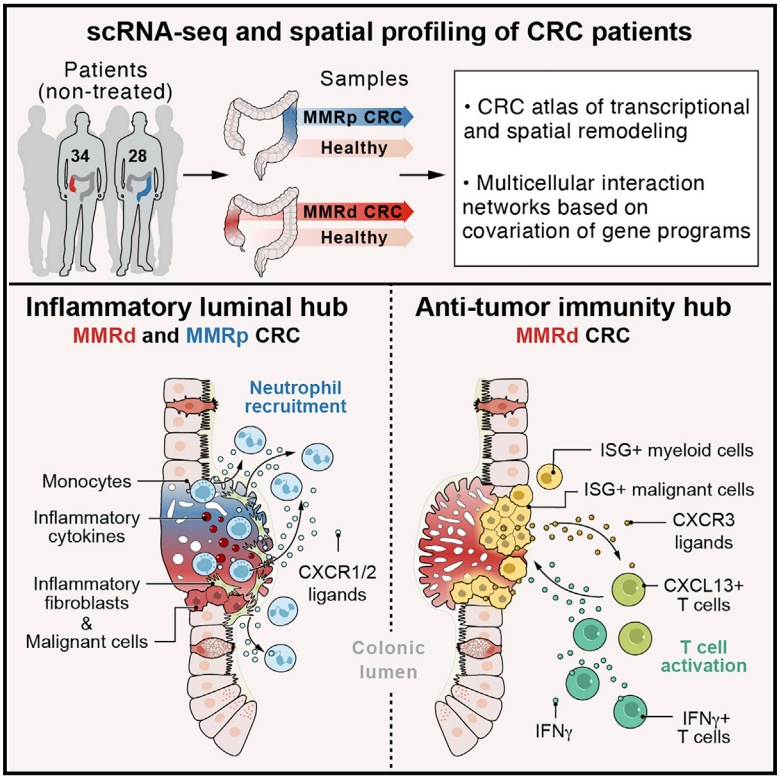
Cell

Spatially organized multicellular immune hubs in human colorectal cancer

Graphical abstract



Highlights

- A scRNA-seq study reveals shared and distinct features of human MMRd and MMRp CRC
- Co-variation of single-cell transcriptional programs across specimens predicts immune hubs
- A myeloid-rich inflammatory hub is identified below the colonic lumen in human CRC
- CXCR3-ligand+ cells form foci with activated T cells in human MMRd CRC

Authors

Karin Pelka, Matan Hofree, Jonathan H. Chen, ..., Orit Rozenblatt-Rosen, Aviv Regev, Nir Hacohen

Correspondence

acanderson@bwh.harvard.edu (A.C.A.), orit@broadinstitute.org (O.R.-R.), aviv.regev.sc@gmail.com (A.R.), nhacohen@mgh.harvard.edu (N.H.)

In brief

Single-cell transcriptomics-based co-variation analysis of human colorectal cancer identifies a spatially resolved myeloid-rich inflammatory hub that is shared by mismatch repair-deficient (MMRd) and mismatch repair-proficient (MMRp) tumors and CXCR3-ligand+ multicellular foci distinct for MMRd tumors.

Pelka et al., 2021, Cell *184*, 4734–4752
September 2, 2021
https://doi.org/10.1016/j.cell.2021.08.003

CellPress

CellPress
OPEN ACCESS

Cell

Article

Spatially organized multicellular immune hubs in human colorectal cancer

Karin Pelka,[1,2,20,25] Matan Hofree,[3,25] Jonathan H. Chen,[1,2,4,25] Siranush Sarkizova,[1] Joshua D. Pirl,[1] Vjola Jorgji,[2,4] Alborz Bejnood,[3] Danielle Dionne,[3] William H. Ge,[1] Katherine H. Xu,[2] Sherry X. Chao,[1,5] Daniel R. Zollinger,[6] David J. Lieb,[1] Jason W. Reeves,[6] Christopher A. Fuhrman,[6] Margaret L. Hoang,[6] Toni Delorey,[3] Lan T. Nguyen,[3] Julia Waldman,[3] Max Klapholz,[7] Isaac Wakiro,[8] Ofir Cohen,[1,8,9] Julian Albers,[1] Christopher S. Smillie,[3] Michael S. Cuoco,[3] Jingyi Wu,[8] Mei-ju Su,[8] Jason Yeung,[8] Brinda Vijaykumar,[10] Angela M. Magnuson,[10] Natasha Asinovski,[10] Tabea Moll,[11] Max N. Goder-Reiser,[11] Anise S. Applebaum,[11] Lauren K. Brais,[12] Laura K. DelloStritto,[8] Sarah L. Denning,[12] Susannah T. Phillips,[11] Emma K. Hill,[12] Julia K. Meehan,[12] Dennie T. Frederick,[11] Tatyana Sharova,[11] Abhay Kanodia,[8]

(Author list continued on next page)

[1]Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Cambridge, MA, USA
[2]Massachusetts General Hospital (MGH) Cancer Center, Harvard Medical School (HMS), Boston, MA, USA
[3]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[4]Department of Pathology, MGH, Boston, MA, USA
[5]Department of Biomedical Informatics, HMS, Boston, MA, USA
[6]NanoString Technologies Inc., Seattle, WA, USA
[7]Evergrande Center for Immunologic Diseases, HMS and Brigham and Women's Hospital (BWH), Boston, MA, USA
[8]Center for Cancer Genomics, Department of Medical Oncology, Dana-Farber Cancer Institute (DFCI), Boston, MA, USA
[9]Department of Medical Oncology, DFCI, Boston, MA, USA
[10]Department of Immunology, HMS, Boston, MA, USA
[11]Clinical Research Center, MGH, Boston, MA, USA
[12]Clinical Research Center, DFCI, Boston, MA, USA
[13]Department of Molecular Biology, MGH, Boston, MA, USA
[14]Department of Surgery, BWH, Boston, MA, USA
[15]Department of Surgery, MGH, Boston, MA, USA
[16]Department of Pathology, BWH, Boston, MA, USA

(Affiliations continued on next page)

SUMMARY

Immune responses to cancer are highly variable, with mismatch repair-deficient (MMRd) tumors exhibiting more anti-tumor immunity than mismatch repair-proficient (MMRp) tumors. To understand the rules governing these varied responses, we transcriptionally profiled 371,223 cells from colorectal tumors and adjacent normal tissues of 28 MMRp and 34 MMRd individuals. Analysis of 88 cell subsets and their 204 associated gene expression programs revealed extensive transcriptional and spatial remodeling across tumors. To discover hubs of interacting malignant and immune cells, we identified expression programs in different cell types that co-varied across tumors from affected individuals and used spatial profiling to localize coordinated programs. We discovered a myeloid cell-attracting hub at the tumor-luminal interface associated with tissue damage and an MMRd-enriched immune hub within the tumor, with activated T cells together with malignant and myeloid cells expressing T cell-attracting chemokines. By identifying interacting cellular programs, we reveal the logic underlying spatially organized immune-malignant cell networks.

INTRODUCTION

Almost all tumors are infiltrated with immune cells, but the types of immune responses and their effects on tumor growth, metastasis, and death vary greatly between different cancers and individual tumors (Thorsson et al., 2018). Which of the numerous cell subsets in a tumor contribute to the response, how their interactions are regulated, and how they are spatially organized within tumors remain poorly understood (Cardenas et al., 2021; Saltz et al., 2018).

Colorectal tumors show a large dynamic range of immune responsiveness, with a striking difference between two genetically distinct subtypes (Boland and Goel, 2010; Li and Martin, 2016): mismatch repair-deficient (MMRd) colorectal tumors have a high mutational burden, often contain cytotoxic T cell infiltrates, and have an ~50% response rate to immune checkpoint blockade, whereas mismatch repair-proficient (MMRp) tumors have a low mutational burden and are largely unresponsive to immunotherapy (André et al., 2020; Le et al., 2015, 2017; Overman et al., 2018).

