ARTICLES
https://doi.org/10.1038/s41588-020-0636-z

nature genetics

Check for updates

Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer

Hae-Ock Lee[1,2,25], Yourae Hong[1,3,25], Hakki Emre Etlioglu[4,25], Yong Beom Cho[3,5,25], Valentina Pomella[4], Ben Van den Bosch[4], Jasper Vanhecke[4], Sara Verbandt[4], Hyekyung Hong[5], Jae-Woong Min[1], Nayoung Kim[1,2], Hye Hyeon Eum[1,2], Junbin Qian[6,7], Bram Boeckx[6,7], Diether Lambrechts[6,7], Petros Tsantoulis[8,9,10], Gert De Hertogh[11,12], Woosung Chung[1], Taeseob Lee[1,13], Minae An[1,3], Hyun-Tae Shin[1], Je-Gun Joung[1], Min-Hyeok Jung[14], Gunhwan Ko[15], Pratyaksha Wirapati[16], Seok Hyung Kim[17], Hee Cheol Kim[5], Seong Hyeon Yun[5], Iain Bee Huat Tan[18,19,20], Bobby Ranjan[21], Woo Yong Lee[5], Tae-You Kim[22], Jung Kyoon Choi[23], Young-Joon Kim[14,24], Shyam Prabhakar[21], Sabine Tejpar[4,26] ✉ and Woong-Yang Park[1,2,3,26] ✉

Immunotherapy for metastatic colorectal cancer is effective only for mismatch repair-deficient tumors with high microsatellite instability that demonstrate immune infiltration, suggesting that tumor cells can determine their immune microenvironment. To understand this cross-talk, we analyzed the transcriptome of 91,103 unsorted single cells from 23 Korean and 6 Belgian patients. Cancer cells displayed transcriptional features reminiscent of normal differentiation programs, and genetic alterations that apparently fostered immunosuppressive microenvironments directed by regulatory T cells, myofibroblasts and myeloid cells. Intercellular network reconstruction supported the association between cancer cell signatures and specific stromal or immune cell populations. Our collective view of the cellular landscape and intercellular interactions in colorectal cancer provide mechanistic information for the design of efficient immuno-oncology treatment strategies.

Extensive multiomics studies have revealed the molecular landscape of colorectal cancer (CRC), elucidating key genetic events and enabling the classification of patients with CRC according to consensus molecular subtypes (CMS). The CMS classification reflects both the tumor and its associated microenvironment, and reduces the molecular complexity to immune, epithelial or mesenchymal phenotypes[1,2]. Further characterization of tumor complexity using single-cell transcriptome analysis would help in dissecting the molecular landscape by cellular component and offer a deeper understanding of tumor heterogeneity and the microenvironment[3]. For example, cancer-associated fibroblasts, which are the hallmarks of the mesenchymal phenotype CMS4, are associated with poor patient survival[4]. Moreover, T-cell receptor tracking and subtype analysis revealed tumor-specific T-cell development and migration patterns[5]. Since many current anticancer therapies target nontumor components, such as the extracellular matrix, immune system and vascular system[6], understanding cellular components and how their dynamic interactions shape the tumor landscape is particularly important. In this study, we used massively parallel single-cell RNA sequencing (scRNA-seq) and revealed the diversity of, and the dynamic relationships between, cellular components determining the molecular subtypes of CRC.

[1]Samsung Genome Institute, Samsung Medical Center, Seoul, Korea. [2]Department of Molecular Cell Biology, Sungkyunkwan University School of Medicine, Suwon, Korea. [3]Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Korea. [4]Molecular Digestive Oncology, Department of Oncology, Katholieke Universiteit Leuven, Leuven, Belgium. [5]Department of Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea. [6]Center for Cancer Biology, VIB, Leuven, Belgium. [7]Laboratory for Translational Genetics, Department of Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium. [8]Centre d'Oncologie, Hôpitaux Universitaires de Genève, Geneva, Switzerland. [9]Service d'Oncologie, Hôpitaux Universitaires de Genève, Geneva, Switzerland. [10]Université de Genève, Geneva, Switzerland. [11]Department of Pathology, University Hospitals Leuven, Leuven, Belgium. [12]Department of Imaging and Pathology, Leuven, Belgium. [13]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Korea. [14]Department of Biochemistry, College of Life Science and Technology, Yonsei University, Seoul, Korea. [15]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea. [16]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [17]Department of Pathology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea. [18]Division of Medical Oncology, National Cancer Center, Singapore, Singapore. [19]Agency for Science, Technology and Research (A*STAR), Genome Institute of Singapore, Singapore, Singapore. [20]Duke-National University of Singapore Medical School, Singapore, Singapore. [21]Systems Biology and Data Analytics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [22]Cancer Research Institute, Seoul National University College of Medicine, Seoul, Korea. [23]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea. [24]Department of Integrated OMICS for Biomedical Science, Yonsei University Graduate School, Seoul, Korea. [25]These authors contributed equally: Hae-Ock Lee, Yourae Hong, Hakki Emre Etlioglu, Yong Beom Cho. [26]These authors jointly supervised this work: Sabine Tejpar, Woong-Yang Park. ✉e-mail: sabine.tejpar@uzleuven.be; woongyang.park@samsung.com

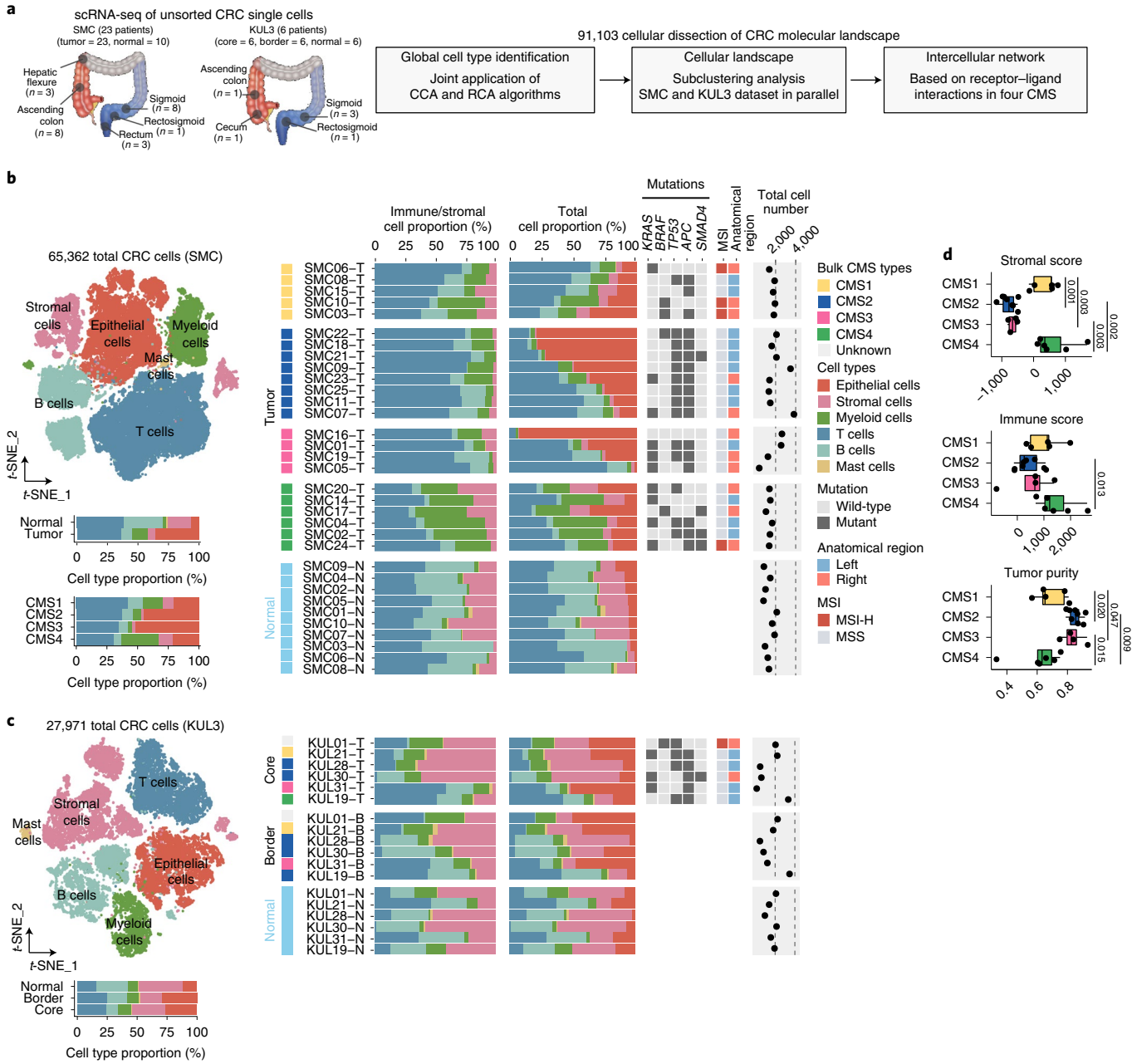

Fig. 1 | Cell type identification in CRC using joint application of RCA and CCA. a, Workflow of sample collection and single-cell transcriptome analysis from Korean and Belgian patients with CRC. **b**,**c**, *t*-distributed stochastic neighbor embedding (*t*-SNE) plot of 65,362 cells in the SMC dataset (**b**) or 27,971 cells in the KUL3 dataset (**c**) identified by joint application of RCA and CCA and color-coded by global cell type (left top). Proportions of the global cell types in CRC tissue and normal mucosa on average (left bottom) or in individual samples (right) are shown. The KUL3 dataset contains core and border samples for tumors (**c**). Samples are grouped by tissue CMS class and aligned with clinical data and selected mutations. Mutations were detected by whole-genome sequencing (**b**) or genotyping (**c**). **d**, Stromal score, immune score and tumor purity of 23 bulk RNA-seq data (CMS1, 5 samples; CMS2, 8 samples; CMS3, 4 samples; CMS4, 6 samples) using the ESTIMATE algorithm. The box plots describe the median and interquartile range (IQR) of each score. The whiskers depict the 1.5× IQR. The *P* values from a two-sided Student's *t*-test are shown. Only significant *P* values are shown.

Results

Global cellular landscape in CRC. To elucidate the cellular landscape of colorectal tumors, we analyzed a total of 91,103 CRC single cells from 23 Korean patients with CRC from the Samsung Medical Center (SMC) and 6 Belgian patients with CRC from Katholieke Universiteit Leuven (this dataset is named KUL3; Fig. 1a–c and Supplementary Table 1). To increase the accuracy of cell type designation, we jointly applied reference component analysis (RCA)[4] and canonical correlation analysis (CCA)[7] before cell type identification (Supplementary Fig. 1a,b). With marker-based annotations, six major cell types were identified in both SMC and KUL3 datasets (Fig. 1b,c and Supplementary Fig. 1b): epithelial cells; fibroblasts/endothelial cells; myeloid cells; T/natural killer (NK)/NK T lymphocytes; B lymphocytes/plasma cells; and mast cells. Marker-based annotation correlated well with the cell types obtained via SingleR v.1.0 (ref. [8]) (Supplementary Fig. 1c,d).

