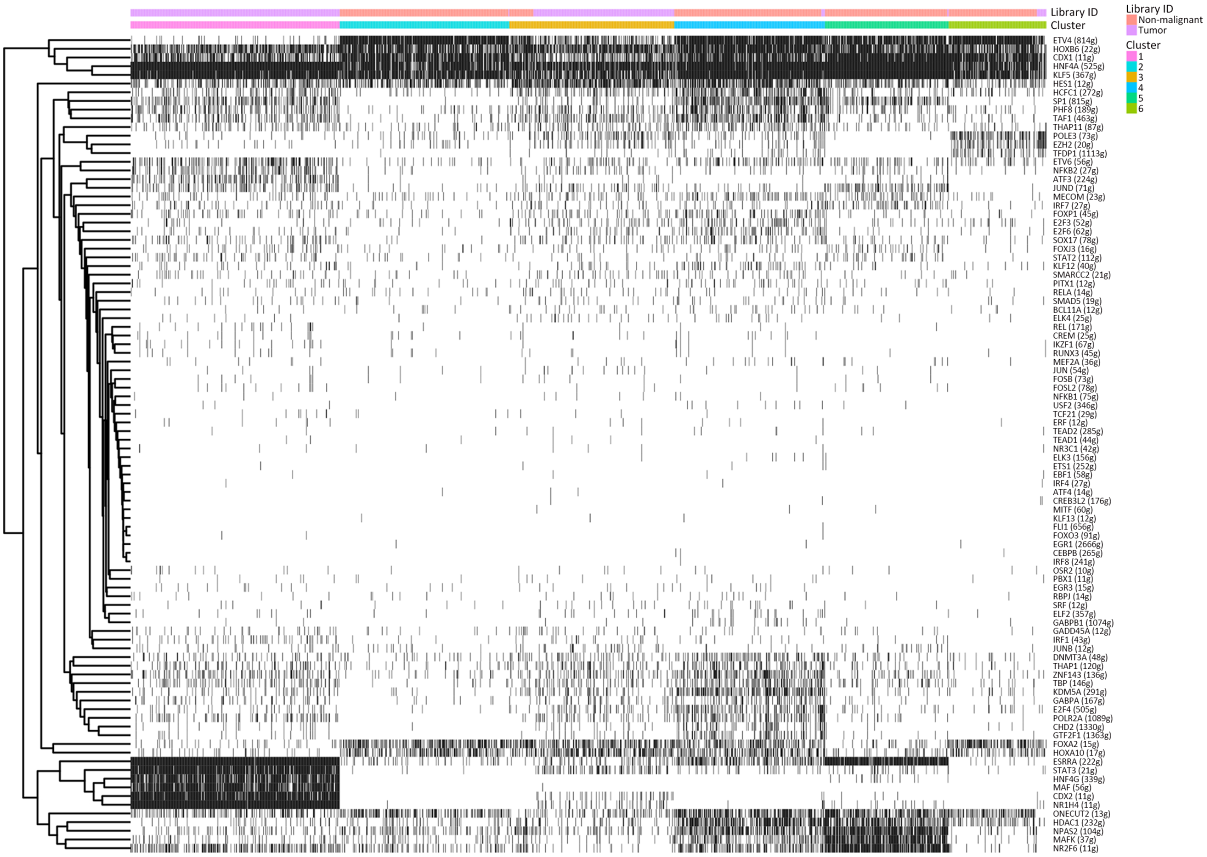

Figure S5 SCENIC analysis of transcription factors involved among the epithelial clusters. The transcription factors CDX2 and STAT3 were significantly up-regulated in cluster 1, but showed almost no expression in other clusters besides cluster 3. SCENIC, single-cell regulatory network inference and clustering.