CellPress
OPEN ACCESS

Ellen Z. Todres,[1] Judit Jané-Valbuena,[3] Moshe Biton,[3,13] Benjamin Izar,[3,8,9,21,22,23] Conner D. Lambden,[7] Thomas E. Clancy,[14] Ronald Bleday,[14] Nelya Melnitchouk,[14] Jennifer Irani,[14] Hiroko Kunitake,[15] David L. Berger,[15] Amitabh Srivastava,[16] Jason L. Hornick,[16] Shuji Ogino,[1,16] Asaf Rotem,[8] Sébastien Vigneau,[8] Bruce E. Johnson,[8,9] Ryan B. Corcoran,[2,18] Arlene H. Sharpe,[7,17] Vijay K. Kuchroo,[7] Kimmie Ng,[9] Marios Giannakis,[1,9] Linda T. Nieman,[2] Genevieve M. Boland,[2,15] Andrew J. Aguirre,[1,9] Ana C. Anderson,[7,*] Orit Rozenblatt-Rosen,[3,24,*] Aviv Regev,[3,19,24,26,*] and Nir Hacohen[1,2,10,26,27,*]

[17]Department of Immunology, Blavatnik Institute, HMS, Boston, MA, USA
[18]Department of Medicine, HMS, Boston, MA, USA
[19]Howard Hughes Medical Institute and Koch Institute for Integrative Cancer Research, Department of Biology, MIT, Cambridge, MA, USA
[20]Present address: Gladstone-University of California San Francisco Institute of Genomic Immunology, San Francisco, CA, USA
[21]Present address: Columbia Center for Translational Immunology, New York, NY, USA
[22]Present address: Columbia University Medical Center, Division of Hematology and Oncology, New York, NY, USA
[23]Present address: Program for Mathematical Genomics, Columbia University, New York, NY, USA
[24]Present address: Genentech, 1 DNA Way, South San Francisco, CA, USA
[25]These authors contributed equally
[26]These authors contributed equally
[27]Lead contact
*Correspondence: acanderson@bwh.harvard.edu (A.C.A.), orit@broadinstitute.org (O.R.-R.), aviv.regev.sc@gmail.com (A.R.), nhacohen@mgh.harvard.edu (N.H.)
https://doi.org/10.1016/j.cell.2021.08.003

Transcriptional profiles of bulk tumors (Cancer Genome Atlas Network, 2012; Guinney et al., 2015; Mlecnik et al., 2016) or single cells (Lee et al., 2020; Li et al., 2017; Zhang et al., 2018, 2020) have been used to classify colorectal cancer (CRC) into subtypes, define their cellular composition, and infer interaction networks between cell types based on the expression of receptor-ligand pairs. However, these studies focused on discrete cell clusters and did not capture the full spectrum of transcriptional programs, which can exist as continuous gradients of program activities within or across clusters (Bielecki et al., 2021; Kotliar et al., 2019). Recently, imaging-based studies have highlighted cellular interaction networks based on recurrent co-localization of different cells in neighborhoods (Schürch et al., 2020). However, these studies were limited by the number of pre-selected markers that resolve key cell types but not their finer features.

Here we developed a systematic approach to discover cell types, their underlying programs, and cellular communities based on single-cell RNA sequencing (scRNA-seq) profiles and applied it to study the distinguishing features of human MMRd and MMRp CRC. We identified 88 cell subsets across immune, stromal, and malignant cells and 204 associated gene expression programs. We revealed multicellular interaction networks based on co-variation of gene program activities in different cell subsets across individual tumors and imaged key molecules for predicted cell subsets and programs to localize these interaction networks in matched tissues from affected individuals. We found stromal remodeling that resulted in a reduction of bone morphogenetic protein (BMP)-producing fibroblasts in MMRd tumors and mislocalization of fibroblast-derived stem cell niche factors throughout the tumor. We discovered an inflammatory interaction network of malignant cells, monocytes, fibroblasts, and neutrophils at the luminal margin of primary MMRd and MMRp tumors and MMRd-specific hotspots of immune activity comprised of chemokine-expressing malignant and non-malignant cells adjacent to activated T cells. Our study demonstrates a path to discovering multicellular interaction networks that underlie immunologic and tumorigenic processes in human cancer.

RESULTS

A comprehensive atlas of cell subsets, programs, and multicellular interaction networks in MMRd and MMRp CRC

To discover how malignant, immune, and stromal cells interact in MMRd and MMRp CRC, we analyzed primary untreated tumors from 34 MMRd and 28 MMRp individuals (with an additional lesion collected for 2 individuals) as well as adjacent normal colon tissue for 36 of the individuals (Figure 1A; Table S1A). We performed droplet-based scRNA-seq on dissociated fresh tissues, retaining 371,223 high-quality cells (STAR Methods), including 168,672 epithelial (non-malignant and malignant), 187,094 immune, and 15,457 stromal cells (Figures S1A and S1B).

We defined cell subsets and transcriptional programs by a two-step graph-clustering approach. First we clustered all cells into 7 major partitions (T/natural killer [NK]/innate lymphoid cell [ILC], B, plasma, mast, myeloid, stroma/endothelial, and epithelial cells). Second, within each partition, we derived clusters (prefix "c") and transcriptional programs (sets of genes with co-varying expression, prefix "p") using consensus non-negative matrix factorization (NMF) (Kotliar et al., 2019; Lee and Seung, 1999; Figure 1B and 1C; Figures S1C and S1D; Tables S2, S3, and S4; STAR Methods). Cell clusters and gene programs were numbered independently of each other. *De novo* identification of programs by NMF enabled several key analyses: (1) simultaneous identification of programs shared across multiple cell types (e.g., proliferation, metabolic, and immune programs), specific to a cell type (e.g., plasmacytoid dendritic cell [pDC] program), and/or expressed in continuous gradients within or across clusters; (2) finding of shared biological properties of malignant cells across individuals despite strong individual-specific transcriptional states (Patel et al., 2014; Puram et al., 2017); and (3) identification of co-varying programs across multiple tumors to find networks of coordinated cells or states that reflect cell interactions or response to a common trigger.