Global cell type annotations for each patient with CRC were categorized according to the molecular subtype of tumor tissues[1]

and supplemented with genetic and clinical information (Fig. 1b,c). Compared to normal tissues, an overall increase in myeloid and decrease in B-cell populations was observed in tumor tissues, suggesting a redirected immune response. For the SMC dataset with larger sample numbers, we found differences in cellular proportions among the CMS subgroups: CMS2 and CMS3 tumors had reduced immune and stromal cell proportions compared with CMS1 and CMS4 tumors (Fig. 1b, left bottom). The cellular proportion determined by scRNA-seq was biased toward an under-representation of tumor cells (Supplementary Fig. 1e)[9]; bulk tissue RNA-seq (Fig. 1d) clearly demonstrated reduced immune cell content in CMS2 and CMS3 tumors, in line with previous studies reporting those subtypes as immune-cold[10]. When comparing the SMC and KUL3 datasets, we observed an enrichment of fibroblasts and myeloid cells in the KUL3 dataset (Supplementary Fig. 2), possibly due to the different sampling processes. Subsequently, the SMC and KUL3 datasets were explored in parallel for cell type-specific subclustering analysis for reciprocal validation.

Single-cell CMS signatures. It has been suggested that human colon cancer cells recapitulate the multilineage differentiation processes of normal colon epithelia[11]. Subclustering analysis of normal epithelial cells in the SMC dataset revealed divergent differentiation states (Fig. 2a and Supplementary Fig. 3a,b), namely stem-like/transit amplifying cells, colonocytes and goblet cells[11,12] (Fig. 2a and Supplementary Fig. 3c). Epithelial cell population diversity was similar in the KUL3 dataset, but additional subpopulations were also identified: *BEST4*^{hi} colonocytes[12] and tuft cells[13] (Supplementary Fig. 3d–f). Trajectory analysis[14] using intestinal reference genes[4,11] further demonstrated differentiation paths originating from stem-like/transit amplifying cells with branching toward colonocytes or goblet cells (Supplementary Fig. 3b,e). Projection of tumor epithelial cells along the normal epithelial cell differentiation trajectory revealed cosegregation of tumor cells with normal stem-like/transit amplifying populations compared to differentiated cell types (Fig. 2b and Supplementary Table 2), suggestive of the regenerative/proliferative potential of tumor cells.

To focus on transcriptional programs categorizing tumor cells, we performed clustering analysis. As shown in Supplementary Fig. 4a,b, most tumor cells formed patient-specific clusters, suggesting highly variable transcriptional states in each patient compared to that in normal cell types[15]. To overcome this individual variation and find common transcriptional programs consolidating and stratifying tumor cells, we utilized the reversed graph embedding technique from Monocle v.2 (ref. [14]) and reconstructed an unsupervised tumor cell trajectory. The trajectory revealed a transcriptional hierarchy, defining nine molecular states (Fig. 2c and Supplementary Fig. 4c) and converging into two discrete transcriptional groups (Fig. 2d, left and Supplementary Fig. 4c). One was highly enriched for transport and Wnt signaling gene expression (*FABP1*, *ID1*, *ID3*, *OLFM4* and *ASCL2*) and the other for secretory and migratory gene expression (*TFF1*, *TFF2*, *TFF3*, *SPINK4* and *AGR2*). These transcriptional features reflected the CMS of CRC, specifically the CMS2 canonical subtype and the CMS3 metabolic subtype, respectively[1] (Fig. 2d and Supplementary Fig. 4d). Intriguingly, similarity scores of single epithelial cells to the reference CMS gene expression programs recapitulated the trajectory, assigning the upper trajectory branch as CMS2-like and the lower branch as predominantly CMS3-like (Fig. 2c, middle) tumor cells. In contrast, projection of tissue CMS classification onto the trajectory assigned the upper right branch to CMS1 and the lower right branch to the CMS4 tissue groups. The discrepant CMS classifications at the cellular and tissue levels reflect the contribution of cancer-associated fibroblasts and immune cell types to the transcriptomes of CMS1 and CMS4 CRCs as suggested previously[4,10,16–18].

Furthermore, sample arrangement along the single-cell CMS (Fig. 2e) revealed patient groups with dominant CMS2-like tumor cells or with mixed CMS1-like and CMS3-like cells. The CMS1-like and CMS3-like cells coexpressed varying levels of the CMS1 immune and CMS3 secretory epithelial feature genes (Supplementary Fig. 4d), suggesting a close relationship between the two subtypes. CMS1-like tumor cells were predominant in many but not all CMS1 tissues, suggesting two different modes of immune cell activation driven by tumor or nontumor components. Notably, the single-cell CMS patterns were associated with the anatomical region[19], microsatellite instability (MSI) and tumor mutation profiles (Fig. 2e and Supplementary Fig. 4e). CMS2-dominant samples frequently harbored *TP53* and *APC* comutations, whereas *KRAS* or *BRAF* mutations were associated with the CMS1/CMS3-enriched samples. *SMAD4* mutations had no association with single-cell CMS, yet three out of four mutations were found in CMS4 tissues.

In contrast to the robust single-cell representation of the CMS1–3 classes, we found very few CMS4-like tumor cells, even from CMS4 tissues (Fig. 2c, middle). When we estimated the cell type influences on each tissue molecular subtype, epithelial tumor cells showed a negative correlation with CMS4 genes, but stromal cells (fibroblasts and endothelial cells included) had the highest positive correlation (Fig. 2f and Supplementary Fig. 5e). Consistently, strong stromal influences were detected in the transforming growth factor-β (TGF-β) and cognate receptor expression (Supplementary Fig. 4f). Constant exposure of tumor cells to the TGF-β-rich CMS4 microenvironment might have induced an elevation of the levels of TGF-β pathway components (Supplementary Fig. 4g, TGF-β signaling pathway).

It is worth noting that tumor cells in individual patients demonstrated differing intratumoral heterogeneity in CMS signatures (Fig. 2e and Supplementary Fig. 5a,b,d) and in other cancer-related signatures, such as proliferation, stemness, hypoxia and chemotherapy resistance (Supplementary Figs. 4g and 5c). Collectively, our tumor cell analysis demonstrated lineage-associated oncogenic events, tuning of transcriptional signatures by the tumor microenvironment and intratumoral heterogeneity, implying functional diversification and environmental adaptation.

Stromal cell dynamics supporting tumor growth. Among stromal cell subclusters, fibroblasts and endothelial cells were the major cell types (Fig. 3a, Supplementary Figs. 6 and 7 and Supplementary Table 3). Additionally, pericytes, vascular smooth muscle cells and enteric glial cells were identified as minor stromal cell types (Fig. 3a). Pericytes were characterized by the coexpression of contractile genes (*TAGLN*, *ACTA2*) and *RGS5* (ref. [20]); vascular smooth muscle cells by contractile and cytoskeletal gene expression (*TAGLN*, *ACTA2*, *MYH11*, *SYNPO2*, *CNN1* and *DES*)[21]; and enteric glial cells by *SOX10*, *S100B* and *PLP1* (ref. [22]) (Fig. 3b).

Fibroblast subclusters were defined by the expression of *COL3A1*, *DCN* and *THY1* (ref. [23]) (Fig. 3b, top left). In normal mucosa, stromal fibroblast types, denoted as S1, S2 and S3, were identified[23] (Fig. 3b), resembling lipofibroblasts, PDGFRA/B^{hi} and matrix fibroblasts from lung tissue[24–26]. The S2 fibroblasts reside in proximity to the epithelial monolayer[23]. Thus, expression of BMP and Wnt pathway genes (*BMP2*, *BMP5*, *FRZB*, *POSTN* and *WNT5A*) by S2 fibroblasts may support the proliferation and function of epithelial progenitor cells[27] (Fig. 3b). In contrast, myofibroblasts exclusively populated CRC tissues (clusters 1, 10, 12, 21 and 22 in Fig. 3a and Supplementary Figs. 6c and 7d,e).

Myofibroblast presence was most prominent in CMS1 and CMS4 tissues (Fig. 3a,c) and manifested in heterogeneous gene expression phenotypes (MF1–4; Fig. 3b,d). MF1 was enriched for extracellular matrix synthesis, MF2 for extracellular matrix degradation and inflammatory chemokines, MF3 for *RGS5* (suggesting a pericyte

