Landscape of Cell Heterogeneity and Evolutionary Trajectory in Ulcerative Colitis-associated Colon Cancer Revealed by Single-cell Rna Sequencing

Quan Wang
Peking University People's Hospital
Zhu Wang
Shandong Provincial Hospital
Zhen Zhang
Peking University People's Hospital
Wei Zhang
Peking University People's Hospital
Mengmeng Zhang
Peking University People's Hospital
Zhanlong Shen
Peking University People's Hospital
Yingjiang Ye
Peking University People's Hospital
Kewei Jiang
Peking University People's Hospital
Shan Wang (shanwang60@sina.com)
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 https://orcid.org/0000-0003-3396-5936

Primary research

Keywords: ulcerative colitis-associated colon cancer, single-cell RNA sequencing, cell heterogeneity, evolutionary trajectory

Posted Date: August 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-60886/v1

Version of Record: A version of this preprint was published at Chinese Journal of Cancer Research on January 1st, 2021. See the published version at https://doi.org/10.21147/j.issn.1000-9604.2021.02.13.

Abstract

Background: Patients with colitis-associated cancer (CAC), a particular kind of colorectal cancer that develops from inflammatory bowel diseases (IBDs), have an earlier morbidity and a poorer prognosis. However, in CAC, single cell transcriptome analysis of the microenvironment composition and characteristics has yet to be performed. To understand the intra-tumor heterogeneity in CAC and to reveal a potential evolutionary trajectory from ulcerative colitis (UC) to CAC at the single cell level.

Methods: Fresh samples of tumor- and adjacent tissue, from a CAC patient with pT3N1M0, 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 prognosis.

Results: Ultimately, 4,777 single-cell transcriptomes (1220 genes per cell) were studied, which composed of 2,250 (47%) and 2,527(53%) originated from tumor- and non-malignant tissue, respectively. And we defined the composition of cancer-associated stromal cells and identified six cell clusters included myeloid, T and B cells, fibroblasts, endothelial and epithelial cells. Likewise, the notable pathways and transcription factors (TFs) involved of these cell clusters were analyzed and described. Moreover, we graphed the precise cellular composition and developmental trajectory from UC to UC-associated colon cancer, and predicted that *CD74, CLCA1* and *DPEP1* had a potential role in the disease progression.

Conclusions: scRNA-seq technology could reveal intratumor cell heterogeneity in ulcerative colitis-associated colon cancer, and might provide a promising direction to seek the novel potential therapeutic targets in the evolution from IBD to CAC.

Novelty And Impact

The different subgroups of tumor cells were illustrated in the tumor microenvironment. There were six clusters of cells displayed and described their related notable pathways involved in the disease of colitis-associated colorectal cancer at the single cell level. At the meantime, *CD74, CLCA1* and *DPEP1*, as the top three differentially expressed genes, were identified to have a potential role in the disease progression, which may represent a promising strategies to seek a novel potential therapeutic target on the evolution from ulcerative colitis to colitis-associated colorectal cancer.

Introduction

Colorectal cancer (CRC) is the third most common cancer type worldwide, and was estimated approximately 147,950 new cases and 53,200 deaths in 2020 in the United States, which has been the second most common cause of cancer death, even including 17,930 cases and 3,640 deaths under 50 years old(1). In parallel, patients with colitis-associated cancer (CAC), a particular kind of CRC that develops from inflammatory bowel diseases (IBDs), have an earlier morbidity and a poorer prognosis(2). CAC is often thought to arise from flat dysplasia with indistinct margins, in a field of concomitant

inflammation, scarring and pseudopolyposis, rather than developing from a polypoid adenoma, which is the major cause of sporadic CRC (S-CRC)(3). Furthermore, the sequence of events leading to CAC is distinct from S-CRC at the molecular level. A distinct set of genes, including *TP53*(4), *APC*(5) and *KRAS*(6) contain more mutations than in CACs. However, we still have a poor understanding of the molecular processes underlying colorectal carcinogenesis in IBDs. Ulcerative colitis (UC), the most common form of IBD, has become increasingly prevalent worldwide(7). In a 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(8), thus indicating the requirement for intensive study.

In the past, research on tumor origin targeted only 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-tumor 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(9). Obtaining accurate TME information not only helps to gain a better understanding of the tumor origin and development, but also contributes to the development of new therapeutic targets.

Genomic and transcriptomic studies have revealed driver mutations, aberrant regulatory programs, and disease subtypes for major human tumors (10). However, these studies relied on profiling technologies that measure tumors in bulk, resulting in data that represent an 'average' of all cells present, thus limiting their ability to capture intra-tumor heterogeneity. Recent advances in single-cell sequencing provide an avenue to explore genetic and functional heterogeneity at a cellular resolution(11). Single-cell RNA sequencing (scRNA-seq) combined with computational methods for functional clustering of cell types provides a less biased approach to understanding cellular heterogeneity. It has been used in many studies of human tumors (12), circulating tumor cells (CTCs)(13) and patient-derived xenografts (14) 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 determine the evolutionary process of UC-associated colon cancer. To our knowledge, this is the first study depicting the cell landscape of TME in UC-associated colon cancer at the single-cell level.

Materials And Methods

Human specimen collection

Fresh tumor and non-malignant tissues were taken from a 43-year-old female patient with UC related colon adenocarcinoma who had been a history of UC for eight years, and the postoperative pathology was classified as pT3N1M0 (median differentiation ulcerative adenocarcinoma, MSS+). The written informed consents were 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 operative resection, a tumor tissue sample and a non-malignant colon tissue sample, which was at least 5 cm away from the neoplastic foci (**Fig. S1**).

Protocols of scRNA-seqand Data quality control

In terms of preparation of single-cell suspensions, Droplet-based scRNA-seq and others methods related single cells analysis, their protocols were described in ***Supplementary Materials.***Single cells were filtered for related analysis basing upon the criteria, all cells were removed that had either fewer than 201 UMIs (unique molecular identifier), over 6,000 or under 101 expressed genes, or over 10% UMIs derived from the mitochondrial genome. Gene expression (in UMI) was scale-normalized and transformed in log2 (UMI+1).

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

PCA was used to summary the resulting variably expressed genes to reduce the dimensionality of this data set. And tSNE, to recalculate the sample distance by the conditional probability of the random neighbour fitting based on the Student T distribution in the high dimensional space, 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 the high-level functions and effects of the biological system (http://www.genome.jp/kegg/), which was performed via DAVID (https://david.ncifcrf.gov/). We defined the enriched pathways with a Q value ≤ 0.05 were regarded as significant difference, as same as the Functional annotation through the Gene Ontology database, which the Fisher exact test was used to select only significant categories including biological process, cellular component and molecular function classifications, and Q value ≤ 0.05 were also considered to be 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 TCGA (http://www.cbioportal.org/). Specifically, we downloaded preprocessed gene expression data as well as clinical data for primary solid tumors and normal solid tissue, for colon and rectal adenocarcinoma, using the Bioconductor TCGA biolinks package (version 2.2.10). The CRC samples were divided into high and low expression groups based on gene median expression level. Z scores from TCGA and the validation cohort were combined using the weighted Z method. For Kaplan–Meier plots, marker gene expression categorization was optimized.

Gene Immunohistochemical staining (IHC) in The Human Protein Atlas

IHC 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/), to confirm the expression change in these genes along with disease progression.

Results

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

Once the non-malignant and CAC tissues were obtained, we rapidly digested to a single-cell suspension and analyzed using scRNA-seq involving a single-tube protocol with unique transcript counting through barcoding with unique molecular identifiers (UMIs) (**Fig. S2A**). To obtain detailed cellular genetic information on this tumor, we took more than 1.6 billion post-normalization reads for subsequent analysis, which were obtained from 4,777 cells and a median of 1220 genes per cell were expressed. Of the sample cells, 2,250 (47%) originated from tumor tissue and 2,527(53%) originated from non-malignant tissue (**Fig. S2B**). Following gene expression normalization for read depth and mitochondrial read count, we applied PCA to genes variably expressed across all 4,777 cells (n= 1,220 genes). Subsequently, we classified the cells into groups 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, we identified myeloid cells, T cells, B cells, fibroblasts, endothelial cells and epithelial cells (**Fig. S2B & S2C**; ***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 cell source and numbers between the diverse subgroups (**Fig. S2D**).