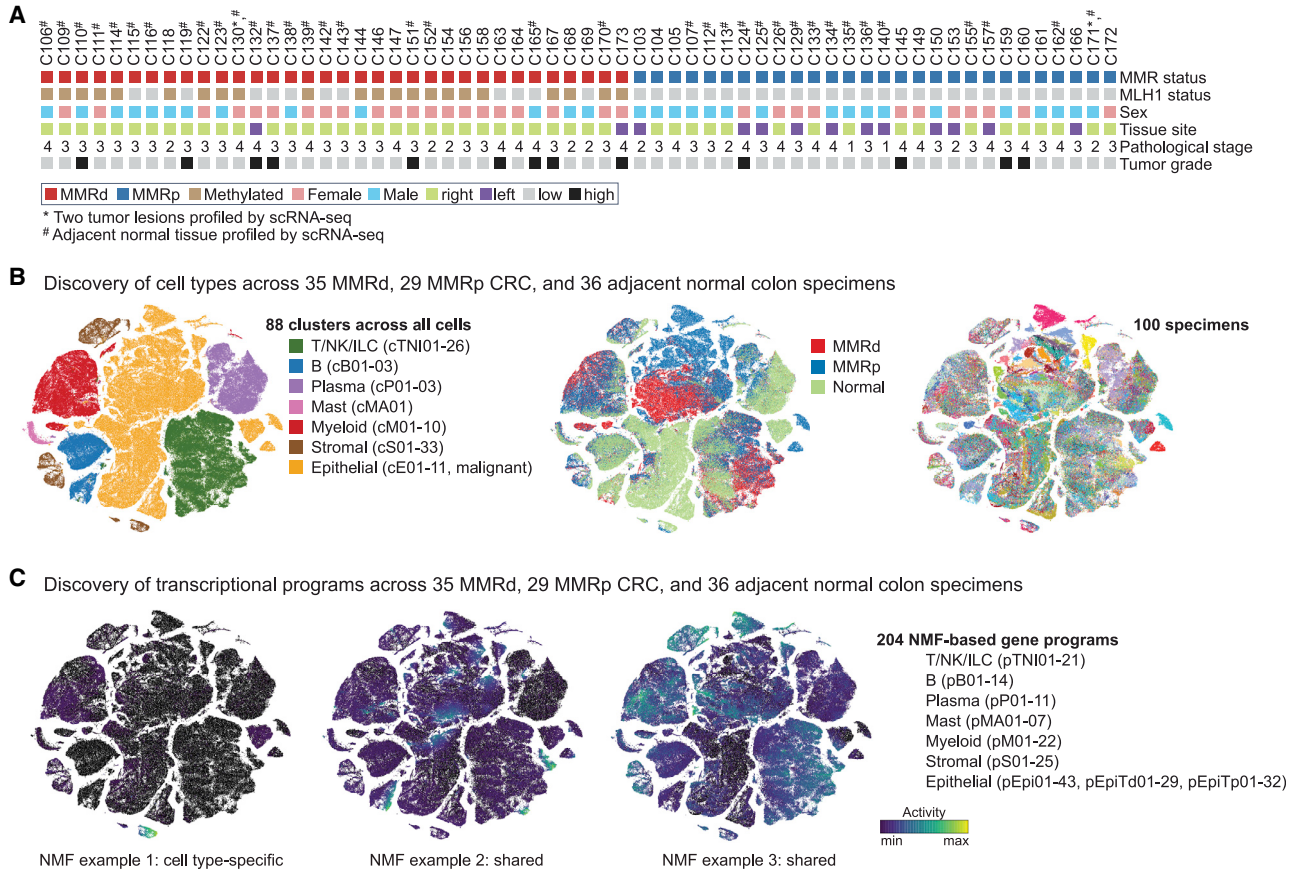


Figure 1. Cohort and atlas of cell subsets and programs in MMRd and MMRp CRC

(A) MMR status and clinical characteristics of primary untreated individuals with CRC.

(B) tSNEs (t-distributed stochastic neighbor embedding) by major cell partitions (left), tissue type (middle), or specimen (right).

(C) NMF-based gene programs can be cell type specific (example 1, pS02-Fibro matrix/stem cell niche) or shared (example 2, pTNI03-proliferation; example 3, pEpi30-ISGs).

See also Figure S1 and Table S1.

Remodeling of the immune cell compartment in MMRd and MMRp CRC

To understand the basis for differential immune responses in CRC, we first compared the immune composition of MMRd and MMRp CRC and normal colon tissue, finding dramatic remodeling between tumor and normal tissue and between MMRd and MMRp tumors. Specifically, 37 of 43 immune cell clusters (manually curated cluster markers in Figure S2A) were differentially abundant as a fraction of all immune cells between tumor (MMRd or MMRp) and normal colon tissue (Figure 2A; Figure S2B; Table S2). Tumors were depleted of immunoglobulin A (IgA)-producing plasma cells, B cells, *IL7R*+ T cells, and γδ-like T cells, and enriched with regulatory T cells (Tregs), monocytes, macrophages, and likely neutrophils relative to normal colon (Figure 2A).

There was a significant expansion of monocytes/macrophages in tumors (Figures 2A and 2B). Monocytes and macrophages upregulated tumor-specific NMF-derived transcriptional programs (Figures 2B and 2C), characterized by genes that can amplify inflammation (*MMP12* and *MMP9* in pM02), recruit myeloid cells (chemokines *CCL2* and *CCL7* in pM10), stimulate growth (growth factors *VEGFA* and *EREG* in pM14), and resolve inflammation (*APOE* in pM06). Myeloid cells from MMRd tumors showed higher activities of programs with genes in glycolysis (pM03), immune-activating alarmins such as *S100A8/9/12* (pM16), and chemokines that attract monocytes and neutrophils (pM20). Overall, monocytes and macrophages were remodeled in tumors and expressed more immune-activating programs in MMRd tumors.

T cell compartment differences between MMRd and MMRp tumors

The predominant change in the immune composition of MMRd versus MMRp tumors was in the T cell compartment (Figures 2A and 2D). Among the clusters enriched in MMRd tumors were *CXCL13*+ T cells and *PDCD1*+ γδ-like T cells, whereas IL17+ T cells were enriched in MMRp tumors (Figure 2A, marked with an asterisk next to the cluster number; Figure S2B). *CXCL13* in T cells has been noted in other CRC and melanoma single-cell studies (Lee et al., 2020; Li et al., 2019; Zhang et al., 2018) and has recently emerged as a marker of human tumor-reactive