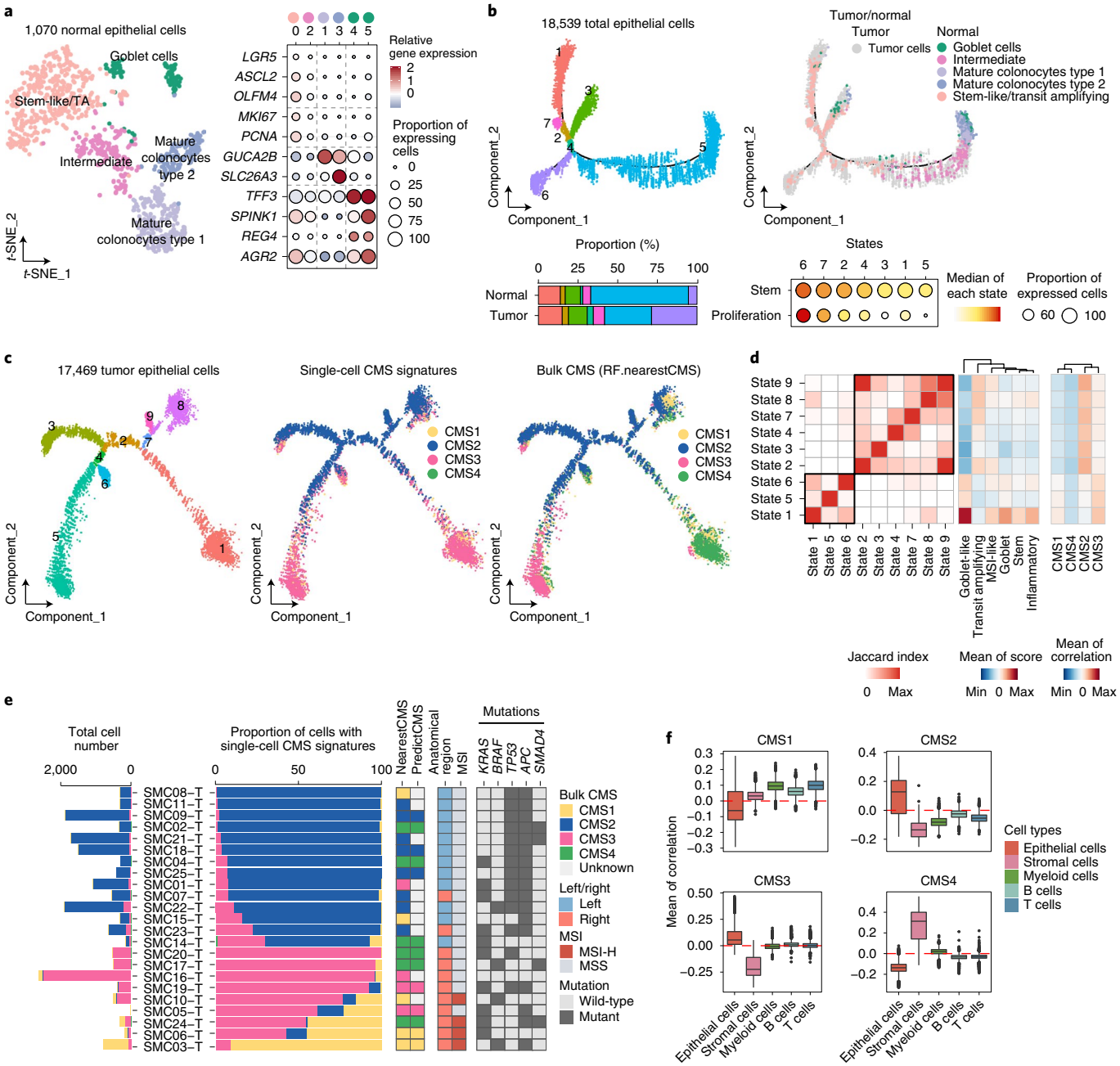

Fig. 2 | Transcriptome signatures and heterogeneity in normal and tumor epithelial cells. a, *t*-SNE plot of normal colon epithelial cells of the SMC dataset, color-coded by intestinal subtype (left). Canonical marker gene expression for intestinal cell types centered to the average expression of each gene across all normal epithelial cells with a scale from 2 to −2. The dot size represents the proportion of expressing cells in each cluster (right). **b**, The semisupervised trajectory of normal epithelial cells inferred by Monocle v.2, color-coded by state (top left) or subtype of normal epithelial and tumor cells (top right). The proportion of each state in CRC tissue and normal mucosa (bottom left) is shown. The median expression of stemness- and proliferation-associated (*MKI67*, *PCNA*) genes in each state (bottom right) is shown. **c**, Unsupervised trajectory of tumor cells, visualizing cell states (left), CMS signatures at the single-cell level (middle) and corresponding tissue CMS classes (right). **d**, Similarity matrix for significant DEGs in each state ($n = 2{,}754$ cells for state 1; 1,685 for state 2; 4,191 for state 3; 242 for state 4; 4,521 for state 5; 575 for state 6; 217 for state 7; 2,285 for state 8; 999 for state 9). The color indicates the Jaccard index between DEGs (left). Cluster average scores for CRC feature genes or Pearson's correlation to CRC molecular subtype (right) are shown. **e**, Number and relative proportions of four CMS-like tumor subpopulations (left, middle) ordered by prevalence of CMS2-like cells to CMS1-like cells. CMS classes from bulk RNA-seq, clinical information and mutations from bulk whole-genome sequencing data are aligned (right). **f**, Box plots representing the average correlation values with the CMS centroid data (5 sets) for each cell type in the tissue CMS groups ($n = 17{,}469$ cells for tumor epithelial cells; 2,736 for stromal cells; 6,400 for myeloid cells; 3,938 for B cells; 16,739 for T cells). Each box plot describes the median and IQR of each score. The whiskers depict the 1.5× IQR. The *P* value of the one-way ANOVA and all *P* values are less than 2×10^{-16}.

origin) and MF4 for proliferation with MF2-like features (MF4, cluster 22 in Supplementary Fig. 6d)[28]. All myofibroblast clusters highly expressed Wnt signaling genes such as *POSTN* and *WNT5A* (Fig. 3b). These data suggest that myofibroblasts stimulate tumor growth by extensive tissue remodeling and by supporting cancer stem cell survival through Wnt signaling[29].

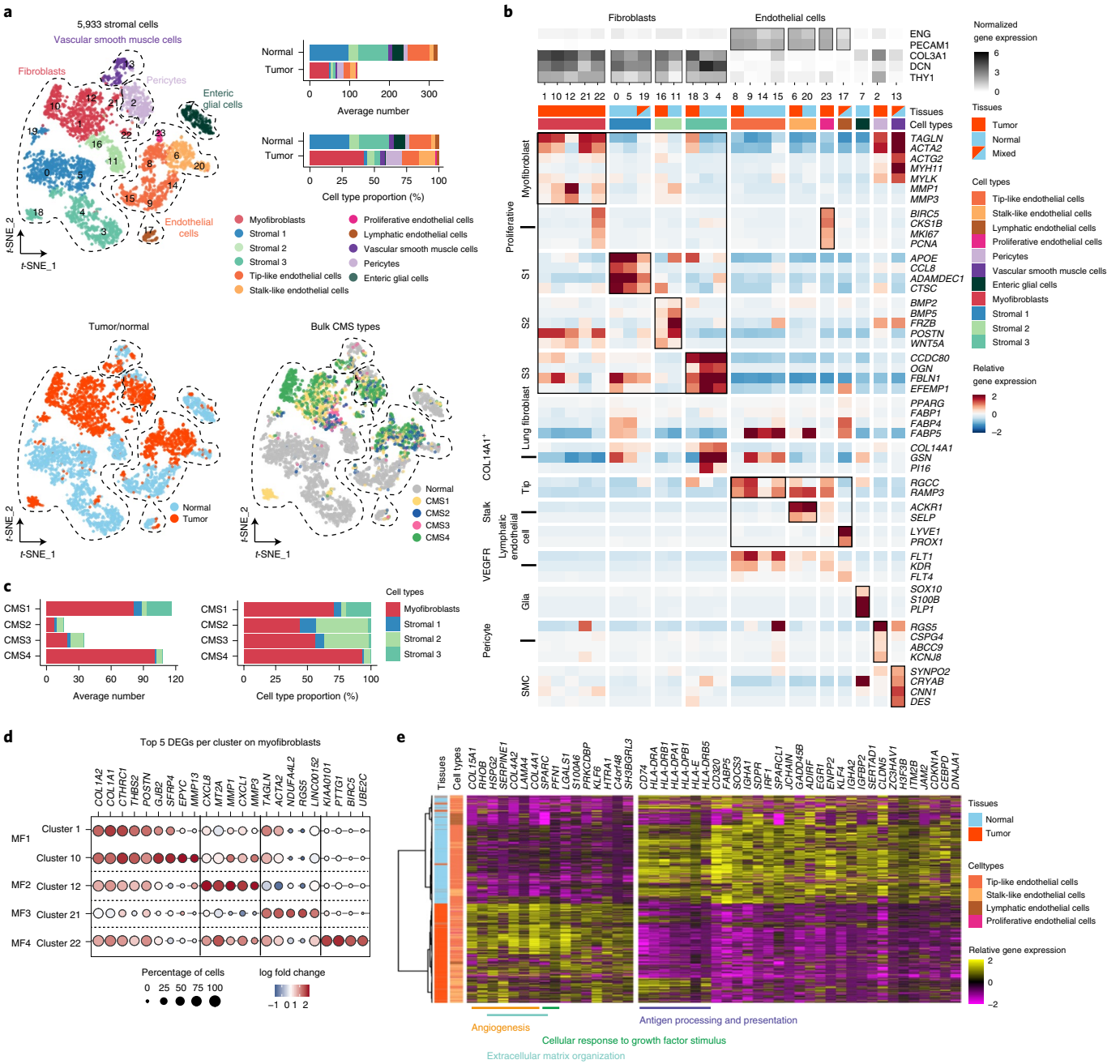

Fig. 3 | Stromal cell dynamics in normal mucosa and CRC tissue. a, *t*-SNE plot of 5,933 stromal cells derived from the SMC dataset, color-coded by sub-cell type (top left), sample origin and tissue CMS class of the tumor samples (bottom). Comparison of the average cell numbers and proportions of sub-cell types in tumor tissue and normal mucosa (top right) is shown. **b**, Heatmap of canonical lineage marker genes for fibroblasts and endothelial cells. The color of each square indicates the average gene expression for endothelial cells and fibroblasts in the natural log scale (white to black). The average gene expression heatmap of selected canonical markers for stromal cell subsets on each cluster is shown. The expression was centered to the average expression across all single cells with a scale from 2 to −2. VEGFR, vascular endothelial growth factor receptor. **c**, Proportion and average cell number of sub-cell types among fibroblasts for each tissue CMS class. **d**, Top 5 DEGs per cluster on myofibroblasts. MF1, stromal clusters 1 and 10; MF2, stromal cluster 12; MF3, stromal cluster 21; MF4, stromal cluster 22, respectively. **e**, Hierarchical clustering heatmap of 1,507 endothelial cells showing dysregulated genes in cancer for both tip-like and stalk-like endothelial cells (greater than twofold, two-sided *t*-test, $P < 0.01$, adjusted $P < 0.01$, and percentage of expressed cells >0.25). The expression of each gene was centered to the average expression across endothelial cells with a scale from 2 to −2. The color indexes on the bottom show the representative biological gene ontology terms for the marked genes.

Endothelial cell clusters were characterized by *ENG* and *PECAM1* gene expression without fibroblast markers[23] (Fig. 3a,b, top right). We assigned eight endothelial cell clusters to four endothelial cell types, including tip and stalk cells[30,31]. One of the subpopulations, lymphatic endothelial cells, was found in both tumor and normal samples (cluster 17 in Fig. 3a,b). Vascular endothelial cell clusters included tip- and stalk-like endothelial cells, the majority of which were either tumor-derived (clusters 6 and 8 in Fig. 3a,b) or normal colon-derived (clusters 9, 14, 15 and 20). Similar endothelial subpopulations were identified in the SMC and KUL3 datasets (Fig.3b and Supplementary Fig. 7a–d). Notably, comparison of the tip and stalk cell types between