2. Different angiogenesis pathways in tumor and non-malignant endothelial cells

We detected 228 endothelial cells and revealed four clusters (**Fig. 1A**). We next attempted to identify marker genes for each of these clusters and to assign them to known endothelial cell types (**Fig. 1B**; ***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*$^{+}$). The other set of 25 lymphatic endothelial cells was found in the tumor sample (cluster 4; marker genes *LYVE1*$^{+}$ and *CCL21*$^{+}$), but interestingly, no distinct cluster of lymphatic endothelial cells from the non-malignant sample was found.

Except for lymphocytes-composed cluster 4, we used hallmark pathway gene signatures analysis to find the different characteristics among the other three clusters, which were all from vascular endothelial cells (**Fig. 1C**). Interestingly, 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 (**Fig. 1D**). Detailed analysis showed that the two pathways were both angiogenesis-related, but the different tissue-derived clusters seemed to have diverse mechanisms. The role of *KRAS* oncogenes in promoting cellular transformation is well established, and KRAS modulates tumor-stroma interactions and supports cancer invasiveness by influencing the expression of metalloproteinases and cytokines involved in angiogenesis(15). In our study, the result of the high enrichment of the *KRAS* pathway in tumor tissues was in accordance with the hyper-vascular nature of tumors.

We next applied single-cell regulatory network inference and clustering (SCENIC)(16), which scans differentially expressed genes for overrepresented transcription factor binding sites, and analyzes the co-

expression of transcription factors (TFs) and their putative target genes (**Fig. 1E & Fig. 1F**). Many TFs took part in attending angiogenesis, including *FOXP1* and *ETS1*. Of note, *FOXP1* can stimulate angiogenesis by repressing semaphorin 5B in endothelial cells, and it regulates angiogenesis through the circ-SHKBP1/miR-544a/FOXP1 pathway (17). While *ETS1* enables angiogenesis in several ways (18).

3. 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 hitherto remained unexplored, as fibroblast phenotypes are considered highly context-dependent and unstable in culture. In our samples, 857 fibroblasts were detected (**Fig. 2A**). Subclustering revealed six distinct subtypes: clusters 1 and 4 (clusters 1, *PCOLCE2*$^+$ and cluster 4, *CXCL6*$^+$) were strongly enriched in tumor tissue, while cluster 2 (*PLAT*$^+$) was totally enriched in non-malignant tissue. And other clusters (cluster 3, *ARHGDIB*$^+$; cluster 5, *MYH11*$^+$; cluster 6, *STMN2*$^+$), however, were derived from a mixture of tumor and nonmalignant tissue, but mostly enriched in non-malignant tissue (**Fig. 2B**). Remarkably, fibroblasts (*CD34*$^+$ and *KLF4*$^+$) were generally enriched in tumor tissue. The marker genes of CAFs, PCOLCE2 and CXCL6, were confirmed to be significantly up-regulated in tumor tissues by the bulk RNA-seq data from TCGA data (**Fig. 2C**).

The first role of CAFs is promoting the proliferation of cancer cells. The clusters 1 and 4 were enriched in Wnt and KRAS signaling, which have a close relationship with tumor proliferation (19) (**Fig. 2D& 2E**). SCENIC analysis showed that *KLF12*, which promotes CRC growth, was highly expressed in cluster 1(20) (**Fig. 2G**, **Fig. S3**). And the cluster 4 showed high expression in TGFβ and Wnt signaling, which are also related to cancer invasion and metastasis(21). Another remarkable characteristic of CAFs is extracellular matrix (ECM) remodeling, and collagens participate in tumor progression(22). We found that various collagens were highly expressed in fibroblasts, and different clusters seemed to have different expression inclinations (**Fig. 2F**). In addition to collagens, many other matrix components, such as fibronectin, periostin, hyaluronan and proteoglycan (marker genes *PRG4, POSTN, HAS2* and *FN1*, respectively) were also up-regulated by CAFs (23, 24). The corresponding genes were highly expressed in CAFs (**Fig. 2G**). CAFs also influenced the drug resistance of tumors, and CXCR4 expression predicts patient outcome and recurrence patterns after hepatic resection for colorectal liver metastases. *CXCL12* can mediate drug resistance by combining with CXCR4 expressed in cancer cells(25). Additionally, *FAP*$^+$ CAF can mediate immune suppression by *CXCL12*(26).

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

We detected 1100 B lymphocyte cells. B cells are the most prevalent stromal cell type (**Fig. 3A**). Clustering revealed five clusters, of these, two clusters (clusters 4 and 5; *REG3A*$^+$ and *MS4A1*$^+$, respectively) were mostly tumor enriched, and the other three clusters (the cluster 1, *IGHM*$^+$; cluster 2, *IGLL5*$^+$ and cluster 3, *IGHGP*$^+$, respectively) were composed of both tumor and non-malignant cells (**Fig. 3B**). Moreover, the clusters 1, 2 and 3 showed plasma properties and were not grouped into distinct clusters. Although all of them expressed immunoglobulin A, the tumor-derived cells had higher IgM expression, while

nonmalignant-derived cells preferred higher IgA expression. Cluster 4 significantly expressed a large number of innate immunity-related genes, such as *REG3A, PRSS2, ITLN2*, and *LYZ*, but had negative CD5 expression. Therefore, we identified them as B1 cell-like cells. Cluster 5 had high expression of *MS4A1, LTB,* and *HLA-DRB1*, which were characteristic of follicular B cells (**Fig. 3C**).

Pathway analyses showed that peroxisome signaling was highly expressed in tumor cells (**Fig. 3D**). Peroxisome signaling was closely related to CRC risk (27) and has a positive correlation with lymph node metastasis and poor prognosis of CRC(28). The results also showed that the oxidative phosphorylation pathway was enriched in nonmalignant-derived cells, which might attributed to the impact of colitis(29). SCENIC analysis failed to identify differences between nonmalignant tissue-derived and tumor-derived plasma cells (**Fig. S4)**). There was also no difference in transcript numbers in tumor-associated versus non-malignant tissue-associated plasma cells, but the plasma cells showed a higher transcription trend than clusters 4 and 5 (**Fig. 3A**).

5. Different derived myeloid cells showed diverse expressing properties.

The 365 myeloid cells clustered into four subsets, which were not completely separated (**Fig. 4A**). One cluster corresponded to macrophages (cluster 1, *CAPG*$^{+}$), and one cluster to monocytes (cluster 2, *CXCL2*$^{+}$). There was also one dendritic cell cluster (cluster 3, *IDO1*$^{+}$) and a granulocyte cluster (cluster 4, *CC20*$^{+}$) (**Fig. 4B**). The dendritic cells and granulocytes were typically more abundant in tumors than in non-malignant tissues, while macrophages were more abundant in non-malignant tissues. The monocytes were detected at similar numbers in both tissues.

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 the different derived cells (**Fig. S5**). Furthermore, we analyzed the pathways of the differently derived cells and revealed tumor-associated 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 preferred the oxidative phosphorylation and biomass production pathways (pathways associated with phosphorylation and protein secretion) (**Fig. 4C**).

6. Several typical species of T cells identified in multiple analysis.

With 318 cells detected, T cells were mainly divided into three clusters, which were designated as naive T cells (cluster 1, *YPEL5*$^{+}$), cytotoxic T cells (cluster 2, *GZMA*$^{+}$) and natural killer T cells (cluster 3, *PIGR*$^{+}$) (**Fig. 5A&5B**). In cluster 1, the significantly expressed genes, such as *YPEL5* and *GPR18*, were closely related to proliferation and cell differentiation(30). 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 (**Fig. 5C**). SCENIC analysis also showed that the T cell-specific differentiation-related transform factor Elf-1 was significantly up-regulated (31) (**Fig. 5D& 5E**). In cluster 2, we found that cytotoxic T cell-specific genes, such as *GZMA* and *GNLY*, were highly expressed (32). Additionally, the glycolysis pathway was most highly expressed in cluster 2 among the three clusters, and a related gene,

PKM, was also highly expressed (33). In cluster 3, we detected minor populations of cells expressing higher levels of the immune checkpoint molecule HAVCR, which acts in the tolerance and exhaustion of T cells (34) and is currently targeted in clinical trials of immunotherapy for cancers (35). Its nearby populations of cells had a high expression of MKI67, which encodes proliferation-related proteins(36), and the notable transcription numbers also reflected their high proliferative activity (**Fig. 5A& 5B**).