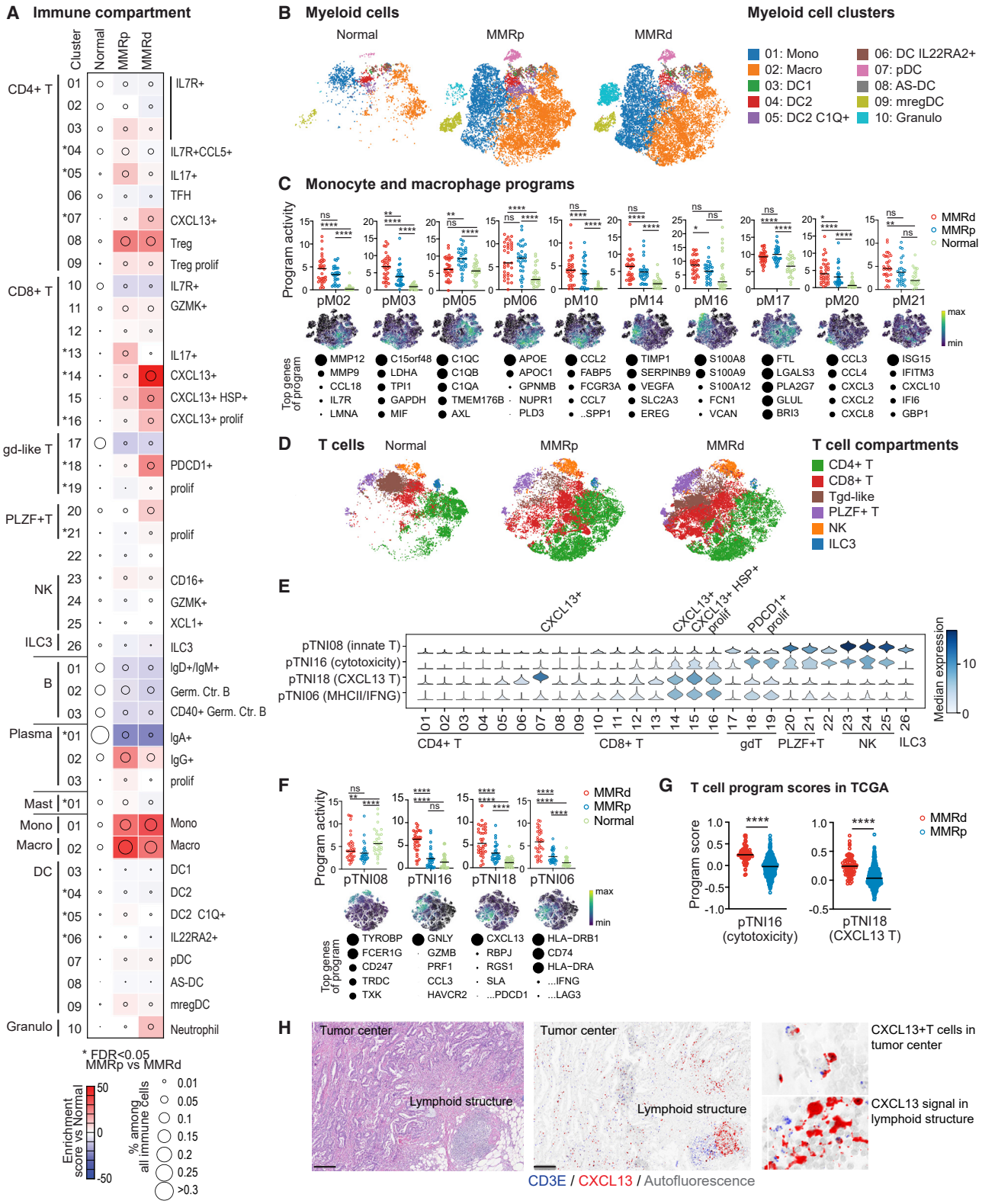


Figure 2. The immune compartment in MMRd and MMRp CRC

(A) Compositional changes in immune cell clusters in MMRp and MMRd tumors relative to adjacent normal tissue. Kruskal-Wallis false discovery rate (FDR) < 0.05 for MMRp versus MMRd are marked with asterisks.

(legend continued on next page)

CD8+ T cells and response to immunotherapy (Ayers et al., 2017; Llosa et al., 2019; Thommen et al., 2018; Oliveira et al., 2021). Thus, we hypothesize that anti-tumor T cell immunity may have developed often in MMRd but rarely in MMRp tumors (Figure S2B).

Programs enriched in MMRd versus MMRp T cells (Table S2E) included two programs (pTNI18 with *CXCL13*, *PDCD1*, *TOX*; pTNI06 with major histocompatibility complex [MHC] class II, *IFNG*, and *LAG3*) with high and moderate activity in T cell receptor [TCR] αβ- and TCRγδ-like T cells, respectively, and one cytotoxicity program (pTNI16) shared among CD8+, γδ-like, *PLZF*+ (*ZBTB16*) T cells and NK cells. *PLZF*+ T cells and NK/ILC3 cells were selectively marked by an innate T cell program (pTNI08) that was reduced in MMRd and MMRp tumors compared with normal tissue (Figures 2E and 2F). We confirmed the higher MMRd activity of the *CXCL13* and cytotoxicity programs (which can be attributed only to the T/NK/ILC partition, allowing us to analyze bulk data; Figure S2C) in three external CRC cohorts (Figure 2G; Figure S2D; Cancer Genome Atlas Network, 2012; Jorissen et al., 2008; Marisa et al., 2013). Thus, in MMRd tumors, subsets of T and NK cells acquire cytolytic properties (*GNLY*, *GZMB*, and *PRF1*), and T cells acquire exhaustion markers associated with chronic stimulation (e.g., *PDCD1*, *TOX*, *LAG3*, and *HAVCR2*).

CXCL13+ T cells localize within MMRd tumors

Given the enrichment of *CXCL13*+ T cells in MMRd tumors and their previous association with immunotherapy responses as well as localization to tertiary lymphoid structures (TLSs) in lung cancer (Thommen et al., 2018), we stained tissue sections from our cohort with RNA probes targeting *CXCL13* and *CD3E*. We found abundant *CXCL13*+ T cells throughout MMRd tumors, outside of TLSs (Figure 2H), which are usually found at the invasive border (Posch et al., 2017). TLS-associated CXCL13 was largely in non-T (*CD3E*-negative) cells in a reticular pattern, consistent with reports of stromal and follicular dendritic cells as sources of *CXCL13* in TLSs (Cyster et al., 2000). *CXCL13*-expressing conventional CD4+ and CD8+ T cells were localized outside of lymphoid structures, but in close proximity to carcinoma cells, consistent with effector activity.

Highly altered endothelial cells in MMRd and MMRp tumors

The stromal compartment was remodeled in both tumor types (Figures 3A and 3B; Figures S3A–S3C; Table S3), with an increase in endothelial cells and pericytes as a fraction of stromal cells (Figure 3C) and a reduction in lymphatic endothelial cells as a fraction of endothelial cells in tumor versus normal cells (Figure 3B). Along with one cluster shared between tumor and normal cells, we found 8 tumor-specific clusters of endothelial cells (marked in red in Figure 3B) with no significant differences between MMRd and MMRp tumors. Quantifying the similarity between endothelial clusters in tumor versus normal colon cells (using partition-based graph abstraction [PAGA]; Wolf et al., 2019), we found altered versions of arterial and venous cells and several clusters that did not map back to normal cells, such as tip cells and proliferating cells (Figure 3D). Interestingly, these proliferating endothelial cells expressed *HIF1A* and *CSF3* (Figure S3A), suggesting metabolic and inflammatory changes.

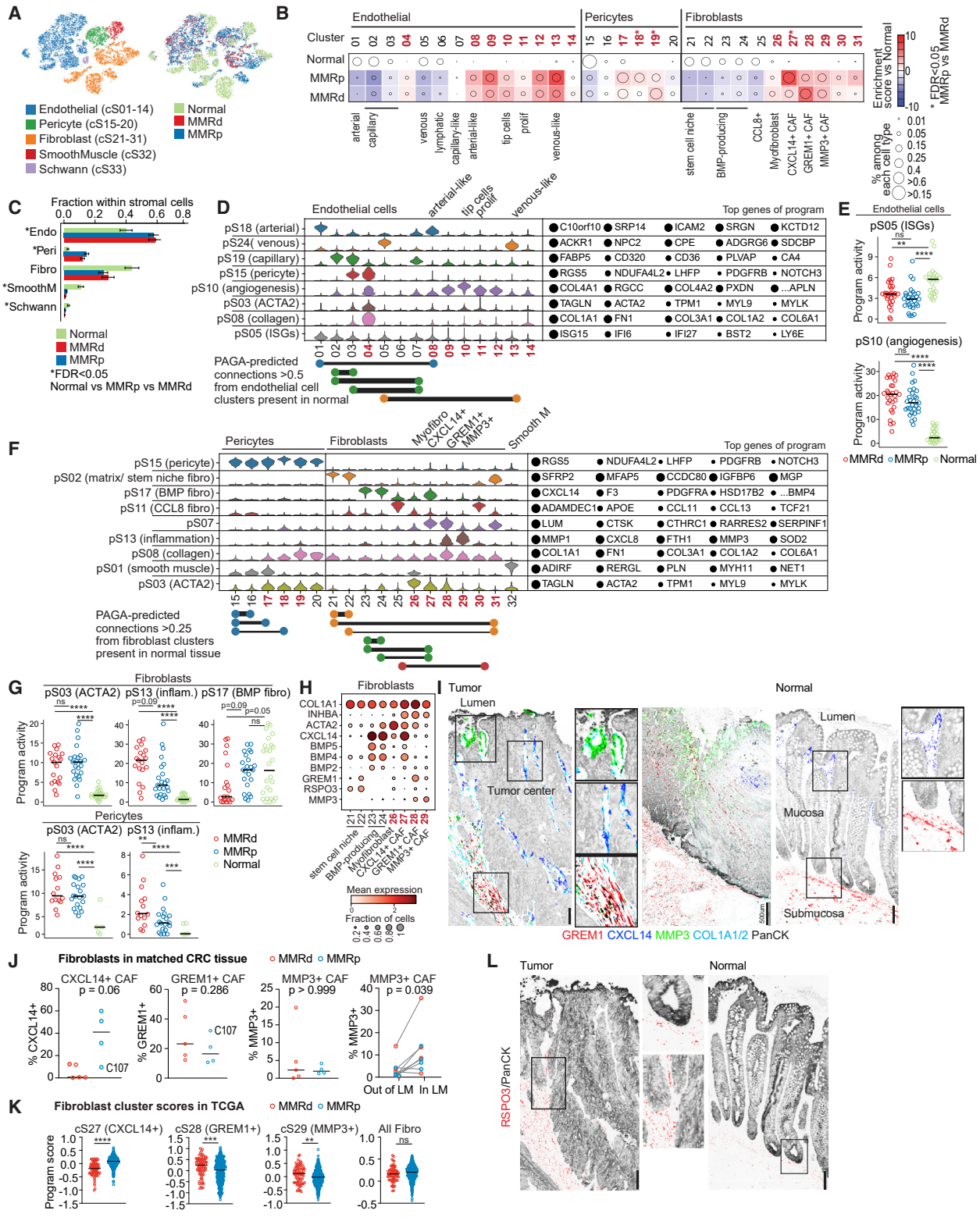
Program pS10 with basement membrane collagens, pro-angiogenic molecules, and a tip cell marker (Table S3) was upregulated across all tumor-specific clusters, whereas a program of interferon-stimulated genes (ISGs)/antigen presentation (pS05) was repressed (Figures 3D and 3E), as observed previously (Lee et al., 2020). Thus, endothelial cells are highly altered in tumors, with more angiogenesis program activity and changes in immune-related gene expression.