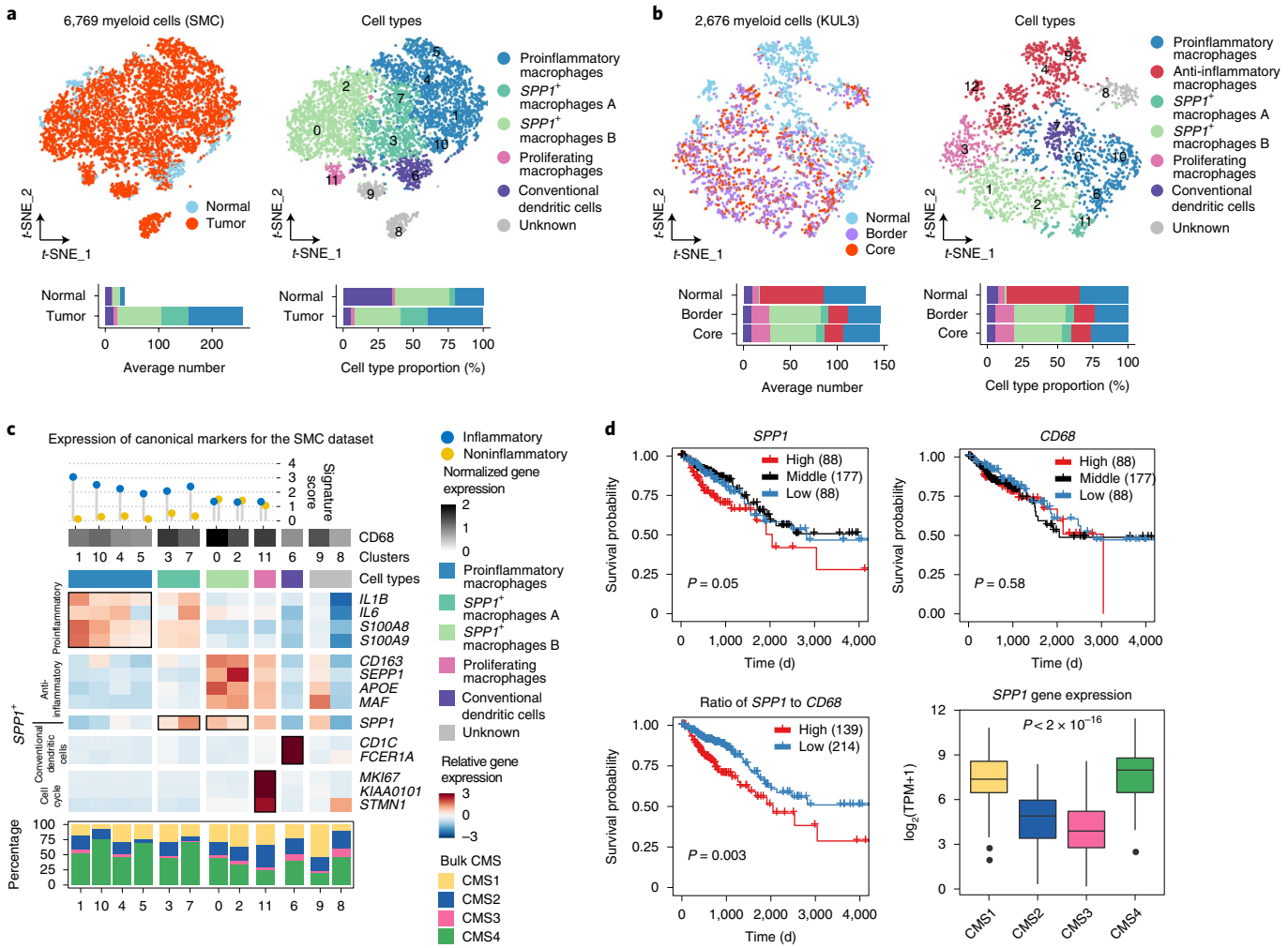

Fig. 4 | Transcriptional reprogramming in tumor-associated macrophages. a,**b**, *t*-SNE plot of 6,769 myeloid cells from the SMC dataset (**a**) and 2,676 myeloid cells from the KUL3 dataset (**b**), color-coded by sample origin and cell subtype and annotated by cluster number (top). Comparison of cell number and proportion of specified cell subtypes between tumor (core or border in KUL3) and normal tissues (bottom) is shown. **c**, Average expression of canonical marker genes for myeloid subpopulations on each cluster identified the *SPP1*+ subset. Comparison of inflammatory and noninflammatory signature levels in each cluster indicates pro- or anti-inflammatory macrophage groups. Signature levels were calculated as the mean expression of inflammatory (*IL1B*, *IL6*, *S100A8* and *S100A9*) or anti-inflammatory genes (*CD163*, *SEPP1*/*SELENOP*, *APOE* and *MAF*) genes (top). Normalized expression of myeloid lineage marker gene (*CD68*) is shown. Gene expression was centered to the average expression across all myeloid cells (middle). The percentage of CMS tissue classes per cluster is shown (bottom). **d**, Survival analysis using 353 TCGA COAD and READ samples according to *SPP1* and *CD68*, and the ratio of *SPP1* to *CD68*. *SPP1*/*CD68* > 1, high; *SPP1*/*CD68* < 1, low (bottom left). The *P* value in the box was calculated using a two-sided log-rank test. Comparison of *SPP1* gene expression between CMS classes from 353 TCGA COAD and READ RNA-seq data is shown ($n = 63$ samples for CMS1; 110 for CMS2; 76 for CMS3; 104 for CMS4). The box plot describes the median and IQR of each score. The whiskers depict the 1.5× IQR. The *P* value of the one-way ANOVA is shown (bottom right).

normal and tumor tissues revealed the overrepresentation of 'regulators of angiogenesis' in tumors and 'antigen processing and presentation' in normal mucosa (Fig. 3e and Supplementary Fig. 7f). Moreover, we identified a highly proliferative cell population with overexpression of survivin (coded by *BIRC5*) and *CKS1B* as a unique tumor-derived cluster (proliferative endothelial cells, cluster 23 in Supplementary Fig. 6d and Fig. 3b). Previous studies associated these genes with endothelial cell apoptosis inhibition and proliferation[32,33]. The highly angiogenic nature of endothelial cells reflects activated neovascularization in colon cancer. In addition, under-represented antigen-presenting vascular endothelial cell numbers suggest attenuated immune stimulation (Fig. 3e and Supplementary Fig. 7). Downregulation of *ICAM1* gene expression[34] (Supplementary Table 4) also supports immune attenuation of the endothelial components in CRC.

Shift from mucosal to suppressive cellular immunity. Myeloid expansion in CRC tissue (Fig. 1b,c and Supplementary Fig. 2) suggests their active role in shaping the tumor microenvironment. Subclustering analysis classified myeloid cells as either macrophages or dendritic cells (Fig. 4a,b and Supplementary Table 5). The macrophage clusters included both proinflammatory and anti-inflammatory subpopulations[35–37] (Fig. 4a–c and Supplementary Fig. 8). The dendritic cell cluster expressed transcripts for *CD1C* and *FCER1A*, consistent with conventional dendritic cells[38]. We recovered very few myeloid cells from normal mucosa in SMC patients compared to KUL3 patients (Fig. 4a,b). Nonetheless, we found similar myeloid cell types in both datasets (Fig. 4a,b). The macrophages prevalent in the tumor tissues had a mixed phenotype, expressing both proinflammatory and anti-inflammatory genes (Fig. 4c and Supplementary Fig. 8a,b).

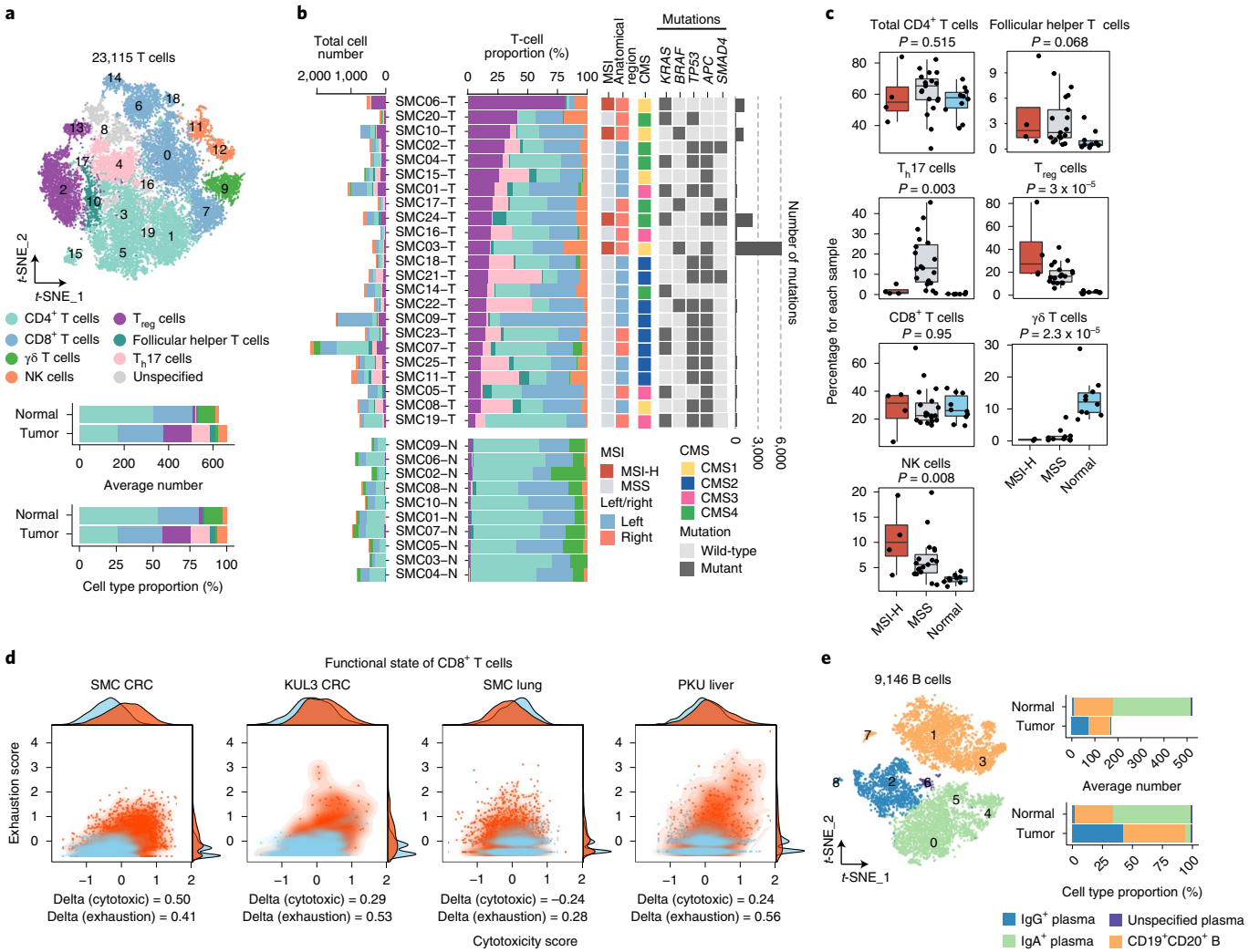

Fig. 5 | Adaptive immune cell profiles in normal mucosa and CRC tissue. a, *t*-SNE plot of 23,115 T lymphocytes color-coded by cell type and annotated by cluster number (top). Comparison of average cell numbers and proportions (excluding unspecified cells) in normal tissues and tumors color-coded by T-lymphocyte subset (bottom). **b**, Cell numbers (left) and relative proportions (middle) of T-lymphocyte subsets color-coded by subtype as in Fig. 5a. Clinical and genetic information is aligned on the right-hand side. **c**, Box plots comparing the proportion of T-lymphocyte subsets among 4 MSI-H or 19 MSS tumor samples and 10 normal samples. The box plot describes the median and IQR of each score. The whiskers depict the 1.5× IQR. The *P* value of the one-way ANOVA is shown. **d**, Comparisons of CD8+ T-cell phenotype in multiple cancer types: SMC CRC cohort; KUL3 CRC cohort; SMC lung cancer cohort from GSE131907; and PKU liver cancer cohort, color-coded by sample origin (orange, tumor samples; sky blue, normal samples). The delta scores on the bottom represent changes in mean expression for cytotoxicity or exhaustion scores from normal to tumor. **e**, *t*-SNE plot of 9,146 B cells from the SMC dataset, color-coded by cell type and annotated for cluster labels. Comparison of average cell numbers and proportions for B-cell subsets (including plasma cells) between tumor and normal tissues is shown.