7. 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), while four clusters almost exclusively from non-malignant tissues (cluster 2, *PI3*$^{+}$; cluster 4, *MUC1*$^{+}$; cluster 5 *CA4*$^{+}$; cluster 6, *HMGB2*$^{+}$, respectively) (**Fig. 6A&6B**). Pathway analysis showed significant differences between the tumor- and nonmalignant-derived tissues (**Fig. 6C**). In particularly, cluster 1 existed typically exhibited malignant properties. High proliferation- and embryonic developmental process-related pathways were highly expressed (Wnt signaling, Myc targets and EMT signaling) (37, 38), and lesions repair-associated pathways, such as the DNA repair pathway, were down-regulated(39) (**Fig. 6D**).

Through SCENIC analysis, we identified that the transcription factors CDX2 and STAT3 were significantly up-regulated in cluster 1, but showed almost no expression in other clusters besides cluster 3 (**Fig. S6**). CDX2 inhibits aggressive phenotypes of colon cancer cells in vitro and in vivo (40) and could be a prognostic marker related to the benefit of adjuvant chemotherapy(41). STAT3 is essential for the transduction of tumor-promoting signals of the IL-6/STAT3 pathway, which is highly activated in CAC (42). Collectively, different clusters exhibited obviously diverse properties even in the same tissues (**Fig. 7E**).

8. Evolutionary trajectory of disease development and internal variation in crucial genes

The complete transcriptome data for a large number of 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(43), therefore, we could use the dynamic information of gene expressions to track the developing progress of disease. We applied transcriptional similarity-based pseudotime analysis(44) to order epithelial cells and indicated their developmental trajectories (**Fig. S7A**). All the 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 (nonmalignant-derived cells), followed by states 2, 3 and 5 (mixed derived cells), and ending with states 1, 6 and 9 (tumor-derived cells) (**Fig. S7A& S7B**). Following this trajectory, the differentially expressed genes were identified, and these genes might play crucial roles in the evolution from colitis to cancer. Thus, we focused on the top three differentially expressed genes, *CD74, CLCA1* and *DPEP1*, in the subsequent analysis (**Fig. S7C**). Different degrees of IHC between normal and tumor tissues for CD74, CLCA1 and DPEP1 from the public

data website of The Human Protein Atlas, confirmed the expression change in these genes along with disease progression (**Fig. S7D**).

Cluster of differentiation 74 (CD74), also known as HLA-DR antigen-associated invariant chain and encoded by the *CD74* gene(45), is a polypeptide involved in the formation and transport of MHC class II protein, which is found on a number of cancer cell types(46, 47). In CRC, stimulation of CD74 by MIF induces a signaling cascade leading to up-regulation of Bcl-2, resulting in significantly increased survival of patients with colon cancer, and the MIF/CD74 axis is a target for novel therapies (48). Our data demonstrated that CD74 was down-regulated as tumorigenesis progressed, which was consistent with previous reports. TCGA data showed that the expression of CD74 was significantly high in normal tissues ($p<0.05$) (**Fig. S7D& S7E**), and the patients with high CD74 expression had better survival ($p=0.0067$) (**Fig. S7F**).

Calcium-activated chloride channel regulator 1 (CLCA1) 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(49), regulation of tissue inflammation in the innate immune response(50) and tumor suppression in CRC(51). *CLCA1* can suppress CRC aggressiveness via inhibition of the Wnt/β-catenin signaling pathway and low expression of CLCA1 predicts a poor prognosis in CRC (52). Our data demonstrated that CLCA1 was upregulated firstly, and downregulated during the late period of the pseudotime analysis. TCGA data showed that there was no significance for the expression of CLCA1 between tumor and normal tissues ($p=0.0715$) (**Fig. S7D& S7E**), and no relevance between CLCA1 expression and the prognosis of patients ($p=0.11$) (**Fig. S7F**).

Dipeptidase 1 (DPEP1), encoded by the *DPEP1* gene, hydrolyzes a wide range of dipeptides and participated in many biological processes, such as the metabolism of glutathione and β-lactam hydrolysis in the kidney(53). In CRC, DPEP1 expression affects 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(54). Our data demonstrated that *DPEP1* was upregulated as tumorigenesis progressed, which was consistent with previous reports. TCGA data showed that the expression of DPEP1 was significantly high in tumor tissues ($p<0.05$) (**Fig. S7D& S7E**), but there was no relevance between DPEP1 expression and the prognosis of patients ($p=0.093$) (**Fig. S7F**).

Discussion

In previous studies, researchers attempted to define cell identity 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 on the basis of a comprehensive and quantitative readout of mRNA (55). Thus, scRNA-seq technology has attracted great attention since its inception, and a large number of applied research results have been published in

just a few years(56). In the context of human cancer, scRNA-seq was used to reveal the intra-tumor heterogeneity and the transcriptional trajectories of malignant transformation (57).

In this study, we analyzed 4,777 single-cell transcriptomes of human colon tumorous and non-malignant tissues from UC-associated colon cancer. Meanwhile, we defined the composition of cancer-associated stromal cells, analyzed the different subgroups of tumor cells and described the notable pathways and transcription factors involved in the disease. Many of the cell types tumor-derived identified by a scRNA-seq approach, including B cells, T cells, endothelial cells, myeloid cells, fibroblasts and epithelial cells (summarized in **Fig. S8**), presented an altered gene expression profile with pro-tumoral properties compared with the non-malignant cells. Of note, we graphed the evolutionary trajectory of tumor development and the pseudotime analysis revealed the cellular composition of CAC and its developmental trajectory, and the tumor microenvironment might play a crucial role in the evolutionary process from an IBD to a CAC. Moreover, we identified the top three differentially expressed genes that have a role in the disease progression, *CD74, CLCA1* and *DPEP1*. These results may represent a promising strategy that identifies novel potential therapeutic targets in the evolution from IBD to CAC and may prevent the development of CRC.

Gene expression changes in the tumorigenesis trajectory suggest directions for the design of therapies(58). For instance, CD74 is abundant in non-malignant cells, but down-regulated in tumor endothelial cells. Additionally, 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 benefit by blocking these conversion processes. Distinctive features of tumor cells may represent vulnerabilities and provide potential entry points for the design of novel therapies.

To our knowledge, this is the first study depicting the cell landscape of UC-associated colon cancer at the single-cell level, which describes the intra-tumor heterogeneity mainly from three different viewpoints: gene expression, pathway enrichments and transcription factor analysis. It provides a more accurate perspective for analyzing the evolutionary progress of UC-associated colon cancer than looking at average calculation (59).

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 lack more patients with CAC to compare the obtained results. Secondly, the methods to classify the cell types are based on distinct and highly expressed genes, coupled with previous reports. An authorized or unified standard should be established. Thirdly, the cohort of the TCGA database was CRC patients, which might have a bias for UC-associated colon cancer. Lastly, a larger and specific cohort will be helpful to yield more accurate and convincing results in our future studies.

This study primarily elucidates the tumor microenvironment composition 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 IBD to CAC and may prevent the development of

CAC. In the future, the researchers should concern whether the associations between the transcriptome of the tumor cells and either of the previously described CRC CMS subtypes exist, and determine the transcriptome of the tumor cells a signature related to ulcerative colitis among more patients.