Inflammatory fibroblasts localize to the luminal surface of tumors

Fibroblasts partitioned into 11 subsets, with 6 predominant in tumor and 5 in normal colon samples (Figure 3B). Analogous to the previously described myofibroblastic cancer-associated fibroblasts (Dominguez et al., 2019; Elyada et al., 2019; Öhlund et al., 2017), 3 cancer-associated fibroblast subsets (cS26–cS28) (and tumor pericytes) expressed a contractile program (pS03) that included smooth muscle actin (*ACTA2*) (Figures 3F and 3G; Table S3), with one subset (cS26, myofibroblasts) expressing it very highly along with the smooth muscle program (pS01), which was shared with smooth muscle cells and pericytes (Figure 3F).

Two cancer-associated fibroblast (CAF) subsets (cS28 and cS29) expressed an inflammatory program (pS13) (Figure 3F; Table S3) in both tumor types, with higher activity in MMRd tumors (Figure 3G; Table S3). This program, mirroring those of previously described inflammatory CAFs (Dominguez et al., 2019; Elyada et al., 2019; Öhlund et al., 2017) and inflammatory fibroblasts in inflammatory bowel disease (Elmentaite et al., 2020; Haberman et al., 2014; Huang et al., 2019; Olsen et al., 2009; Smillie et al., 2019), included tissue remodeling factors (*MMP1* and *MMP3*) and neutrophil-attracting chemokines (*CXCL8* and *CXCL1*). Tissue staining for *MMP3* and the ubiquitous fibroblast

(B) tSNEs of myeloid cells in all normal and tumor samples.
(C) Activities of selected myeloid gene programs with high activities in monocytes and macrophages. Each dot indicates the 75th percentile of the program activity in the myeloid cells of one specimen. GLME (generalized linear mixed model) FDR: **** ≤ 0.0001, *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, not significant (ns) for > 0.05. tSNEs below show program activities within the myeloid compartment. For each program, the top genes are listed below, with circle size indicating the relative weight of each gene within the program.
(D) tSNEs of the T/NK/ILC cell partition colored by major cell subsets.
(E) pTNI08, pTNI16, pTNI18, and pTNI06 activities within each of the T/NK/ILC cell clusters.
(F) pTNI08, pTNI16, pTNI18, and pTNI06 activities displayed as in (C). GLME FDR reported as in (C).
(G) pTNI16 and pTNI18 gene signature scores in bulk RNA-seq from TCGA-CRC (COADREAD, colon and rectum adenocarcinoma) specimens. Mann-Whitney-Wilcoxon test, ****p ≤ 0.0001.
(H) Localization of *CXCL13*+ T cells in tumor center versus lymphoid structure. Left: H&E. Right: *CD3E* and *CXCL13* RNA *in situ* hybridization (ISH). Scale bar, 200 μm.
See also Figure S2 and Table S2.



(legend on next page)

marker *COL1A1* (Figure 3H) in 8 CRC specimens (4 MMRd and 4 MMRp) revealed that these highly inflammatory fibroblasts were strongly enriched around dilated blood vessels at the colonic luminal margin (LM) of MMRd and MMRp tumors (Figures 3I and 3J; Figure S3D).

BMP-expressing CAFs are reduced in MMRd CRC, whereas CAF-derived stem cell niche factors are abnormally present throughout tumors

To further understand the functional alterations in CAFs, we compared the CAFs with fibroblasts from normal colon tissue based on shared programs and PAGA-based similarity between clusters (Wolf et al., 2019; Figure 3F).

We identified a CAF equivalent (cS27) of BMP-expressing fibroblasts (c23,c24), cells that line normal colon epithelial cells and drive differentiation of epithelial cells through Wnt inhibition via BMPs and Wnt antagonists such as *FRZB*. These may correspond to the PDGFRA-high subset of telocytes in the small intestine (McCarthy et al., 2020). The BMP-expressing CAFs were distinguished from other CAF subsets by *CXCL14* expression (Figure 3H). *CXCL14*+ fibroblasts lined the epithelium in normal tissues and tumors (Figure 3I). A previous bulk RNA-seq study reported reduced *CXCL14* expression in MMRd versus MMRp CRC but suggested that this was due to differential expression in malignant epithelial cells (Mlecnik et al., 2016). Although there was a significant but modest (1.25-fold reduction) reduction in *CXCL14* expression in MMRd versus MMRp malignant cells, *CXCL14* was rarely expressed in malignant cells (~9.2% of MMRp and ~1.5% of MMRd malignant cells), with one exception (MMRp individual C103; Figure S3E). Instead, MMRd individuals (as well as MMRp individual C107, who had high T cell activity) had reduced *CXCL14*+ CAFs (Figure 3B), which we confirmed in imaging-based quantification (Figure 3J) and external cohorts (Jorissen et al., 2008; Marisa et al., 2013; Figure 3K; Figure S3F).

CAFs also contributed expression of stem cell niche factors, such as *RSPO3* and *GREM1*, which were broadly expressed throughout tumors (Figures 3I, left, and 3L, left), in contrast to their crypt-associated expression in normal tissue (Figures 3I, right, and 3L, right). Specifically, in non-neoplastic tissue, *RSPO3* and *GREM1* expression is strictly limited to areas below the bottom of the crypt (Harnack et al., 2019; Karpus et al., 2019; Stzepourginski et al., 2017), most prominently along a distribution similar to that of the *muscularis mucosa* (Figure S3G), as described previously (Davis et al., 2015; Harnack et al., 2019; Worthley et al., 2015). In contrast, *GREM1*+ and *RSPO3*+ cells (Figure 3I,L) were found in stromal bands that reached far upward from the base into the tumor body. In MMRd specimens, these cells also occupied positions similar to the epithelial cell-lining *CXCL14*+ *BMP*-expressing fibroblasts (Figure 3I, middle image). High expression of *RSPO3* drives tumor growth and can arise from *PTPRK-RSPO3* fusion events in a small fraction of human CRC (Hilkens et al., 2017; Seshagiri et al., 2012). Our data suggest that perhaps a more common mechanism to increase access to stem cell niche factors, like *RSPO3*, occurs via spatial redistribution of stromal cells and/or their programs, especially CAFs.