In particular, the secreted phosphoprotein 1 (*SPP1*)+ macrophage fraction was enriched in the tumor core and border compared with normal tissues (Supplementary Table 6). *SPP1* gene expression, which produces osteopontin, can be induced by proinflammatory molecules, such as lipopolysaccharide, nitrogen oxide, interferon gamma and interleukin-1β[39]. However, *SPP1*+ macrophages also express apolipoprotein E (*APOE*), which is associated with an anti-inflammatory phenotype[35]. Given the myeloid cell profile of CRC, marked expansion of *SPP1*+ macrophages might play a central role in immune suppression and tumor progression through osteopontin[40,41]. These results suggest that tumor-associated macrophages may undergo additional transcriptional reprogramming distinct from the pro- and anti-inflammatory differentiation axis (Supplementary Fig. 8c). *SPP1*+ macrophages show increased presence in patients with the CMS1 and CMS4 types (Fig. 4c), which was validated using The Cancer Genome Atlas (TCGA) colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) datasets (Fig. 4d and Supplementary Fig. 8e). Furthermore, patients with higher levels of *SPP1*+ gene expression showed worse prognosis (Fig. 4d), reflecting the clinical impact of expanded *SPP1*+ macrophage cluster in CRC.

T lymphocytes comprised the largest immune cell cluster in both normal mucosa and CRC tissues (Fig. 1b,c). Twenty T-cell subclusters in the SMC dataset represented diverse subpopulations of CD4+, CD8+, γδ T cells and NK cells (Fig. 5a, Supplementary Fig. 9 and Supplementary Table 7). Both CD4+ and CD8+ T-cell clusters contained several subclusters representing the transition from naïve to activated or exhausted states (Supplementary Fig. 9b). We separately labeled the regulatory (T_{reg}) and T helper 17 (T_h17) subsets from other CD4+ T-cell clusters.

Many T-cell subtypes showed unequal distributions in normal and tumor tissues: T_h17 and T_{reg} cells were found predominantly in

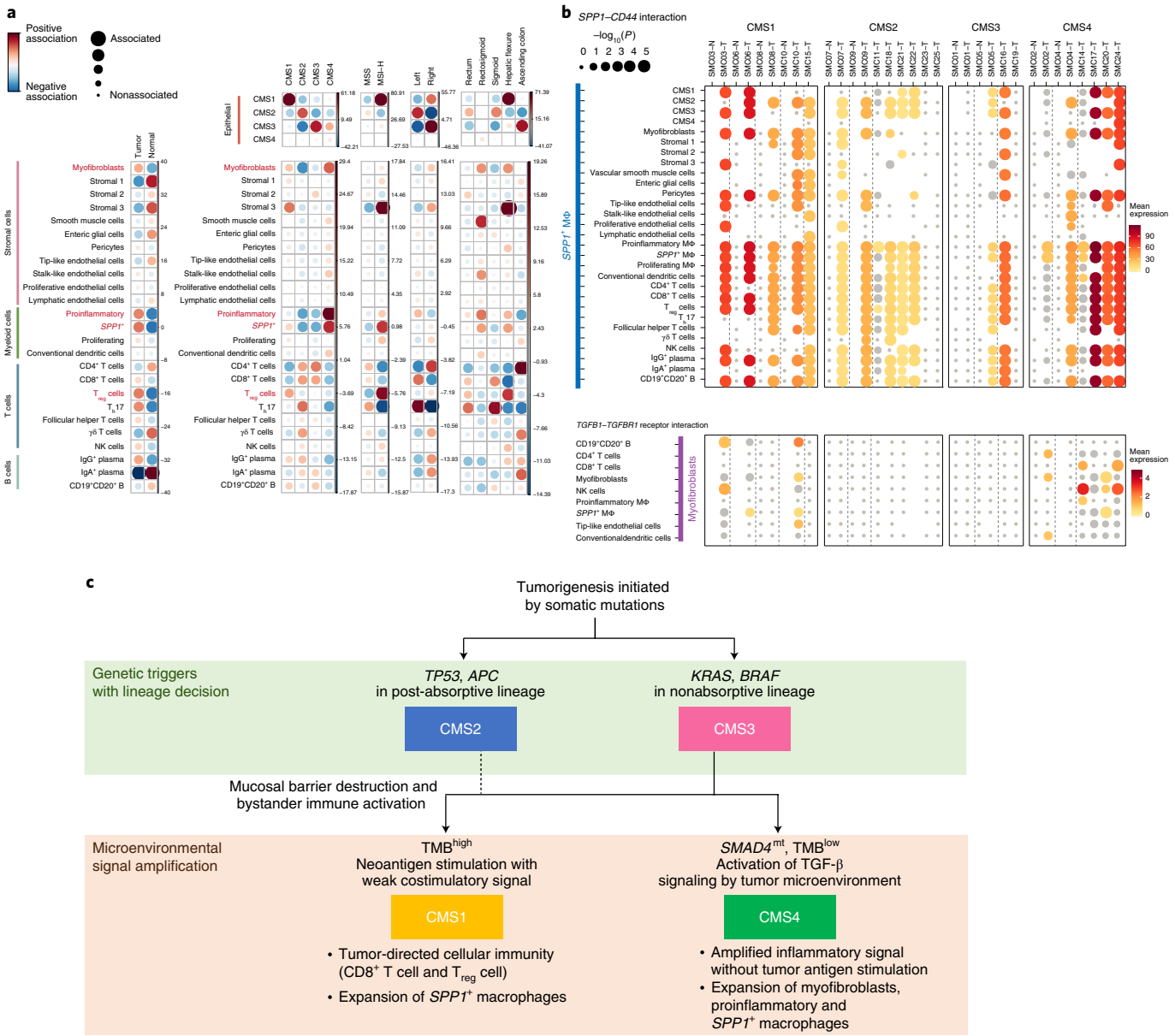

Fig. 6 | Differential reprogramming of the cellular interaction network in CRC for tissue CMS classes. a, Alterations in cellular dynamics between 23 tumors and 10 normal tissues and between bulk CMS classes, MSI status and anatomical region in 23 tumor tissues (left to right) are shown. Dot size represents Pearson's residual of the chi-squared test and the color represents the degree of positive or negative association from Pearson's residual of the chi-squared test. **b**, Receptor–ligand pair expression per sample using CellPhoneDB in 23 tumors and 10 normal tissues: *SPP1–CD44* (top); *TGFB1–TGFBR1* (bottom). The dot color represents the mean value of the average ligand expression in a cell type and the average receptor expression in a cell type. The dot size represents $-\log_{10}(P)$. The *P* value was calculated using the proportion of the mean value for specific receptor–ligand pairs compared to a randomly permuted mean distribution. The gray dot indicates a nonsignificant interaction. **c**, Schematic model of CRC landscape shaping by genetic triggers during tumorigenesis and signal amplification by microenvironmental factors. *SPP1*$^+$ MΦ, SPP1$^+$ macrophages; proinflammatory MΦ, proinflammatory macrophages; proliferating MΦ, proliferating macrophages.

tumor tissues, but γδ T cells were enriched in the normal mucosa (Fig. 5a,b). When we evaluated immune cell composition in relation to MSI status as a surrogate marker for a positive response to immune checkpoint inhibitors, MSI-high (MSI-H) samples were enriched in patients with higher T_{reg} populations and lower T_h17 populations in the SMC dataset (Fig. 5b,c). We also evaluated the functional status of CD8$^+$ T-cell subpopulations by scoring for cytotoxicity and exhaustion[42] (Fig. 5d). CD8$^+$ T cells in tumor tissues had higher cytotoxicity and exhaustion scores compared to those in normal mucosa (Fig. 5d and Supplementary Fig. 9b). This equivalent increase was not found in lung cancer, where exhaustion surpassed cytotoxicity in tumor-derived CD8$^+$ T cells (Fig. 5d, middle). Of note, MSI-H patients in CRC sustained slightly lower exhaustion scores compared to microsatellite stable (MSS) patients (Supplementary Fig. 9c), indicating greater functionality. The KUL3 dataset demonstrated T-cell subtypes similar to those observed in the SMC dataset and corroborated tumor-associated alterations (Supplementary Fig. 10a–e) with the enrichment of T_{reg} (cluster 0 and 6 in Supplementary Fig. 10e) and cytotoxic/exhausted CD8$^+$ T cells (cluster 13) in tumor tissues.

Within the B-cell cluster of the SMC dataset, follicular B cells and plasma cells were identified (Fig. 5e and Supplementary Fig. 11a).

In the CRC tissue, we found a significant reduction in the B/plasma cell compartment (Figs. 1b,c and 5e) and different antibody isotype expression[43] (Fig. 5e and Supplementary Fig. 11b). The decreased production of *IGHA* and increased expression of *IGHG* in CRC was observed in most of the light chain-specific plasma cells (Supplementary Fig. 12a), suggesting a systemic change in the immune microenvironment. An overall shrinkage of *IGHA* plasma cells was observed in the core and borders of tumors in the KUL3 dataset, but a shift to *IGHG* gene expression was more evident at the tumor core (Supplementary Figs. 11c–f and 12b).

CMS-specific cellular interaction network. To statistically evaluate the alterations in cellular dynamics, we performed chi-squared tests in 25 cell subtypes for the normal versus tumor tissue CMS groups, MSS versus MSI-H or anatomical location groups (Fig. 6a). We observed an enrichment of myofibroblasts, macrophages, $CD8^+$ T cells, T_{reg} cells, T_h17 cells and immunoglobulin G $(IgG)^+$ plasma cells in tumors. By comparison, stromal fibroblasts, $CD4^+$ T cells, γδ T cells and IgA^+ plasma cells were enriched in normal mucosa. The CRC cellular landscape was differentially formulated by CMS, MSI-H status and anatomical region. In CMS1 tumors, including most of the MSI-H samples, T_{reg} cells were significantly enriched, whereas myofibroblasts and proinflammatory/$SPP1^+$ macrophages were most prominent in CMS4 tumors.

To understand the interactions among different cell populations and how they jointly create the colorectal tumor microenvironment, we inferred a putative cellular interaction network based on the receptor–ligand database[44]. In normal mucosa, the strongest interaction nodes were formed around IgA^+ plasma cells and $CD4^+$ T cells, supporting the activation of mucosal antibody response (Supplementary Fig. 13a). The IgA network was diminished in CRC tissue, but new interaction edges emerged around tumor cells, connecting to myofibroblasts, $SPP1^+$ macrophages, $CD8^+$ T cells and T_{reg} cells. Furthermore, we analyzed the CRC cellular network by CMS tissue subtype, accommodating gross differences in the tumor microenvironment (Supplementary Fig. 13b). In CMS1 and CMS4 CRC, the top 20 predicted interactions involved mostly nontumor cell types, such as T_{reg} cells for CMS1 or myofibroblasts and $SPP1^+$ macrophages for CMS4 tumors. In CMS2 and CMS3 tissue, the top 20 interactions were all predicted to occur between tumor and nontumor cell types, demonstrating the dominant influence of tumor cells in shaping the tumor microenvironment.