Abbreviations

CAC: Colitis-associated cancer;

IBDs: Inflammatory bowel diseases;

UC: Ulcerative colitis;

scRNA-seq: Single cell RNA sequencing;

TCGA: The Cancer Genome Atlas;

TFs: Transcription factors;

CRC: Colorectal cancer;

S-CRC: Sporadic CRC;

TME: Tumor microenvironment;

CTCs: Circulating tumor cells;

MSS: Microsatellite stability;

UMIs: Unique molecular identifier;

PCA: Principle component analysis;

tSNE: t-distributed stochastic neighbor embedding;

KEGG: Kyoto Encyclopedia of Genes and Genomes;

IHC: Immunohistochemical staining;

SCENIC: Single-cell regulatory network inference and clustering;

CAFs: Cancer associated fibroblasts;

CD74: Cluster of differentiation 74;

CLCA1: Calcium-activated chloride channel regulator 1;

DPEP1: Dipeptidase 1

Declarations

Acknowledgement

We thank the *CapitalBio Technology* to perform the procedures of single-cell sequencing and help us to analyze the data. At the meantime, the assistance of data a professional copyediting agency, American Journal Experts (AJE) to improve the quality of language in our manuscript.

Authors' contributions

WQ, WZ, JKW and WS made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data and was a major contributor in writing the manuscript. ZZ, ZW, ZMM, YYJ and SZL have been involved in drafting part of the manuscript or revising it critically for important intellectual content. WS have given final approval of the version to be published. All authors read and approved the final manuscript.

Funding

This work was supported by National Key Research and Development Program of China (2017YFC1308805), Industry-University-Research Innovation Fund in Ministry of Education of the People's Republic of China (2018A01013) and National Natural Science Foundation of China (Grant No. 81871962).

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Protocols involving human samples collection were reviewed and approved by the Ethics Committee of the Peking University People's Hospital. And a written consent was signed by the patient before the surgery.

Patient consent for publication

Identifying information, including names, initials, date of birth or hospital numbers, images or statements are not included in the manuscript.

Publication of clinical datasets

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC *et al.* Colorectal cancer statistics, 2020. CA Cancer J Clin 2020. doi: [10.3322/caac.21601]

2. Eaden JA, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut 2001; 48: 526-535. doi:

3. Ullman T, Odze R, Farraye FA. Diagnosis and management of dysplasia in patients with ulcerative colitis and Crohn's disease of the colon. Inflammatory bowel diseases 2009; 15: 630-638. doi: [10.1002/ibd.20766]

4. Hao XP, Frayling IM, Sgouros JG, Du MQ, Willcocks TC, Talbot IC *et al.* The spectrum of p53 mutations in colorectal adenomas differs from that in colorectal carcinomas. Gut 2002; 50: 834-839. doi:

5. Tarmin L, Yin J, Harpaz N, Kozam M, Noordzij J, Antonio LB *et al.* Adenomatous polyposis coli gene mutations in ulcerative colitis-associated dysplasias and cancers versus sporadic colon neoplasms. Cancer research 1995; 55: 2035-2038. doi:

6. Burmer GC, Levine DS, Kulander BG, Haggitt RC, Rubin CE, Rabinovitch PS. c-Ki-ras mutations in chronic ulcerative colitis and sporadic colon carcinoma. Gastroenterology 1990; 99: 416-420. doi:

7. Ungaro R, Mehandru S, Allen PB, Peyrin-Biroulet L, Colombel J-F. Ulcerative colitis. Lancet (London, England) 2017; 389: 1756-1770. doi: [10.1016/S0140-6736(16)32126-2]

8. Castaño-Milla C, Chaparro M, Gisbert JP. Systematic review with meta-analysis: the declining risk of colorectal cancer in ulcerative colitis. Alimentary pharmacology & therapeutics 2014; 39: 645-659. doi:

9. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature 2013; 501: 328-337. doi: [10.1038/nature12624]

10. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458: 719-724. doi: [10.1038/nature07943]

11. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature 2017; 541: 331-338. doi: [10.1038/nature21350]

12. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (New York, N.Y.) 2016; 352: 189-196. doi: [10.1126/science.aad0501]

13. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proceedings of the National Academy of Sciences of the United States of America 2013; 110: 21083-21088. doi: [10.1073/pnas.1320659110]

14. Kim K-T, Lee HW, Lee H-O, Kim SC, Seo YJ, Chung W *et al.* Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome biology 2015; 16: 127. doi: [10.1186/s13059-015-0692-3]

15. Sparmann A, Bar-Sagi D. Ras-induced interleukin-8 expression plays a critical role in tumor growth and angiogenesis. Cancer cell 2004; 6: 447-458. doi:

16. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G *et al.* SCENIC: single-cell regulatory network inference and clustering. Nature methods 2017; 14: 1083-1086. doi: [10.1038/nmeth.4463]

17. He Q, Zhao L, Liu Y, Liu X, Zheng J, Yu H *et al.* circ-SHKBP1 Regulates the Angiogenesis of U87 Glioma-Exposed Endothelial Cells through miR-544a/FOXP1 and miR-379/FOXP2 Pathways. Molecular therapy. Nucleic acids 2018; 10: 331-348. doi: [10.1016/j.omtn.2017.12.014]

18. Chen J, Fu Y, Day DS, Sun Y, Wang S, Liang X *et al.* VEGF amplifies transcription through ETS1 acetylation to enable angiogenesis. Nature communications 2017; 8: 383. doi: [10.1038/s41467-017-00405-x]

19. Bahrami A, Amerizadeh F, ShahidSales S, Khazaei M, Ghayour-Mobarhan M, Sadeghnia HR *et al.* Therapeutic Potential of Targeting Wnt/β-Catenin Pathway in Treatment of Colorectal Cancer: Rational and Progress. Journal of cellular biochemistry 2017; 118: 1979-1983. doi: [10.1002/jcb.25903]

20. Kim S-H, Park Y-Y, Cho S-N, Margalit O, Wang D, DuBois RN. Krüppel-Like Factor 12 Promotes Colorectal Cancer Growth through Early Growth Response Protein 1. PloS one 2016; 11: e0159899. doi: [10.1371/journal.pone.0159899]

21. O'Connell JT, Sugimoto H, Cooke VG, MacDonald BA, Mehta AI, LeBleu VS *et al.* VEGF-A and Tenascin-C produced by S100A4+ stromal cells are important for metastatic colonization. Proceedings of the National Academy of Sciences of the United States of America 2011; 108: 16002-16007. doi: [10.1073/pnas.1109493108]

22. Fang M, Yuan J, Peng C, Li Y. Collagen as a double-edged sword in tumor progression. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 2014; 35: 2871-2882. doi: [10.1007/s13277-013-1511-7]

23. Cirri P, Chiarugi P. Cancer associated fibroblasts: the dark side of the coin. American journal of cancer research 2011; 1: 482-497. doi:

24. Silver DJ, Siebzehnrubl FA, Schildts MJ, Yachnis AT, Smith GM, Smith AA *et al.* Chondroitin sulfate proteoglycans potently inhibit invasion and serve as a central organizer of the brain tumor microenvironment. The Journal of neuroscience : the official journal of the Society for Neuroscience 2013; 33: 15603-15617. doi: [10.1523/JNEUROSCI.3004-12.2013]

25. Yopp AC, Shia J, Butte JM, Allen PJ, Fong Y, Jarnagin WR *et al.* CXCR4 expression predicts patient outcome and recurrence patterns after hepatic resection for colorectal liver metastases. Annals of surgical oncology 2012; 19 Suppl 3: S339-S346. doi: [10.1245/s10434-011-1774-4]

26. Fearon DT. The carcinoma-associated fibroblast expressing fibroblast activation protein and escape from immune surveillance. Cancer immunology research 2014; 2: 187-193. doi: [10.1158/2326-6066.CIR-14-0002]

27. Voutsadakis IA. Peroxisome proliferator-activated receptor gamma (PPARgamma) and colorectal carcinogenesis. Journal of cancer research and clinical oncology 2007; 133: 917-928. doi:

28. Yun S-H, Roh M-S, Jeong J-S, Park J-I. Peroxisome proliferator-activated receptor γ coactivator-1α is a predictor of lymph node metastasis and poor prognosis in human colorectal cancer. Annals of diagnostic pathology 2018; 33: 11-16. doi: [10.1016/j.anndiagpath.2017.11.007]

29. Almatov KT, Agzamov K, Rakhimov MM. [Oxidative phosphorylation in mitochondria from digestive organs in chronic allergic colitis]. Voprosy meditsinskoi khimii 1982; 28: 44-49. doi:

30. Lampert F, Stafa D, Goga A, Soste MV, Gilberto S, Olieric N *et al.* The multi-subunit GID/CTLH E3 ubiquitin ligase promotes cell proliferation and targets the transcription factor Hbp1 for degradation. eLife 2018; 7. doi: [10.7554/eLife.35528]

31. Wang CY, Bassuk AG, Boise LH, Thompson CB, Bravo R, Leiden JM. Activation of the granulocyte-macrophage colony-stimulating factor promoter in T cells requires cooperative binding of Elf-1 and AP-1 transcription factors. Molecular and cellular biology 1994; 14: 1153-1159. doi:

32. Stenger S, Hanson DA, Teitelbaum R, Dewan P, Niazi KR, Froelich CJ *et al.* An antimicrobial activity of cytolytic T cells mediated by granulysin. Science (New York, N.Y.) 1998; 282: 121-125. doi:

33. Kuranaga Y, Sugito N, Shinohara H, Tsujino T, Taniguchi K, Komura K *et al.* SRSF3, a Splicer of the PKM Gene, Regulates Cell Growth and Maintenance of Cancer-Specific Energy Metabolism in Colon Cancer Cells. International journal of molecular sciences 2018; 19. doi: [10.3390/ijms19103012]

34. Huang Y-H, Zhu C, Kondo Y, Anderson AC, Gandhi A, Russell A *et al.* CEACAM1 regulates TIM-3-mediated tolerance and exhaustion. Nature 2015; 517: 386-390. doi: [10.1038/nature13848]

35. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nature reviews. Cancer 2012; 12: 252-264. doi: [10.1038/nrc3239]

36. Miele A, Medina R, van Wijnen AJ, Stein GS, Stein JL. The interactome of the histone gene regulatory factor HiNF-P suggests novel cell cycle related roles in transcriptional control and RNA processing. Journal of cellular biochemistry 2007; 102: 136-148. doi:

37. Campisi J, Gray HE, Pardee AB, Dean M, Sonenshein GE. Cell-cycle control of c-myc but not c-ras expression is lost following chemical transformation. Cell 1984; 36: 241-247. doi:

38. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. The Journal of clinical investigation 2009; 119: 1420-1428. doi: [10.1172/JCI39104]

39. Hoeijmakers JH. Genome maintenance mechanisms for preventing cancer. Nature 2001; 411: 366-374. doi:

40. Zheng J, He S, Qi J, Wang X, Yu J, Wu Y *et al.* Targeted CDX2 expression inhibits aggressive phenotypes of colon cancer cells in vitro and in vivo. International journal of oncology 2017; 51: 478-488. doi: [10.3892/ijo.2017.4040]

41. Dalerba P, Sahoo D, Paik S, Guo X, Yothers G, Song N *et al.* CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. The New England journal of medicine 2016; 374: 211-222. doi: [10.1056/NEJMoa1506597]

42. Li Y, de Haar C, Chen M, Deuring J, Gerrits MM, Smits R *et al.* Disease-related expression of the IL6/STAT3/SOCS3 signalling pathway in ulcerative colitis and ulcerative colitis-related carcinogenesis. Gut 2010; 59: 227-235. doi: [10.1136/gut.2009.184176]

43. Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012; 481: 306-313. doi: [10.1038/nature10762]

44. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology 2014; 32: 381-386. doi: [10.1038/nbt.2859]

45. Claesson L, Larhammar D, Rask L, Peterson PA. cDNA clone for the human invariant gamma chain of class II histocompatibility antigens and its implications for the protein structure. Proceedings of the National Academy of Sciences of the United States of America 1983; 80: 7395-7399. doi:

46. Ghoochani A, Schwarz MA, Yakubov E, Engelhorn T, Doerfler A, Buchfelder M *et al.* MIF-CD74 signaling impedes microglial M1 polarization and facilitates brain tumorigenesis. Oncogene 2016; 35: 6246-6261. doi: [10.1038/onc.2016.160]

47. Murayama T, Nakaoku T, Enari M, Nishimura T, Tominaga K, Nakata A *et al.* Oncogenic Fusion Gene CD74-NRG1 Confers Cancer Stem Cell-like Properties in Lung Cancer through a IGF2 Autocrine/Paracrine Circuit. Cancer research 2016; 76: 974-983. doi: [10.1158/0008-5472.CAN-15-2135]

48. Bozzi F, Mogavero A, Varinelli L, Belfiore A, Manenti G, Caccia C *et al.* MIF/CD74 axis is a target for novel therapies in colon carcinomatosis. Journal of experimental & clinical cancer research : CR 2017; 36: 16. doi: [10.1186/s13046-016-0475-z]

49. Gruber AD, Elble RC, Ji HL, Schreur KD, Fuller CM, Pauli BU. 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: 200-214. doi:

50. Toda M, Tulic MK, Levitt RC, Hamid Q. A calcium-activated chloride channel (HCLCA1) is strongly related to IL-9 expression and mucus production in bronchial epithelium of patients with asthma. The Journal of allergy and clinical immunology 2002; 109: 246-250. doi:

51. Bustin SA, Li SR, Dorudi S. Expression of the Ca2+-activated chloride channel genes CLCA1 and CLCA2 is downregulated in human colorectal cancer. DNA and cell biology 2001; 20: 331-338. doi:

52. Yang B, Cao L, Liu J, Xu Y, Milne G, Chan W *et al.* Low expression of chloride channel accessory 1 predicts a poor prognosis in colorectal cancer. Cancer 2015; 121: 1570-1580. doi: [10.1002/cncr.29235]

53. Nakagawa H, Inazawa J, Inoue K, Misawa S, Kashima K, Adachi H *et al.* Assignment of the human renal dipeptidase gene (DPEP1) to band q24 of chromosome 16. Cytogenetics and cell genetics 1992; 59: 258-260. doi:

54. McIver CM, Lloyd JM, Hewett PJ, Hardingham JE. Dipeptidase 1: a candidate tumor-specific molecular marker in colorectal carcinoma. Cancer letters 2004; 209: 67-74. doi:

55. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular cell 2017; 65. doi: [10.1016/j.molcel.2017.01.023]

56. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome medicine 2017; 9: 75. doi: [10.1186/s13073-017-0467-4]

57. Young MD, Mitchell TJ, Vieira Braga FA, Tran MGB, Stewart BJ, Ferdinand JR *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science (New York, N.Y.) 2018; 361: 594-599. doi: [10.1126/science.aat1699]

58. Song M, Willett WC, Hu FB, Spiegelman D, Must A, Wu K *et al.* Trajectory of body shape across the lifespan and cancer risk. International journal of cancer 2016; 138: 2383-2395. doi: [10.1002/ijc.29981]

59. Hirsch D, Wangsa D, Zhu YJ, Hu Y, Edelman DC, Meltzer PS *et al.* Dynamics of Genome Alterations in Crohn's Disease-Associated Colorectal Carcinogenesis. Clinical cancer research : an official journal of the American Association for Cancer Research 2018; 24: 4997-5011. doi: [10.1158/1078-0432.CCR-18-0630]