Malignant cells are actively engaged in the immune response

Because malignant cells typically group by individual (in contrast to normal epithelial cells, which cluster by cell subset) (Figure 4A), it can be more challenging to identify their shared properties. We

Figure 3. Stromal remodeling in MMRd and MMRp CRC

(A) tSNEs of stromal cells in all normal and tumor samples.

(B) Compositional changes in endothelial, pericyte, and fibroblast subsets within their respective compartments in MMRp and MMRd tumors relative to adjacent normal tissue. Tumor-enriched clusters are indicated in bold red. Kruskal-Wallis FDR < 0.05 for MMRp versus MMRd are marked with asterisks. cS30 and cS31 are overwhelmingly from two tumors that grew below non-neoplastic tissue and may not be purely tumor derived.

(C) Fraction of stromal cell subsets per tissue type (mean ± SEM with each specimen being one data point). Kruskal-Wallis FDR < 0.05 for normal versus tumor are marked with asterisks.

(D) Activities of selected programs in each of the endothelial cell clusters. Tumor-enriched clusters are indicated in bold red. Top program genes are listed to the right, with circle size indicating the weight of each gene in the program. Key edges (connectivity) between two normal or one normal and one tumor-associated cluster (weights > 0.5, identified by PAGA) are shown below, and colors are matched to programs with high activity in the respective clusters.

(E) Activity of pS05 (ISG) and pS10 (angiogenesis) in all tumor and normal samples. Each point indicates the 75th percentile of the program activity per specimen in endothelial cells. GLME FDR: **** ≤ 0.0001, *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, ns for > 0.05.

(F) Selected programs in fibroblast and pericyte subtypes shown as in (D). Shown below are PAGA-based connectivity weights > 0.25.

(G) Activities of pS03 (ACTA2), pS13 (inflammation), and pS17 (BMP fibro) in fibroblasts and pS03 and pS13 in pericytes, shown as in (E).

(H) Dot plot showing geometric mean expression (log(TP10K+1)) and frequency (dot size) of key genes in selected fibroblast subtypes. *INHBA* distinguishes cancer-associated fibroblasts (CAFs) from fibroblasts in normal tissue. Tumor-enriched clusters are indicated in bold red.

(I) Representative RNA ISH/immunofluorescence (IF) images of tumor show *MMP3*+ fibroblasts at the luminal surface around dilated vessels, *CXCL14*+ fibroblasts lining malignant cells, and *GREM1*+ fibroblasts in stromal bands reaching far into the tumor (left image). In tumors, *GREM1*+ fibroblasts additionally line malignant epithelial cells (center). In normal tissues (right), *GREM1* signal is restricted to in and below the *muscularis mucosa*, while only *CXCL14+* fibroblasts line the epithelial cells. Scale bar, 100 μm (except where annotated).

(J) Quantification of *CXCL14*+, *GREM1*+, and *MMP3*+ CAFs among *COL1A1*/*COL1A2*+ fibroblasts based on whole-slide scans of 5 MMRd and 4 MMRp CRC specimens from (I); Mann-Whitney-Wilcoxon test. Rightmost graph: *MMP3*+ cells among all *COL1A1*/*COL1A2*+ cells outside or inside of the LM (defined as ≤ 360 μm from the luminal border of the tumor); Wilcoxon matched-pairs signed rank test. Only 8 samples are included on the right because one clinical paraffin block did not contain LM.

(K) Gene signature scores of top differentially expressed genes from *CXCL14*+ CAFs, *GREM1*+ CAFs, *MMP3*+ CAFs, and all fibroblasts in bulk RNA-seq from TCGA-CRC (COADREAD). Mann-Whitney-Wilcoxon test: ****p ≤ 0.0001, ***p ≤ 0.001, **p ≤ 0.01, *p ≤ 0.05, ns for > 0.05.

(L) RNA ISH/IF staining for *RSPO3* on consecutive sections to those shown in (I) shows that the *RSPO3* signal is restricted to the crypt base in normal tissue (right image and top inset) but ascends far into the tumor (left image and bottom inset). Scale bar: 100um.

See also Figure S3 and Table S3.

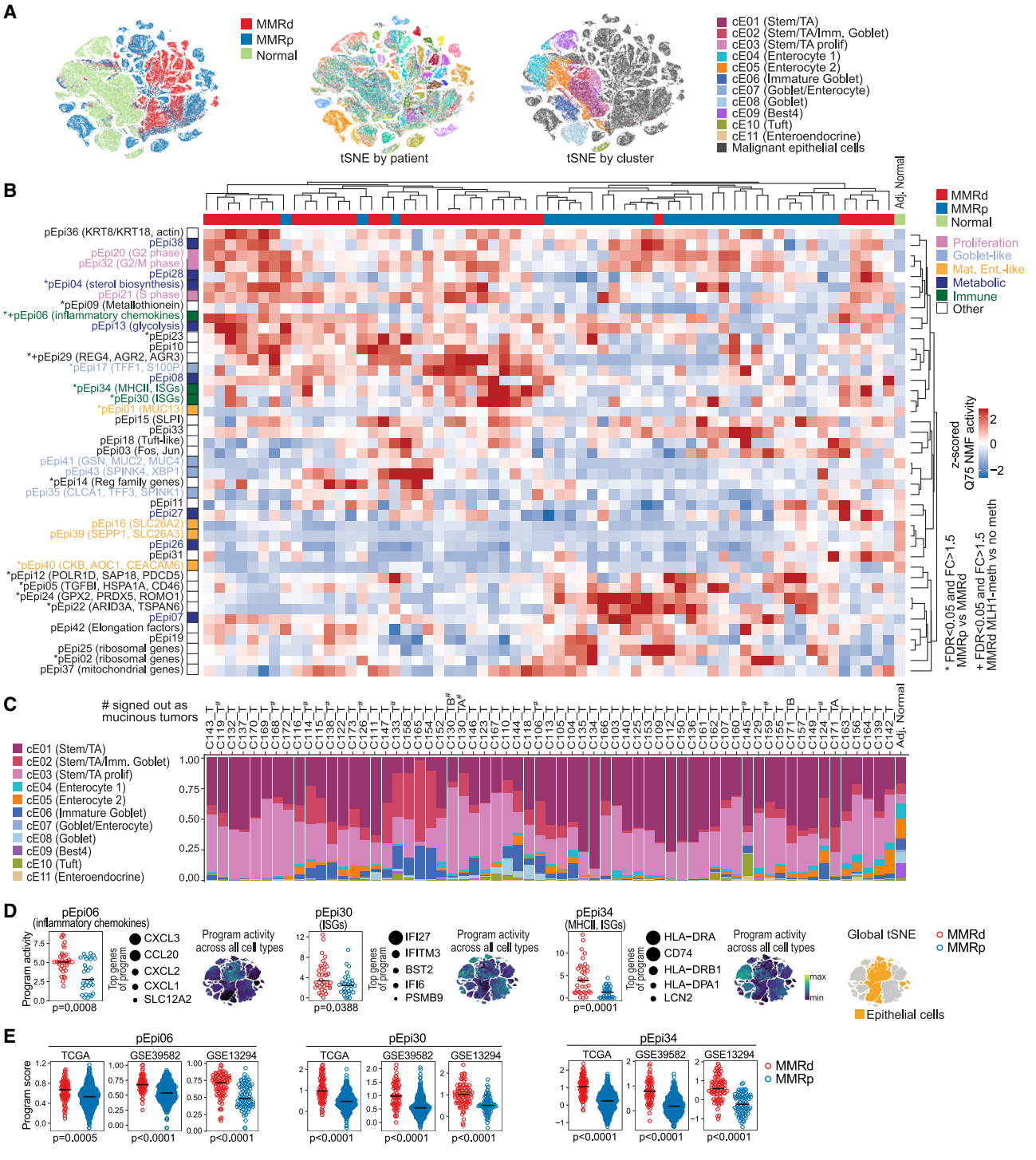


Figure 4. Transcriptional programs in malignant cells differ between MMRd and MMRp CRC

(A) tSNEs of epithelial cells by tissue type (left), individual (middle), and cell type (right).