In the molecular reconstitution for the specific receptor–ligand pairs[45,46], we estimated prominent *SPP1–CD44* and *TGFB1–TGFBR* interactions in CMS4 tumors (Fig. 6b), suggesting that these molecular interactions are crucial in creating an immunosuppressive and prometastatic tumor microenvironment[47–49]. Collectively, our comprehensive cellular landscape and interaction network reveal the molecular characteristics of CRC, sculpted by lineage-associated genetic alterations, together with tumor-driven or bystander alterations in the stromal and immune compartment (Fig. 6c).

Discussion

To date, single-cell transcriptome studies in CRC have characterized very few selected cell types[4,5,11]. In this study, we comprehensively explored the cellular landscape and reconstructed the putative interaction network between tumor cells and their microenvironment. This collective view allowed us to elucidate how these diverse cellular components jointly determine CRC molecular subtypes in individual patients.

In normal mucosa, we illustrated the dominance of IgA-type humoral immunity and γδ T-cell-driven innate immunity, maintaining the gut homeostasis[50,51]. In CRC, mucosal immunity was shifted toward suppressive cellular immunity marked by enrichment of $SPP1^+$ macrophages and T_{reg} cells. Stromal myofibroblasts and endothelial cells also contributed to the immune suppressive microenvironment by TGF-β production or downregulation of major histocompatibility class II and leukocyte adhesion molecules. T_{reg} cells and myofibroblasts have been commonly observed in other cancer types including lung and liver[9,42,52].

Tumor cells are at the center of the cellular interaction network in CRC, defining the molecular characteristics of tumor tissues. The tumor cell signatures primarily represent absorptive or secretory cell lineages, which coincide with CMS2 or CMS3 CRC tissue subtype gene expression features. Tumor cells showing CMS3-like gene expression features often coexpressed CMS1-like immune signatures, but CMS2-like cancer cells were devoid of immune gene expression. This suggested that tumor-derived immune signatures are secondary features upregulated in inherently CMS3-like cancer cells. The prevalence of specific immune cell populations in CMS1 tissues further suggests that T_{reg} cells may have substantiated these secondary features. We speculate that the initial triggers causing the formation of these unique tumor microenvironments were the genetic alterations in tumor cells. First, *KRAS* and *BRAF* wild-type tumors were associated with CMS2-like tumor cells and tissues, but *KRAS* or *BRAF* mutant tumors were more prevalent in non-CMS2-like tumor cells and tissues. *KRAS*/*BRAF* mutations are known to trigger NF-κB activation and inflammatory signaling[53,54]. Second, CMS1-like tumor cells were enriched in MSI-H patients in whom the high mutational burden would have elicited T-cell immune responses. Notably, CMS1 tissues devoid of CMS1-like tumor cells suggest alternative nontumor cell-driven bystander mechanisms of immune cell activation. The latter CMS1 may not respond to immune checkpoint therapies due to the absence of tumor-specific immunity. Third, *SMAD4* mutations may support tumor cell survival in a TGF-β-rich microenvironment formulated by myeloid cells and myofibroblasts[55,56]. The incomplete association between genetic alterations and molecular characteristics of tumor tissues indicates that additional factors are involved in determining CRC's ultimate cellular and molecular landscape. Candidates include unidentified genetic factors, both germline and somatic, and environmental factors that affect inflammatory or costimulatory conditions like the microbiome. Once identified, these missing components would complete the translation of cancer cell signatures into a collective CRC landscape.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0636-z.

Received: 19 September 2019; Accepted: 28 April 2020;
Published online: 25 May 2020

References

1. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
2. Dienstmann, R. et al. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer* **17**, 79–92 (2017).
3. Ren, X., Kang, B. & Zhang, Z. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol.* **19**, 211 (2018).
4. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
5. Zhang, L. et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564**, 268–272 (2018).
6. Binnewies, M. et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).
7. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

8. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
9. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
10. Becht, E. et al. Immune and stromal classification of colorectal cancer is associated with molecular subtypes and relevant for precision immunotherapy. *Clin. Cancer Res.* **22**, 4057–4066 (2016).
11. Dalerba, P. et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29**, 1120–1127 (2011).
12. Parikh, K. et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49–55 (2019).
13. Gerbe, F. et al. Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *J. Cell Biol.* **192**, 767–780 (2011).
14. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
15. Suvà, M. L. & Tirosh, I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. *Mol. Cell* **75**, 7–12 (2019).
16. Calon, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
17. Isella, C. et al. Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.* **47**, 312–319 (2015).
18. Perez-Villamil, B. et al. Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer* **12**, 260 (2012).
19. Thanki, K. et al. Consensus molecular subtypes of colorectal cancer and their clinical implications. *Int. Biol. Biomed. J.* **3**, 105–111 (2017).
20. Berger, M., Bergers, G., Arnold, B., Hämmerling, G. J. & Ganss, R. Regulator of G-protein signaling-5 induction in pericytes coincides with active vessel remodeling during neovascularization. *Blood* **105**, 1094–1101 (2005).
21. Rensen, S. S. M., Doevendans, P. A. F. M. & van Eys, G. J. J. M. Regulation and characteristics of vascular smooth muscle cell phenotypic diversity. *Neth. Heart J.* **15**, 100–108 (2007).
22. Rao, M. et al. Enteric glia express proteolipid protein 1 and are a transcriptionally unique population of glia in the mammalian nervous system. *Glia* **63**, 2040–2057 (2015).
23. Kinchen, J. et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**, 372–386.e17 (2018).
24. Xie, T. et al. Single-cell deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis. *Cell Rep.* **22**, 3625–3640 (2018).
25. Green, J., Endale, M., Auer, H. & Perl, A.-K. T. Diversity of interstitial lung fibroblasts is regulated by platelet-derived growth factor receptor α kinase activity. *Am. J. Respir. Cell Mol. Biol.* **54**, 532–545 (2016).
26. Nabhan, A. N., Brownfield, D. G., Harbury, P. B., Krasnow, M. A. & Desai, T. J. Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science* **359**, 1118–1123 (2018).
27. Vanuytsel, T., Senger, S., Fasano, A. & Shea-Donohue, T. Major signaling pathways in intestinal stem cells. *Biochim. Biophys. Acta* **1830**, 2410–2426 (2013).
28. Otranto, M. et al. The role of the myofibroblast in tumor stroma remodeling. *Cell Adh. Migr.* **6**, 203–219 (2012).
29. Vermeulen, L. et al. Wnt activity defines colon cancer stem cells and is regulated by the microenvironment. *Nat. Cell Biol.* **12**, 468–476 (2010).
30. Kumar, A. et al. Specification and diversification of pericytes and smooth muscle cells from mesenchymoangioblasts. *Cell Rep.* **19**, 1902–1916 (2017).
31. Zhao, Q. et al. Single-cell transcriptome analyses reveal endothelial cell heterogeneity in tumors and changes following antiangiogenic treatment. *Cancer Res.* **78**, 2370–2382 (2018).
32. Lamorte, S. et al. Syndecan-1 promotes the angiogenic phenotype of multiple myeloma endothelial cells. *Leukemia* **26**, 1081–1090 (2012).
33. O'Connor, D. S. et al. Control of apoptosis during angiogenesis by survivin expression in endothelial cells. *Am. J. Pathol.* **156**, 393–398 (2000).
34. Griffioen, A. W., Damen, C. A., Blijham, G. H. & Groenewegen, G. Tumor angiogenesis is accompanied by a decreased inflammatory response of tumor-associated endothelium. *Blood* **88**, 667–673 (1996).
35. Baitsch, D. et al. Apolipoprotein E induces antiinflammatory phenotype in macrophages. *Arterioscler. Thromb. Vasc. Biol.* **31**, 1160–1168 (2011).
36. Benoit, M. E., Clarke, E. V., Morgado, P., Fraser, D. A. & Tenner, A. J. Complement protein C1q directs macrophage polarization and limits inflammasome activity during the uptake of apoptotic cells. *J. Immunol.* **188**, 5682–5693 (2012).
37. Bronte, V. et al. Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards. *Nat. Commun.* **7**, 12150 (2016).
38. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
39. Guo, H., Cai, C. Q., Schroeder, R. A. & Kuo, P. C. Osteopontin is a negative feedback regulator of nitric oxide synthesis in murine macrophages. *J. Immunol.* **166**, 1079–1086 (2001).
40. Castello, L. M. et al. Osteopontin at the crossroads of inflammation and tumor progression. *Mediators Inflamm.* **2017**, 4049098 (2017).
41. Wang, K. X. & Denhardt, D. T. Osteopontin: role in immune regulation and stress responses. *Cytokine Growth Factor Rev.* **19**, 333–345 (2008).
42. Guo, X. et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).
43. Zhang, W. et al. Characterization of the B cell receptor repertoire in the intestinal mucosa and of tumor-infiltrating lymphocytes in colorectal adenoma and carcinoma. *J. Immunol.* **198**, 3719–3728 (2017).
44. Ramilowski, J. A. et al. A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
45. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
46. Efremova, M., Vento-Tormo, M., Teichman, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit receptor–ligand complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
47. Calon, A. et al. Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. *Cancer Cell* **22**, 571–584 (2012).
48. Klement, J. D. et al. An osteopontin/CD44 immune checkpoint controls $CD8^+$ T cell activation and tumor immune evasion. *J. Clin. Invest.* **128**, 5549–5560 (2018).
49. Tauriello, D. V. F. et al. TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* **554**, 538–543 (2018).
50. Gutzeit, C., Magri, G. & Cerutti, A. Intestinal IgA production and its role in host-microbe interaction. *Immunol. Rev.* **260**, 76–85 (2014).
51. Nielsen, M. M., Witherden, D. A. & Havran, W. L. γδ T cells in homeostasis and host defence of epithelial barrier tissues. *Nat. Rev. Immunol.* **17**, 733–745 (2017).
52. Zheng, C. et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356.e16 (2017).
53. Kitajima, S., Thummalapalli, R. & Barbie, D. A. Inflammation as a driver and vulnerability of KRAS mediated oncogenesis. *Semin. Cell Dev. Biol.* **58**, 127–135 (2016).
54. Trinh, A. et al. Tumour budding is associated with the mesenchymal colon cancer subtype and *RAS*/*RAF* mutations: a study of 1320 colorectal cancers with Consensus Molecular Subgroup (CMS) data. *Br. J. Cancer* **119**, 1244–1251 (2018).
55. Fessler, E. et al. TGFβ signaling directs serrated adenomas to the mesenchymal colorectal cancer subtype. *EMBO Mol. Med.* **8**, 745–760 (2016).
56. Levy, L. & Hill, C. S. Smad4 dependency defines two classes of transforming growth factor β (TGF-β) target genes and distinguishes TGF-β-induced epithelial-mesenchymal transition from its antiproliferative and migratory responses. *Mol. Cell. Biol.* **25**, 8108–8125 (2005).