Figures

Fig. 2

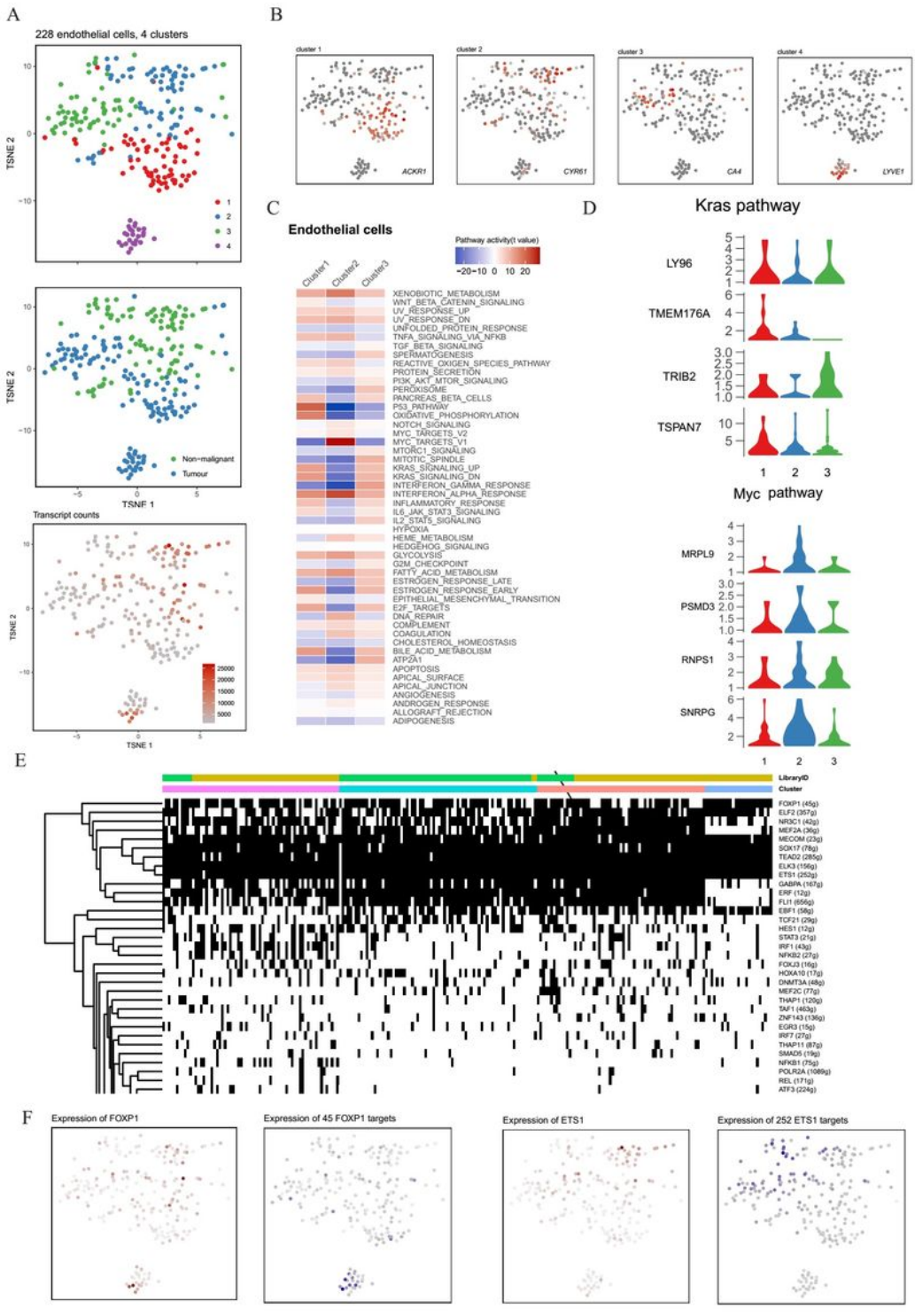

Figure 1

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, (clusters 1, ACKR1+; cluster 2, CYR61+; clusters 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 downregulated in cluster 2, but upregulated in clusters 1 and 3, while the Myc target pathway showed contrasting results. D. Violin plots showing the expression distribution of selected genes involved in the KRAS and Myc pathways. The genes of KRAS signaling pathway were significantly downregulated in cluster 2, but upregulated in clusters 1 and 3, while the genes of Myc target pathway showed contrasting results. 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.

Fig.3

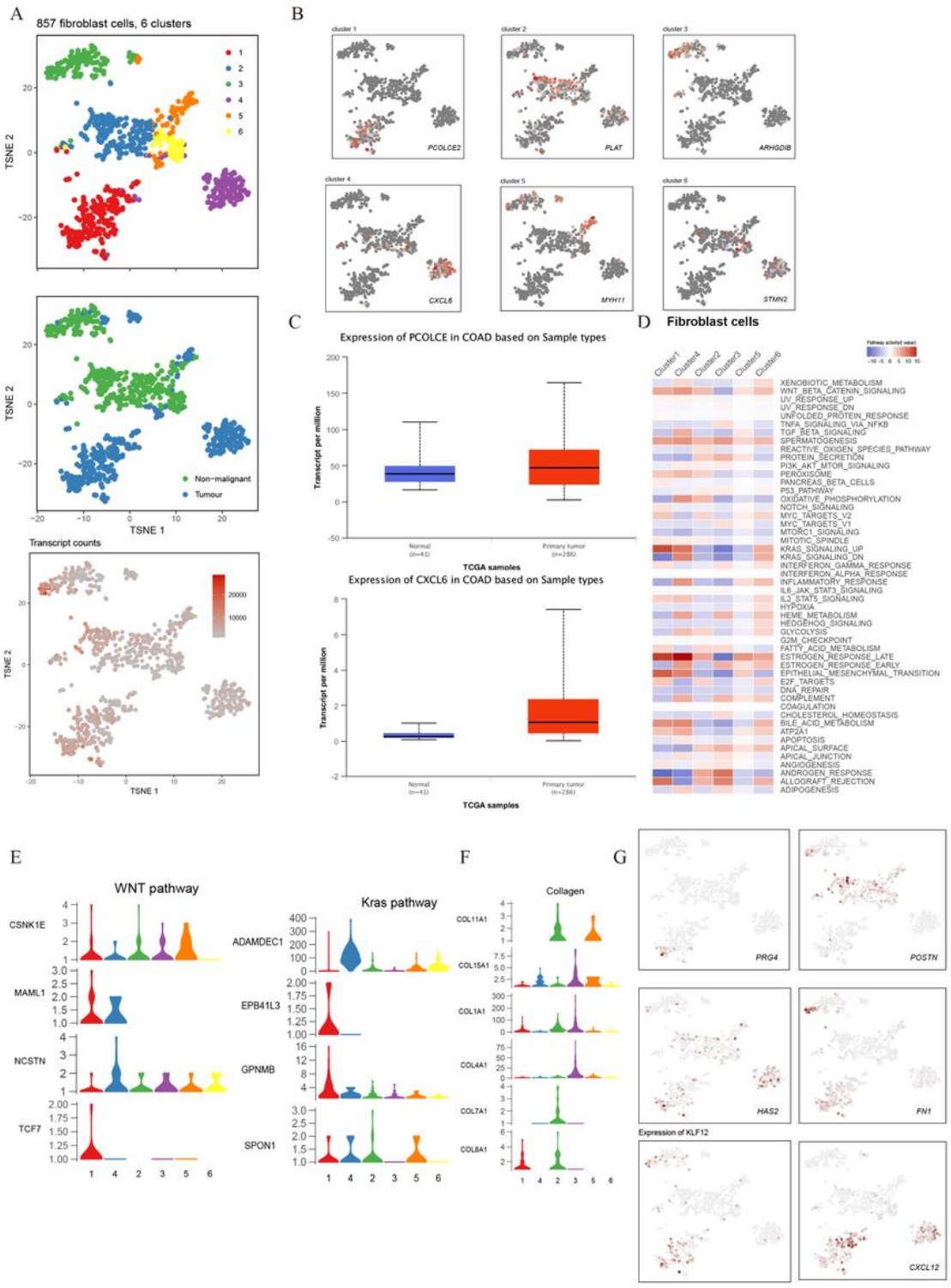

Figure 2

Cancer-associated fibroblasts play 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 (clusters 1, PCOLCE2+; cluster 2, PLAT+; cluster 3, ARHGDIB+; cluster 4, CXCL6+; cluster 5, MYH11+; cluster 6; STMN2+). C. The CAFs marker genes, PCOLCE2 and CXCL6 were confirmed to be significantly

upregulated 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 showing 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, FAP), and transcription factor KLF12 and its target genes, corresponding to the degree of expression