(B) Heatmap showing the 75th percentile of activities from the 43 malignant programs in malignant cells across CRC specimens (rows centered and *Z* scored), hierarchically clustered using average linkage. Gene program activity in normal epithelial cells is shown for comparison (rightmost column). Significant differences in MMRd versus MMRp are indicated by asterisks (GLME, individual as random effects, MMR status as fixed effect, FDR < 0.05). Significant difference between MLH1 promoter-methylated versus MLH1-non-methylated MMRd specimens is indicated with +.

(legend continued on next page)

therefore derived (STAR Methods) and analyzed the activities of 43 expression programs in malignant cells (denoted pEpi*; Figures S1C, S4A, and S4B; Figure 4B; Table S4), which were not specific to single individuals. We also categorized malignant cells based on similarity to normal colon epithelial cell subtypes to better understand their functional properties (Figure 4C; Figures S4C and S4D; STAR Methods).

Many programs were differentially active between malignant and normal epithelial cells. For example, mature enterocyte programs were reduced (Figure 4B, yellow), and proliferation programs increased (Figure 4B, pink) in malignant versus normal epithelial cells, consistent with the vast majority of malignant cells being classified as stem/transit-amplifying (TA)-like cells (Figure 4C). Among the differentially active programs, 10 showed higher and 6 lower activity in MMRd compared with MMRp, a finding that we validated in 3 external datasets (Figure S4A), along with similar grouping of programs across our cohort and in the Cancer Genome Atlas (TCGA) (Figure 4B; Figure S4B).

In particular, 3 immune-related programs showed elevated activity between MMRd and MMRp malignant cells: an ISG (including interferon γ targets; Table S6) and MHC class II gene program (pEpi34) was more active (3.4-fold) in MMRd than MMRp tumors, an ISG (type I interferon targets; Table S6) and MHC class I gene program (pEpi30) was mildly elevated in MMRd versus MMRp (1.6-fold, also with some activity in normal epithelial cells), and a neutrophil and immune-attracting chemokine program (*CXCL1*, *CXCL2*, *CXCL3*, and *CCL20*) (pEpi06) was higher in MMRd versus MMRp tumors (1.6-fold) and in both tumor types compared with normal cells (Figures 4B, dark green, 4D, and 4E; Table S4). Thus, malignant cells, especially in MMRd tumors, express immune-related programs that may mediate interactions with the immune system.

Co-variation of program activities across individuals predicts multicellular immune hubs

We next hypothesized that some of the changes in gene programs within one cell type may be related to changes in another cell type, either because of a direct effect of one cell type on another or because of a shared signal or neighborhood affecting both cell types in concert.

To find such networks of multi-cellular coordinated programs, we searched for program activities that are correlated across specimens from affected individuals (hereafter called co-varying programs), analyzing MMRd and MMRp separately to better capture differences between the two immunologically disparate tumor types. We calculated pairwise correlations of program activities across each set of samples, using the 22 myeloid, 21 T/NK/ILC cell gene programs, and MMRd- or MMRp-derived malignant epithelial programs (pEpiTd* and pEpiTp*; Figure S4E). Stromal cells were not included because the number of stromal cells per sample was insufficient for a co-variation analysis. Finally, we used graph-based clustering of programs (STAR Methods) to identify 7 co-varying multi-cellular hubs in MMRd and 9 in MMRp samples (Figure 5A; Figure S5A). These hubs consist of multiple programs expressed across the range of cell types, revealing multi-cellular interaction networks.

To identify programs that are similar to each other and, thus, more likely to be triggered by a common mechanism, we computed the overlap of the top genes between programs. This analysis revealed immune, metabolic, and other programs that were similar across cell types (Figure 5B; STAR Methods). We note that co-varying programs (Figure 5A) need not be similar to each other (although they can be) and are often characterized by distinct top gene sets.

To study the interactions between malignant cells and immune cells, we focused on 2 MMRd-derived multicellular hubs (hubs 3 and 6; Figure 5A) in which programs active in immune cells co-varied with immune-related programs active in malignant cells.

Malignant cells, fibroblasts, monocytes, and neutrophils engage in inflammatory responses at the luminal surface of primary MMRd and MMRp tumors

Hub 3 featured inflammatory programs in malignant cells and monocytes that co-varied with a neutrophil program, all of which were highly active in MMRd and MMRp tumors compared with normal tissue (Figure 6A; Figure S6A). Treg and *IL17* T cell programs were also found in the hub. Hub 3 was active in MMRp samples (Figures S5A and S6A), and its programs and their correlations were recapitulated in an external single-cell cohort (Lee et al., 2020; Figure S6A). Based on the similarity of inflammatory myeloid, stromal, and malignant programs, which showed overlapping genes and shared transcription factor predictions, such as nuclear factor κB (NF-κB) and CEPBP (Figure 6B), we also included stromal program pS13 (active in *GREM1*+ and *MMP3*+ CAFs; Figure 6C) in our analysis of hub 3.

To understand the communication pathways driving these malignant/immune/stroma cell interactions, we examined all chemokines and cytokines found within the top genes of the inflammatory and co-varying neutrophil programs (Figure 6D). This analysis suggested concerted attraction of *CXCR1/2*+ neutrophils by malignant cells, *GREM1*+ and *MMP3*+ CAFs, monocytes, and neutrophils expressing cognate chemokines (*CXCL1/2/3/5/6/8*) (Figure 6E). The same chemokines were upregulated in CRC-derived fibroblasts and CRC malignant cells when stimulated *in vitro* with cytokines found in the hub 3 inflammatory monocyte and neutrophil programs, such as *IL1B*

(C) Inferred cell type composition of malignant cells in each tumor specimen, classified by supervised learning trained on non-malignant epithelial cells. Epithelial cell composition in normal tissue is shown for comparison (rightmost bar). Morphologically mucinous tumors are indicated with #. Individual order is the same as in (B).

(D) Selected immune-related transcriptional programs in epithelial cells with significantly different activity in MMRd versus MMRp CRC (GLME FDR < 0.05). For each program, the top genes are shown, and circles indicate the relative weight of each gene in the program. tSNEs show program activities across all cell types (global tSNE); location of epithelial cells is indicated on the right in yellow.

(E) Signature gene scores for programs in (D) in bulk RNA-seq from TCGA-CRC(COADREAD), GSE39582, and GSE13294 specimens. Mann-Whitney-Wilcoxon test: ****$p \leq 0.0001$, ***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$, ns for > 0.05.

See also Figure S4 and Tables S4 and S6.