Methods

Patient and tissue sample collection. This study was approved by the institutional review boards of the Samsung Medical Center (approval no. SMC2017-07-131) and Commissie Medische Ethiek UZ KU Leuven/Onderzoek (approval no. S50887-ML4707) for the SMC and KUL3 datasets, respectively. The study was carried out in accordance with ethical guidelines and all patients provided written informed consent. The study participants were 25 Korean and 6 Belgian patients diagnosed with CRC who underwent surgery without previous treatment. After resection, samples from the tumor and nonmalignant colon tissues were collected and immediately transferred for tissue preparation. Half of the tissues were subjected to single-cell isolation and the other half were cryopreserved for subsequent DNA and RNA extraction. Genomic DNA and RNA were extracted from bulk tumors or matched normal tissues by using the AllPrep DNA/RNA Mini Kit (QIAGEN) for DNA analysis and transcriptome sequencing.

For the SMC samples, tissue dissociation was performed using a Tumor Dissociation Kit (Miltenyi Biotec) according to the manufacturer's instructions. Briefly, tissues were cut into 2–4 mm-long pieces and transferred to C tubes containing an enzyme mix (enzymes H, R and A in Roswell Park Memorial Institute (RPMI) 1640 medium). Gentle MACS programs (h_tumor_01, 02 and 03) were run in a MACSmix Tube Rotator (Miltenyi) with two 30-min incubation periods at 37 °C between each run. The digested samples were filtered through a 70-µm strainer, purified using a Ficoll Paque PLUS (GE Healthcare) gradient and cryopreserved in CELLBANKER 1 (Zenoaq Resource) before scRNA-seq.

An enzymatic manual tumor dissociation method was used for the KUL3 samples. After rinsing with cold PBS, tumor specimens were minced to obtain <1 mm^3 pieces. The resulting pieces were incubated in digestion buffer (2 mg ml^{-1} collagenase P and 0.2 mg ml^{-1} DNase I in 10 ml of DMEM) for 10 min at 37 °C. The resulting suspension was filtered using a 40-µm nylon mesh, mixed with 30 ml of ice-cold PBS/FCS and centrifuged at 300*g* for 5 min at 4 °C. The supernatant was removed, the pellet was resuspended in 1 ml of red blood cell lysis buffer and the suspension was incubated at room temperature for 5 min. The suspension was centrifuged at 150*g* for 5 min at 4 °C and the supernatant was removed. The pellet was resuspended in 0.04% bovine serum albumin (BSA)/PBS and filtered using a 40-µm tip strainer. The number of cells in the resulting cell suspension was determined using a LUNA Counter (Logos Biosystems).

To compare sample preparation methods, CRC and matched normal tissues from one patient were processed using three dissociation protocols: SMC; KUL3; or Singapore (Supplementary Fig. 2a,b). The Singapore protocol consisted of tissue digestion using collagenase IV and DNase I in RPMI 1640 media at 37 °C for 1 h, followed by mechanical shearing with a 16-1/2-gauge needle, washing in PBS containing 1% BSA and 2 mM of EDTA and red blood cell lysis using the ACK lysis buffer. A single-cell suspension was generated from another patient sample using the SMC protocol and was subjected to scRNA-seq immediately or after a freeze–thaw cycle. Data were processed separately from the main 23 SMC and 6 KUL3 patient datasets. With this additional test dataset, a total of 109,512 cells from 31 patients were analyzed in the study.

DNA analysis. For the SMC dataset, whole-genome sequencing was performed using TrueSeq DNA Nano (Illumina) with a 150 base pair (bp) paired-end mode on a HiSeq X system (Illumina) at 30× coverage for normal samples and 60× for tumor samples. Sequencing data were aligned to a human genome reference (GRCh38) using the Burrows–Wheeler Aligner (v.0.7.17)[57] and preprocessed using the Genome Analysis Toolkit 4 (GATK4, v.4.0.2.1)[58]. A combination of GATK4 Mutect2 (Broad Institute) and Strelka2 (Strelka v.2.9.7)[59] was used to detect single nucleotide variations.

For the KUL3 dataset, the MSI and mutation status of the tumors were determined using the MSI Analysis System v.1.2 (Promega Corporation) and TruSight Cancer (Illumina) kits, respectively.

Bulk RNA-seq analysis. Total RNA sequencing libraries of the SMC dataset were constructed using a TruSeq Stranded Total RNA Library Preparation Kit with Ribo-Zero Gold (Illumina). Sequencing was performed in 100 bp paired-end mode on a HiSeq 4000 system (Illumina) at 120 million reads per sample. ERCC RNA Spike-In Mixes (Thermo Fisher Scientific) were included for quality assurance and the sequences were aligned to the human reference genome (GRCh38.p10). The RNA reads were aligned using STAR v.2.5.3.a (ref. [60]) and quantified as transcripts per million (TPM) using RSEM v.1.3.0 (ref. [61]). For the KUL3 dataset, FASTQ reads were trimmed for quality and adapter content using Trimmomatic v.0.36 (ref. [62]) and mapped to GRCh38 using STAR v.2.5.3a (ref. [60]). Subsequently, read counts were retrieved using HTseq v.0.10.0 (ref. [63]). To infer tumor purity from bulk tissue samples, we analyzed the stromal and immune scores using the ESTIMATE algorithm[64].

scRNA-seq and data processing. For the SMC dataset, the cryopreserved single-cell dissociates were rapidly thawed, washed and loaded into the Chromium system (10x Genomics) targeting 5,000 cells. For the KUL3 dataset, fresh single-cell suspensions were loaded into the Chromium system. Following the manufacturer's instructions, barcoded sequencing libraries were generated using Chromium Single Cell 3′ v2 Reagent Kits and sequenced across 8 lanes on a HiSeq 4000 platform targeting 100,000 reads per cell for the SMC dataset. RNA-seq data for the KUL3 dataset were generated using the NextSeq 500 and NovaSeq 6000 systems. Sequencing data were aligned to the human reference genome (GRCh38) and processed using the CellRanger 2.1.0 pipeline (10x Genomics). The raw gene expression matrix from the CellRanger pipeline was filtered, normalized using the Seurat R package and selected according to the following criteria: cells with >1,000 unique molecular identifier (UMI) counts; >200 genes and <6,000 genes; and <20% of mitochondrial gene expression in UMI counts. From the filtered cells, the gene expression matrices were normalized to the total UMI counts per cell and transformed to the natural log scale. For batch correction, we used multiple CCA implemented in Seurat v.2.3.4 (ref. [7]). Variable genes were selected as the top 1,000 highly variable genes expressed by more than 0.1% of cells in each sample. The MetageneBicorPlot function was used to select the number of standard correlation vectors, choosing the inflection point in most cases. We aligned the CCA subspaces and clustered and visualized the aligned CCA using *t*-SNE projection. The CCA resolution was selected variably per sample origin. The major cell types were annotated by comparing the canonical marker genes and the differentially expressed genes (DEGs) for each cluster.

Elimination of ambiguous cells using joint application of RCA and Seurat. We combined the Seurat and RCA pipelines for initial clustering and cell type identification, removing discordant cells from the downstream analysis[4,7]. RCA was performed using natural, log-normalized, TPM-like values corrected to library size for each sample. Briefly, the correlations between individual cells and bulk transcriptomes in the RCA global panel were calculated using genes expressed in more than two cells with a value >0.001. RCA calculated the Pearson's correlation coefficient between the natural log TPM-like values and the reference bulk transcriptome. Clustering was performed using the average linkage hierarchical clustering method in the WGCNA package v.1.63, with desepSplit = 1 and mingrupSize = 5. After joint application of Seurat multiple CCA and RCA, we selected intersecting cells with concordant cell type designations in the two algorithms for further analysis.

Unsupervised subclustering for the six global cell types. Subclustering was performed for the six major cell types defined in the initial clustering using the graph-based algorithm in Seurat. Before subclustering, cells with a number of genes exceeding the outliers were removed to eliminate doublets; 63,689 cells from the SMC dataset and 27,414 cells from the KUL3 dataset met the criteria. For clustering, variably expressed genes were selected with a mean expression between 0.0125 and 3 and a dispersion of more than 0.5 using the FindVariableGenes function. Principal component analysis was performed using the variably expressed genes; the number of principal components was differentially selected from the knee point of the scree plot for each cell type to accommodate different population complexities. Resolutions from 0.2 to 1.6 were explored for better subcluster representation.

Survival analysis. TCGA COAD and READ gene expression datasets and clinical dataset from Broad GDAC Firehose (https://gdac.broadinstitute.org/) were collected to analyze survival probability according to gene expression. After downloading gene expression from the Illumina HiSeq platform, we converted raw counts to normalized TPM values with $\log_2$-transformation summed with 1. We selected COAD and READ samples with gene expression, stage and clinical follow-up information and removed one sample with an ambiguous CMS classification. High-, middle- and low-expression groups were categorized using *SPP1* or *CD68* gene expression on 353 COAD and READ samples. The top 25% of samples were classified as the high group, the bottom 25% of samples as the low group and the remaining samples as the middle group. We also performed survival analysis using the *SPP1*/*CD68* ratio: *SPP1*/*CD68* > 1, high group; *SPP1*/*CD68* < 1, low group. The survfit and survdiff functions in R were used to generate Kaplan–Meier survival curves and calculate the *P* value of the log-rank test.

Trajectory analysis using Monocle v.2 and alignment with functionality scores. Monocle v.2 (ref. [14]) was used to illustrate the cell state transition in total epithelial cells, normal epithelial cells, cancer cells or myeloid cells. It applies a reversed graph embedding technique to reconstruct single-cell trajectories. Briefly, we used UMI count matrices and the negbinomial.size() parameter to create a CellDataSet object in the default setting. We used Monocle v.2 variable genes with the following cutoff criteria: dispersion_empirical > dispersion_fit; and mean expression >0.001. In the case of normal epithelial cell trajectories, Monocle v.2 variable genes were substituted with epithelial differentiation marker genes[4] for semisupervised trajectory reconstruction. Dimensional reduction and cell ordering were performed using the DDRTree method and the orderCells function. In the case of total epithelial cell trajectory, Monocle v.2 variable genes were substituted with significant DEGs from trajectory analysis of normal epithelial cells (Supplementary Table 2).

Functionality scores of $CD8^+$ T cells in various cancer types. To identify the functionality score of $CD8^+$ T cells in various cancer types, we obtained the SMC lung cancer samples from GSE131907 and the PKU liver cancer samples from

GSE140228 (ref. [65]). We reselected CD8+ T cells to include only those with *CD8A* or *CD8B* gene expression and no expression of *CD4* or *IL7R*, from the CD8+ T-cell clusters. Six cytotoxic marker genes (*CST7*, *GZMA*, *GZMB*, *IFNG*, *NKG7* and *PRF1*) and five exhaustion marker genes (*CTLA4*, *HAVCR2*, *LAG3*, *PDCD1* and *TIGIT*) were used to calculate mean expressions from the scaled data.