Fig.4

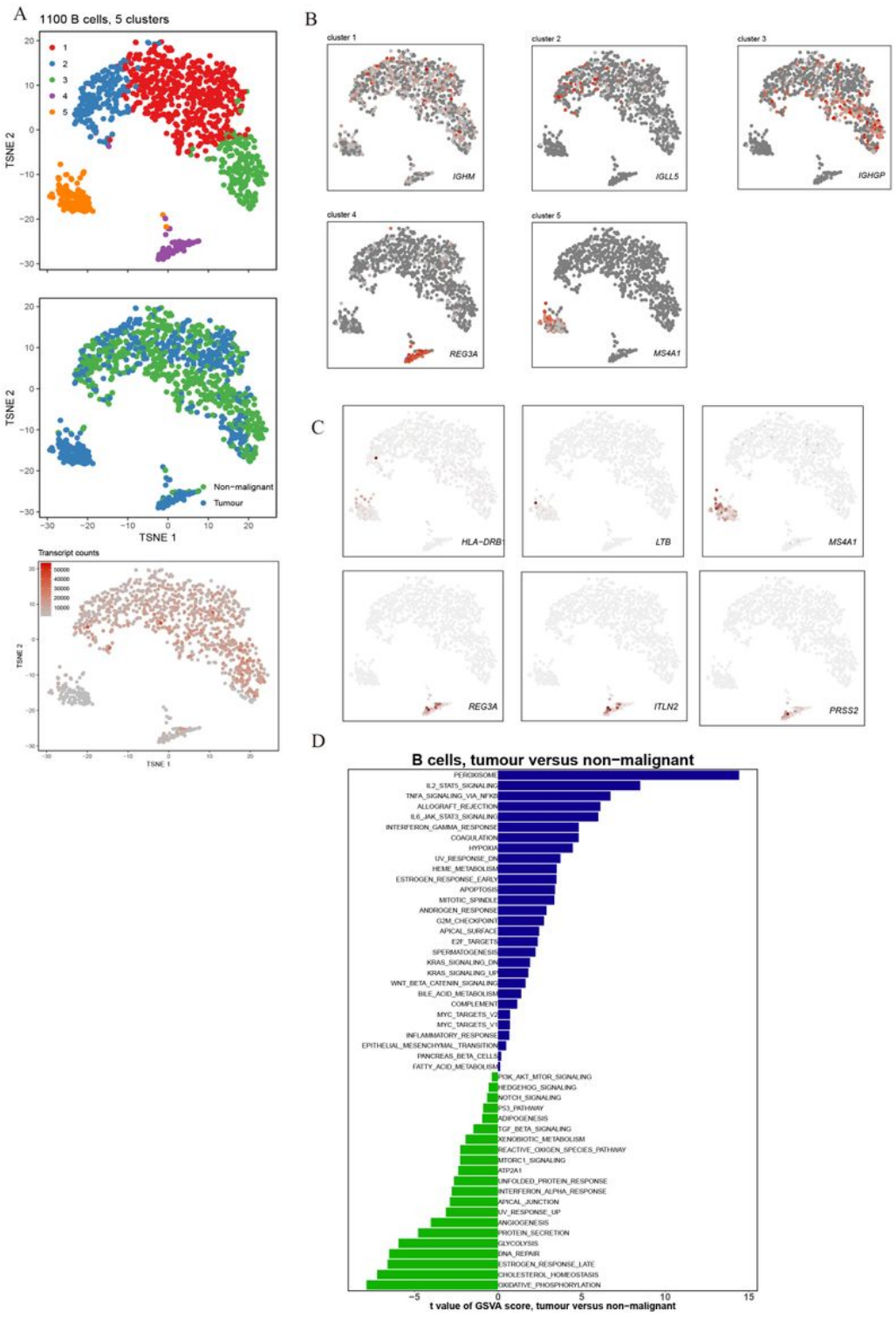

Figure 3

CRC-related pathways are enriched in tumor-derived B cells. A. tSNE plot of 1100 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.

Fig.5

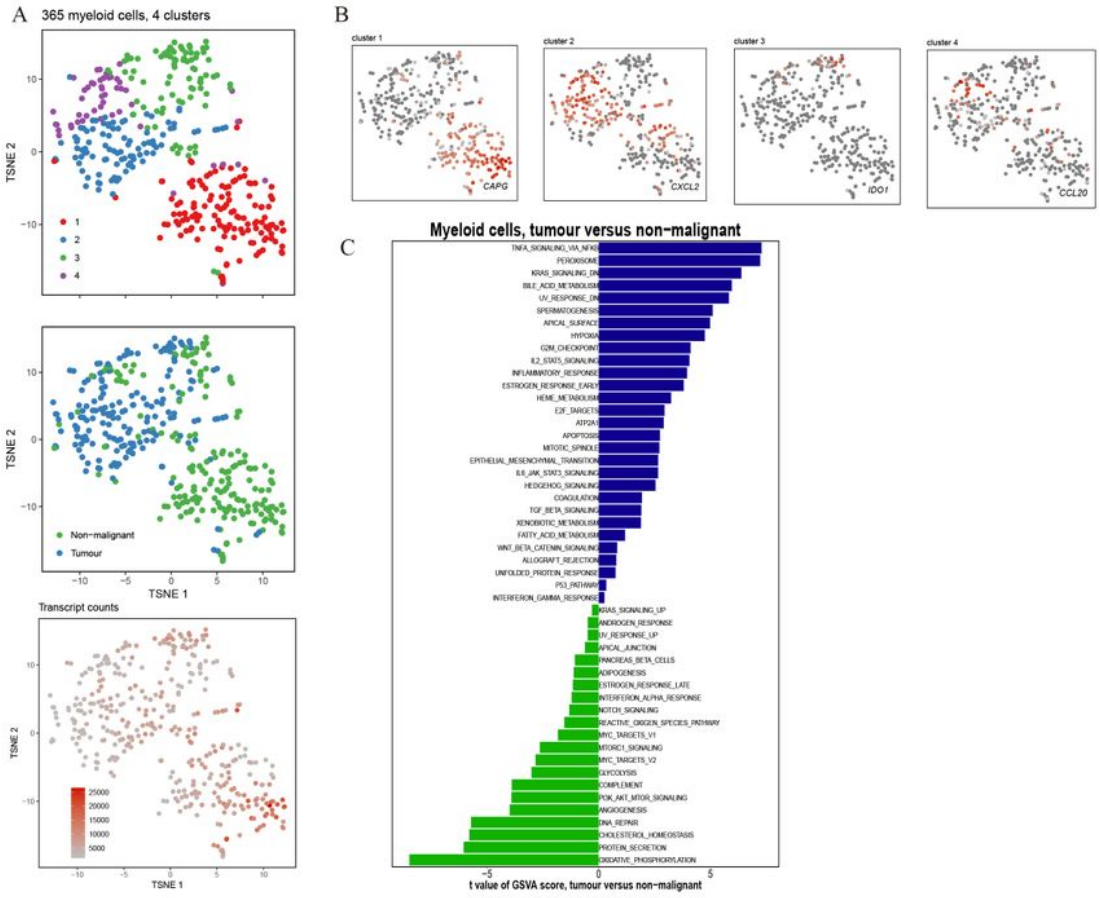

Figure 4

Different derived myeloid cells showed diverse expression properties. A. tSNE plot of 365 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, CC20+). 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).