Table S1 Maker genes for cell types and subclusters

Cluster/sub sets	Annotated name	Cells (N)	Tumor cells [n (%)]	Maker genes	Full spelling of abbreviation
Endothelial cells-1	Vascular endothelial cells	66	50 (75.76)	*ACKR1*	Atypical chemokine receptor 1
				AKAP12	A-Kinase anchoring protein 12
				SELE	Selectin E
				CCL2	C-C motif chemokine ligand 2
Endothelial cells-2	Vascular endothelial cells	74	7 (9.46)	*CYR61*	Cellular communication network factor 1
				IGHG1	Immunoglobulin heavy constant gamma 1 (G1m Marker)
				IGHG3	Immunoglobulin heavy constant gamma 3 (G1m Marker)
				IGHG4	Immunoglobulin heavy constant gamma 4 (G1m Marker)
Endothelial cells-3	Vascular endothelial cells	63	49 (77.78)	*CA4*	Carbonic anhydrase 4
				CD32	Fc fragment of IgG receptor IIa (FCGR2A)
				FLT1	Fms related receptor tyrosine kinase 1
Endothelial cells-4	Lymphatic endothelial cells	25	25 (100)	*LYVE1*	Lymphatic vessel endothelial hyaluronan receptor 1
				CCL21	C-C motif chemokine ligand 21
				EFEMP1	EGF containing fibulin extracellular matrix protein 1
				CLU	Clusterin
				MMRN1	Multimerin 1
Fibroblast-1	CAF	132	128 (96.97)	*PCOLCE*	Procollagen C-endopeptidase enhancer
				MFAP5	Microfibril Associated Protein 5
				SFRP2	Secreted frizzled related protein 2
				PI16	Peptidase inhibitor 16
Fibroblast-2		298	15 (5.03)	*PLAT*	Plasminogen activator, tissue type
				AGT	Angiotensinogen
				CTHRC1	Collagen triple helix repeat containing 1
				CARD16	Caspase recruitment domain family member 16
Fibroblast-3		140	62 (44.29)	*ARHGDIB*	Rho GDP dissociation inhibitor beta
				RGS5	Regulator of G protein signaling 5
				MEF2C	Myocyte enhancer factor 2C
				CSRP2	Cysteine and glycine rich protein 2
Fibroblast-4	CAF	138	135 (97.83)	*CXCL6*	C-X-C motif chemokine ligand 6
				CCL13	C-C motif chemokine ligand 13
				HAPLN1	Hyaluronan and proteoglycan link protein 1
				CCL8	C-C motif chemokine ligand 8
Fibroblast-5		87	52 (59.77)	*MYH11*	Myosin heavy chain 11
				DES	Desmin
				CNN1	Calponin 1
				PLN	Phospholamban
Fibroblast-6		62	4 (6.45)	*STMN2*	Stathmin 2
				ADAM28	ADAM metallopeptidase domain 28

Table S1 (*continued*)

Table S1 (*continued*)

Cluster/subsets	Annotated name	Cells (N)	Tumor cells [n (%)]	Maker genes	Full spelling of abbreviation
B cells-1	Plasma cells	513	231 (45.03)	*IGHG1*	Immunoglobulin heavy constant gamma 1 (G1m marker)
				IGHG2	Immunoglobulin heavy constant gamma 2 (G1m marker)
				IGHG3	Immunoglobulin heavy constant gamma 3 (G1m marker)
				IGHG4	Immunoglobulin heavy constant gamma 4 (G1m marker)
				APOE	Apolipoprotein E
B cells-2	Plasma cells	178	75 (42.13)	*IGHG1*	Immunoglobulin heavy constant gamma 1 (G1m marker)
				IGHG2	Immunoglobulin heavy constant gamma 2 (G1m marker)
				IGHG3	Immunoglobulin heavy constant gamma 3 (G1m marker)
				IGLL5	Immunoglobulin lambda like polypeptide 5
B cells-3	Plasma cells	184	62 (33.70)	*IGHG4*	Immunoglobulin heavy constant gamma 4 (G4m marker)
				IGHA	Immunoglobulin heavy constant alpha 1
B cells-4	B1 cell-like cells	88	80 (90.91)	*REG3A*	Regenerating family member 3 alpha
				DEFA6	Defensin alpha 6
				PRSS2	Serine protease 2
				ITLN2	Intelectin 2
				LYZ	Lysozyme
B cells-5	follicular B cells	137	123 (89.78)	*MS4A1*	Membrane spanning 4-domains A1
				LTB	Lymphotoxin beta
				HLA-DRB1	Major histocompatibility complex, class II, DR beta 1
				CD37	CD37 molecule
				APC	APC regulator of WNT signaling pathway
				CD52	CD52 Molecule
Myeloid cells-1	Macrophages	133	12 (9.02)	*CAPG*	Capping actin protein, gelsolin like
				TREM2	Triggering receptor expressed on myeloid cells 2
				GPNMB	Glycoprotein Nmb
				CAPG	Capping actin protein, gelsolin like
				CHI3L1	Chitinase 3 Like 1
Myeloid cells-2	Monocytes	95	85 (89.47)	*CXCL2*	C-X-C motif chemokine ligand 2
				PRDM1	PR/SET domain 1
Myeloid cells-3	Dendritic cells	84	40 (47.62)	*IDO1*	Indoleamine 2,3-dioxygenase 1
				CD1E	CD1e molecule
				CLEC9A	C-Type lectin domain containing 9A
Myeloid cells-4	Granulocytes	50	36 (72.00)	*CCL20*	C-C motif chemokine ligand 20
				S100A12	S100 calcium binding protein A12
				PTGS2	Prostaglandin-endoperoxide synthase 2
				G0S2	G0/G1 Switch 2--- G0S2
				IL1-B	Interleukin 1 beta
T cells-1	Helper cells	154	132 (85.71)	*YPEL5*	Yippee like 5
				TSC22D3	TSC22 domain family member 3
				GPR183	G protein-coupled receptor 183