CMS classification for bulk RNA-seq and application to single cells. We used the CMSclassifier package v.1.0.0 to identify the CMS of CRC in the SMC bulk RNA-seq dataset[1]. After $\log_2$-transformation of TPM data summed with 0.001 (to avoid an infinite value), we used the random forest algorithm to classify CMS types. After the random forest algorithm, we used the nearestCMS and predictCMS values. TCGA COAD and READ data also used the nearestCMS value from the random forest algorithm. To identify the CMS of CRC in the KUL3 bulk RNA-seq dataset, we used the single sample predictor algorithm from the CMSclassifier package after rlog transformation from DESeq2 v.1.24 (ref. [66]) and scaling the data. After the single sample predictor algorithm, we used the nearestCMS value. For the scRNA-seq dataset from the SMC and KUL3 cohorts, we used the Pearson's correlation coefficient value between CMS centroid data and single cells. The CMS centroid data were obtained from the CMSclassifier package. The data consisted of 695 genes and 20 centroids, each consisting of 4 CMS models in the 5 datasets: TCGA COAD RNASeqV2 GA platform; HiSeq platform; Agilent platform; GSE39582; and E-MTAB-990. After calculation, the CMS type with the highest correlation mean was selected.

Analysis of gene set scores. To make a comparison with the transcriptional signatures of tumor epithelial cells (Fig. 2d), we used the mean expression of goblet, goblet-like, transit amplifying, MSI-like, stem or inflammatory[67] gene sets. To approximate tumor heterogeneity in cancer-related pathways, we used the mean expressions for the TGF-β signaling pathway genes concurrently present in the Kyoto Encyclopedia of Genes and Genomes from the MSigDB[68,69] and PANTHER[70,71] databases, epithelial–mesenchymal transition-associated genes[72], stem-associated genes[4], 5-fluorouracil resistance genes[73], hypoxia from hallmark gene sets and two proliferation genes (*MKI67*, *PCNA*) (Supplementary Figs. 4g and 5c). To calculate the weighted score of the TGF-β family (Supplementary Fig. 4f), we multiplied the mean expression of six genes (*TGFB1*, *TGFB2*, *TGFB3*, *TGFBR1*, *TGFBR2* and *TGFBR3*) and the proportion of epithelial cells, myeloid cells and fibroblasts. For the cell cycle scores in stromal cells (Supplementary Fig. 6), we calculated the mean expression of the gene sets for the cell cycle phases[74].

Intercellular interaction map for tumor and normal data. To analyze receptor–ligand pairing in tumor and normal data, we used curated human receptor–ligand pairs[44]. The value of each receptor–ligand pair was calculated by multiplying the number of ligand-expressing cells and the number of receptor-expressing cells in the counterpart subtypes. We then summed the number of pairs by cell subtype and collected the top 20 pairs. Node size reflects the number of cell subsets. The network plot was drawn using the migest R package v.1.7.4.

Interaction map in each patient sample using CellPhoneDB. To analyze receptor–ligand pairing in each patient, we used CellPhoneDB (www.cellphonedb.org)[45,46], a cell–cell interaction network tool for scRNA-seq datasets. To select only significantly expressed interaction pairs, we selected ligand or receptor genes expressed in more than 25% of the cell types in each patient. In addition, we permuted the change of cell type label for each cell at 1,000 times to calculate the significance of each pair. The *P* value was calculated using the proportion of the mean value for specific receptor–ligand pairs compared to a randomly permuted mean distribution.

Statistics. To determine statistical significance in the analysis of differential gene expression for each cluster, we used a two-sided Student's *t*-test with Bonferroni correction. Statistical significance between multiple cell types or sample types was determined using a one-way analysis of variance (ANOVA).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw scRNA-seq data of the SMC cohort are available in the European Genome-phenome Archive database (EGAS00001003779, EGAS00001003769). The raw scRNA-seq and bulk RNA-seq data of the KUL3 cohort are available in the ArrayExpress under the accession codes E-MTAB-8410 and E-MTAB-8412. Processed scRNA-seq and metadata for the SMC and KUL3 cohorts are available in the NCBI Gene Expression Omnibus (GEO) database under the accession codes GSE132465, GSE132257 and GSE144735. Clusters and gene expression data of the SMC cohort can be found on the User-friendly InteRface tool to Explore Cell Atlas (URECA) website (http://ureca-singlecell.kr). Other datasets referenced in the study are available from the GEO database under the accession codes GSE14028, GSE131907 and GSE81861.

Code availability

The code for the intercellular interaction map has been deposited with GitHub (https://github.com/SGI-CRC/scRNA-seq).

References

57. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
58. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
59. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
60. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
61. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
62. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
63. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
64. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
65. Zhang, Q. et al. Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* **179**, 829–845.e20 (2019).
66. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
67. Sadanandam, A. et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19**, 619–625 (2013).
68. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
69. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
70. Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.* **563**, 123–140 (2009).
71. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
72. De Sousa e Melo, F. et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.* **19**, 614–618 (2013).
73. Kang, H. C. et al. Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays. *Clin. Cancer Res.* **10**, 272–284 (2004).
74. Liu, Z. et al. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat. Commun.* **8**, 22 (2017).

Acknowledgements

This study was supported by the Bio & Medical Technology Development Program of the National Research Foundation funded by the Ministry of Science & ICT (grant no. NRF-2017M3A9A7050803), by the Belgian Federation against Cancer grant nos. 2018-127 and 2016-133 and by a grant from Fondation Roi-Baudouin. S.T. and S.V. are respectively supported by a Senior Clinical Investigator award and a postdoctoral fellowship of the Research Foundation—Flanders.

Author contributions

H.-O.L., Y.H., H.E.E. and Y.B.C. analyzed and interpreted the data. V.P., B.V.B., J.V. and H.H. processed the tumors. S.V., J.-W.M., N.K., H.H.E., J.Q., B.B., D.L., P.T., T.L., M.A. and P.W. provided bioinformatics support. M.-H.J., G.D.H., W.C., H.-T.S. and J.-G.J. set up the server and analyzed the bulk data. Y.H. and G.K. constructed the visualization website. S.H.K. provided pathological examination. H.C.K., S.H.Y., W.Y.L., T.-Y.K., J.K.C. and Y.-J.K. interpreted the clinical data. I.B.H.T., B.R. and S.P. provided critical bioinformatics guidance. S.T. and W.-Y.P. conceived and supervised the study. H.-O.L., Y.H., H.E.E., Y.B.C., S.T. and W.-Y.P. wrote the manuscript with contributions and approval from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0636-z.

Correspondence and requests for materials should be addressed to S.T. or W.-Y.P.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Sabine Tejpar and Woong-Yang Park

Last updated by author(s): Mar 31, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The single-cell data collection pipeline was established using CellRanger (v2.1) in this article. Total RNA sequencing data of SMC cohort was aligned using STAR (2.5.3.a) and quantified using RSEM(v1.3.0). For the KUL3 dataset, FASTQ reads were trimmed for quality and adapter content using Trimmomatic (v0.36) and mapped to GRCh38 using STAR (2.5.3a). Subsequently read counts were retrieved using HTseq(v0.10.0). Whole genome sequencing data of SMC cohort were aligned using Burrows-Wheeler Aligner (BWA 0.7.17) and preprocessed using the Genome Analysis Toolkit4 (GATK4, GATK 4.0.2.1). GATK4 Mutect2 (Broad Institute) and Strelka2 (Streka-2.9.7) were used for variant calling.
Data analysis	For data processing and analysis of single-cell RNA sequencing, Seurat (v2.3), RCA(v1.0), Monocle(v2.6), WGCNA (v1.6), migest (v1.7), and CellphoneDB (v2.0) are used under R (v3.4) or python (v3.7). CMSclassifier (v1.0) was used for cancer cell analysis for RNA-seq data of SMC and KUL3 cohort. The code for intercellular interaction map is deposited in Github (https://github.com/SGI-CRC/scRNA-seq).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw scRNA-seq data of the SMC cohort are available in the European Genome-phenome Archive database (EGAS00001003779, EGAS00001003769). Raw scRNA-seq and bulk RNA-seq data of the KUL3 cohort are available in the ArrayExpress (E-MTAB-8410, E-MTAB-8412). Processed single-cell RNA-seq and metadata for the SMC and KUL3 are available in the NCBI Gene Expression Omnibus database under the accession codes GSE132465, GSE132257, and GSE144735. Clusters and gene

expression data of the SMC cohort can be found on the interactive website http://ureca-singlecell.kr.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. We collected over 90,000 cells from single-cell RNA-seq of 52 samples available from the clinic, from 2 independent data collection sites. For all the study analyses, we had enough number of cells for statistical tests. We provided number of samples or cells in the figure legends.
Data exclusions	The raw gene expression matrix from the CellRanger pipeline was filtered, normalized using the Seurat R package, and selected according to the following criteria: cells with > 1,000 UMI counts, > 200 genes, and < 6,000 genes, and < 20% of mitochondrial gene expression in UMI counts. These are pre-determined cell filtration criteria excluding apoptotic cells and doublets. Next, we combined the Seurat and RCA pipelines for initial clustering and cell type identification, removing discordant cells from the downstream analysis. Before sub-clustering, cells with a number of genes exceeding the outliers were removed to eliminate doublets again.
Replication	Study subjects were colon cancer patients and no biological replicate available.
Randomization	The colorectal cancer samples without prior treatment were obtained without patient selection.
Blinding	All analyses were performed by computational methods.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Tumor and normal tissues of SMC cohort were obtained from stage I-IV Korean colorectal cancer patients after resection surgery at Samsung Medical Center. Tumor tissues are resected on diverse region. The 23 tumor tissues consist of 3 rectums, 3 hepatic flexures, 8 sigmoid regions, 8 ascending regions, 1 rectosigmoid region. Tumor core, border, and normal tissues from stage I-IV KUL3 cohort were obtained from 6 colorectal cancer patient after resection surgery at Universitair Ziekenhuis Leuven (UZ Leuven). The 6 tumor core tissues consist of 1 caecum, 1rectosigmoid region, 3 sigmoid regions, 1 ascending region. All patients are between 38 and 86 years old.
Recruitment	The 25 Korean and 6 Belgian patients diagnosed with colorectal cancer who underwent surgery without prior treatment were recruited in Samsung Medical Center and Universitair Ziekenhuis Leuven, respectively. All patients are selected regardless of age, gender, and stage. There was no bias in recruiting patients with colorectal cancer.
Ethics oversight	The study was approved by the Institutional Review Board (IRB) of Samsung Medical Center (approval no. SMC2017-07-131) and Commissie Medische Ethiek UZ KU Leuven / Onderzoek (approval no. S50887-ML4707) for the SMC and KUL3 datasets, respectively.

Note that full information on the approval of the study protocol must also be provided in the manuscript.