Fig.6

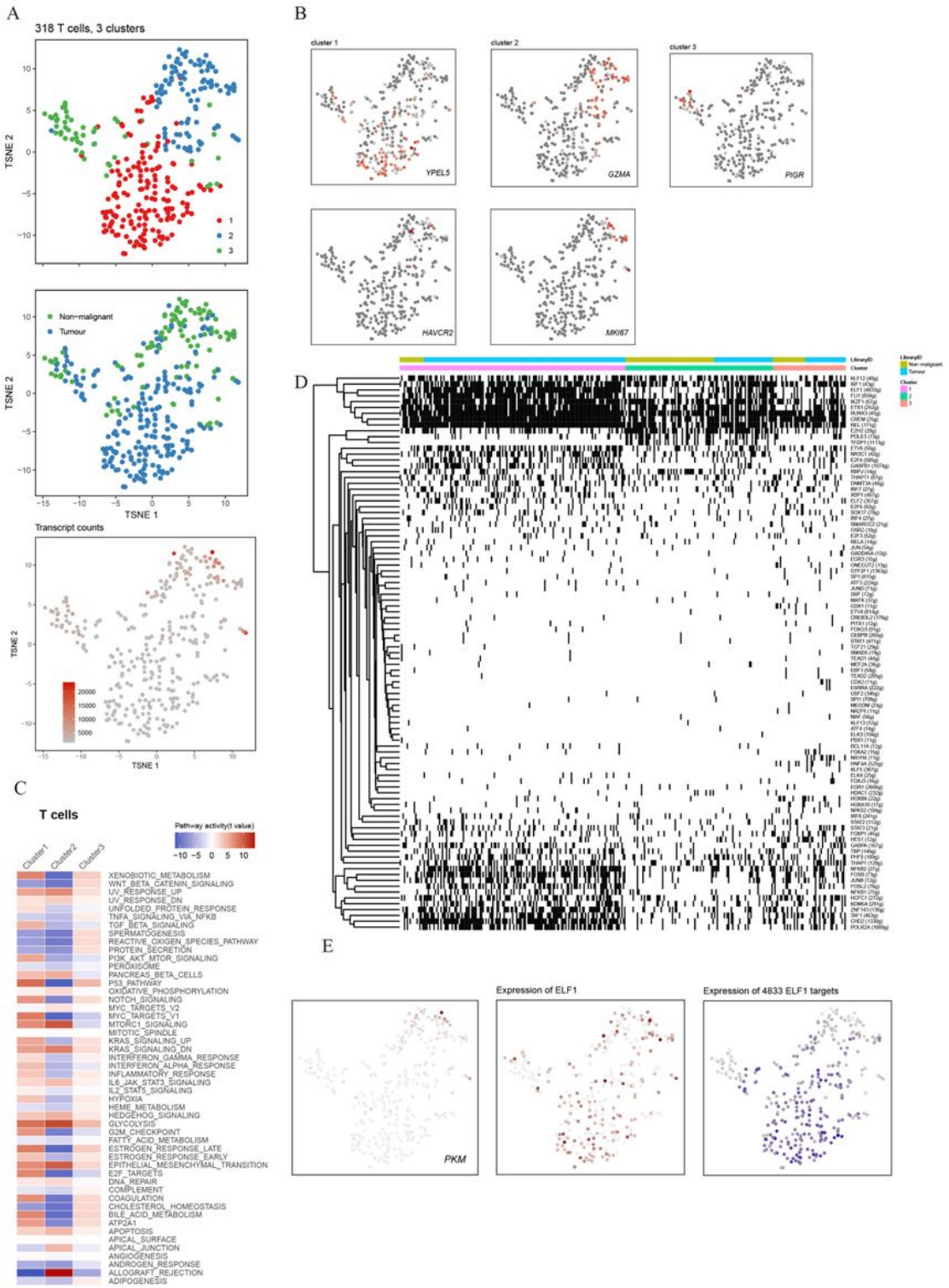

Figure 5

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. 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 upregulated.

Fig.7

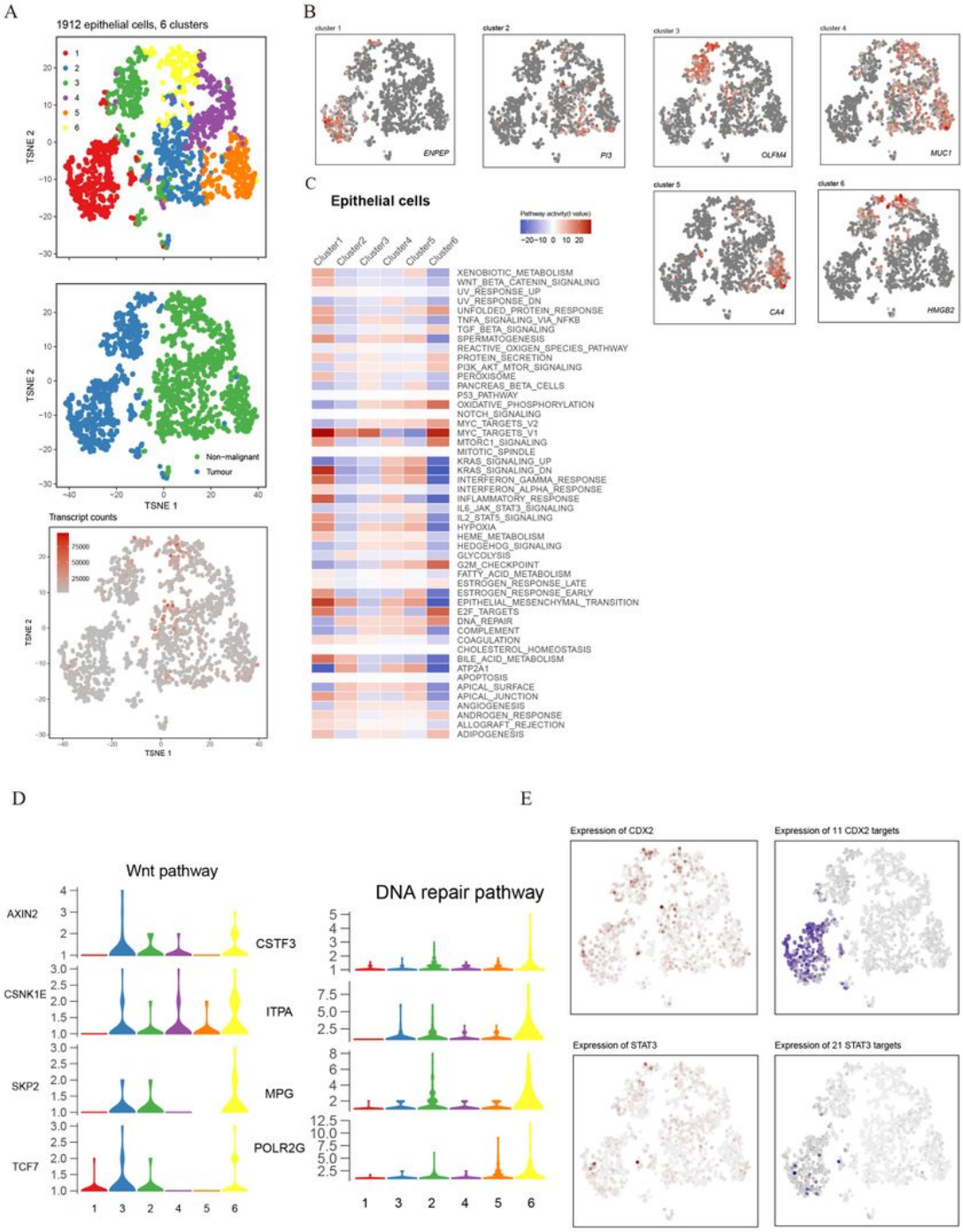

Figure 6

Heterogeneity of epithelial cells was demonstrated 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 (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 nonmalignant-derived tissues. In particularly, cluster 1 and 3 existed typically exhibited malignant properties. D. Violin plots showing 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 epithelial-mesenchymal transition signaling), and lesions repair-associated pathways, such as the DNA repair pathway, were downregulated. E. Exhibition of the involved transcription factors (CDX2, STAT3) and their target genes, corresponding to the degree of expression.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Legendforsupplementarymaterialsrevised2020530.docx
- SupplementaryMaterialsProtocolsofscRNAseq.docx
- TableS1Supplementarytablenew.docx
- FigureS8TME.tif
- FigureS7continue.jpg
- FigureS7.jpg
- Figure.S6.jpg
- FigureS5.jpg
- Figure.S4.jpg
- Figure.S3.jpg
- FigureS2.jpg
- FigureS1.tif