Table S1 (*continued*)

Table S1 (*continued*)

Cluster/sub sets	Annotated name	Cells (N)	Tumor cells [n (%)]	Maker genes	Full spelling of abbreviation
T cells-2	Cytotoxic lymphocytes	125	55 (44.00)	*GZMA*	Granzyme A
				KIAA0101	PCNA clamp associated factor (PCLAF)
				TUBB	Tubulin beta class I
				STMN1	Stathmin 1
				HIST1H4C	Histone cluster 1 H4 family member C
				HMGB2	High mobility group box 2
				PKM	Pyruvate kinase M1/2
				GNLY	Granulysin
T cells-3	Natural killer cells	39	28 (71.79)	*PIGR*	Polymeric immunoglobulin receptor
				NTS	Neurotensin
				TPSB2	Tryptase beta 2
				TPSAB1	Tryptase alpha/beta 1
				KRT8	Keratin 8
				ELF3	E74 Like ETS transcription factor 3
				KRT13	Keratin 13
				PHGR1	Proline, histidine and glycine rich 1
				DEFA5	Defensin alpha 5
Epithelial cells-1	Cancer cells	437	433 (99.08)	*ENPEP*	Glutamyl aminopeptidase
				RBP2	Retinol binding protein 2
				APOC3	Apolipoprotein C3
				APOA1	Apolipoprotein A1
				CYP3A4	Cytochrome P450 family 3 subfamily a member 4
				SLC15A1	Solute carrier family 15 member 1
				MGAM	Maltase-glucoamylase
				KHK	Ketohexokinase
				ALPI	Alkaline phosphatase, intestinal
Epithelial cells-2		354	11 (3.11)	*PI3*	Peptidase inhibitor 3
				TIMP1	TIMP metallopeptidase inhibitor 1
				SLPI	Secretory leukocyte peptidase inhibitor
				STAT3	Signal transducer and activator of transcription 3
				OCIAD2	OCIA domain containing 2
Epithelial cells-3	Cancer cells	344	302 (87.79)	*OLFM4*	Olfactomedin 4
				REG1A	Regenerating family member 1 alpha
				CLCA1	Chloride channel accessory 1
				MT1G	Metallothionein 1G
Epithelial cells-4		315	3 (0.95)	*MUC1*	Mucin 1, cell surface associated
				ABCC3	ATP binding cassette subfamily c member 3
				STT3-B	STT3 oligosaccharyltransferase complex catalytic subunit B
				MT-CO2	Mitochondrially encoded cytochrome C oxidase II

Table S1 (*continued*)

Table S1 (*continued*)

Cluster/sub sets	Annotated name	Cells (N)	Tumor cells [n (%)]	Maker genes	Full spelling of abbreviation
Epithelial cells-5		258	5 (1.94)	*CA4*	Carbonic anhydrase 4
				TM4SF4	Transmembrane 4 L six family member 4
				EMP1	Epithelial membrane protein 1
				SAA1	Serum amyloid A1
				DUOX2	Dual oxidase 2
Epithelial cells-6		204	10 (4.90)	*HMGB2*	High mobility group box 2
				HIST1H4C	Histone cluster 1 H4 family member C
				TUBA1B	Tubulin alpha 1b
				HMGN2	High mobility group nucleosomal binding domain 2

CAF, cancer associated fibroblast.