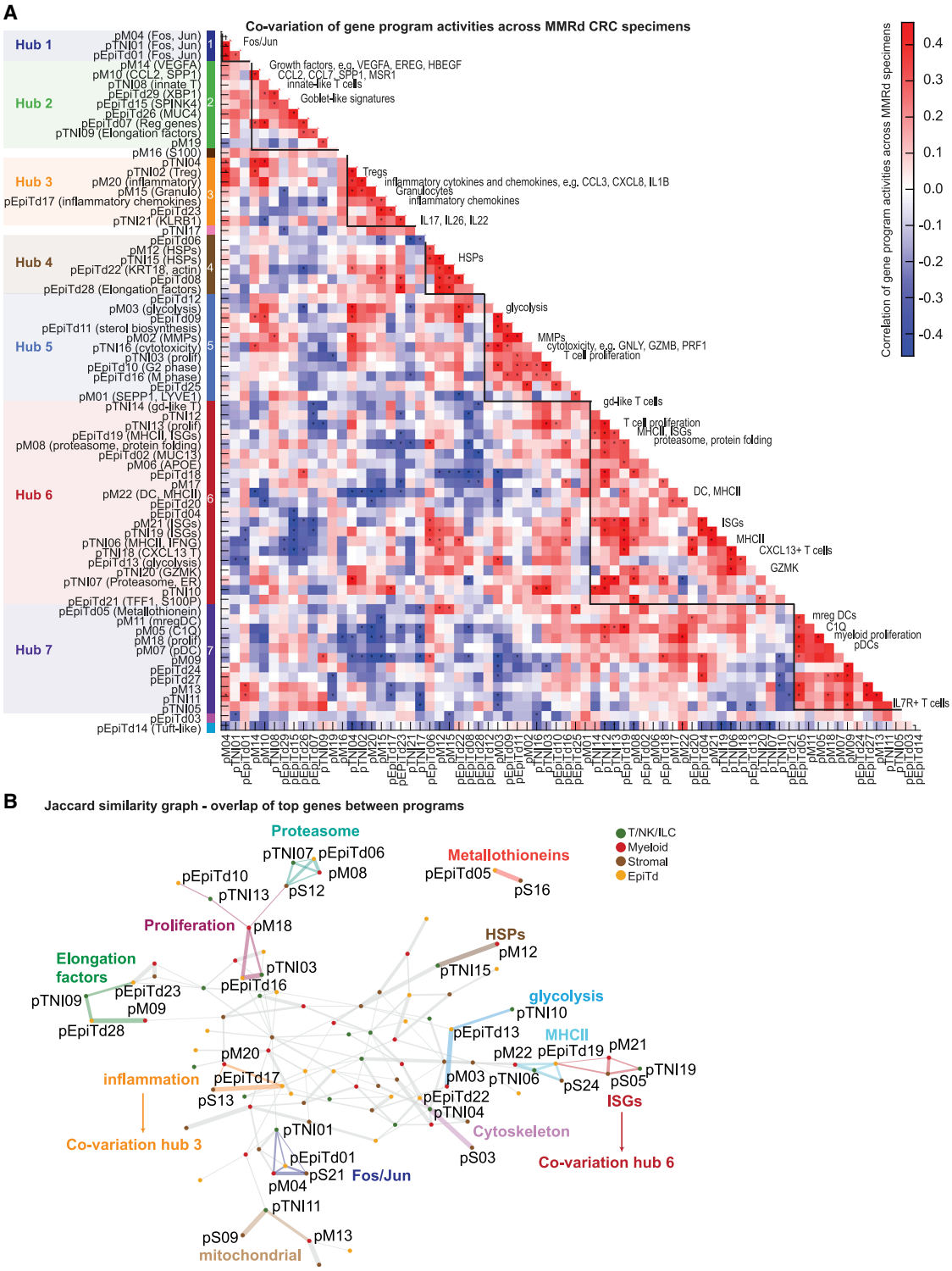


Figure 5. Discovery of multicellular interaction networks in MMRd CRC

(A) Heatmap showing permutation-adjusted pairwise correlation of gene program activities (co-variation score) across MMRd specimens (STAR Methods) using specimen-level activities in T/NK/ILC cell, myeloid, and malignant compartments. Significance is determined using permutation of patient ID and indicated with asterisks (FDR < 0.1). Densely connected modules (hubs) are identified based on graph clustering of significant correlation edges.

(B) Jaccard similarity of gene programs calculated based on the overlap of top weighted genes across T/NK/ILC, stromal, myeloid, and malignant cells. Edge thickness is proportional to program similarity. Edges from selected network neighborhoods are colored and annotated by function.

See also Figure S5 and Table S5.

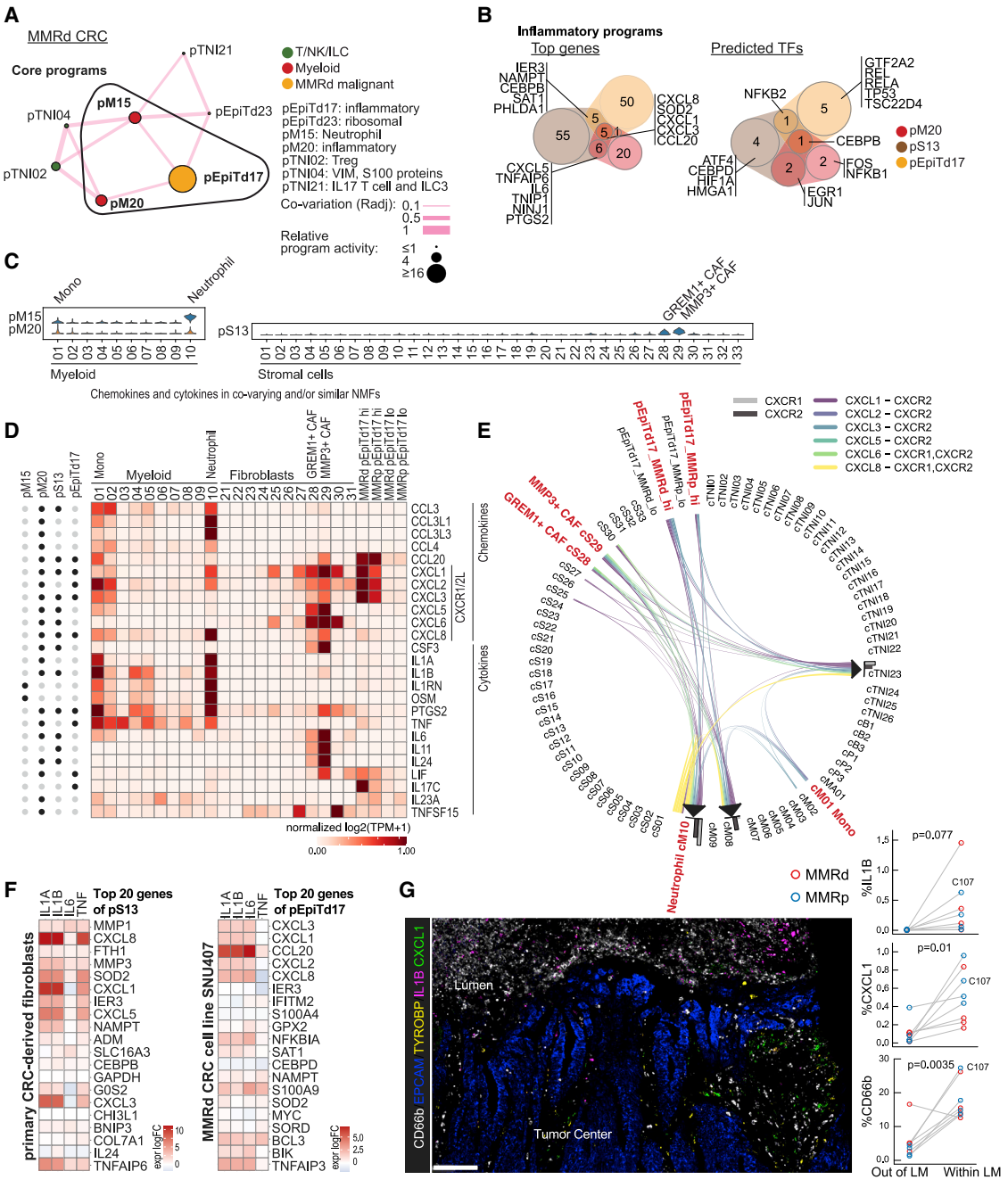


Figure 6. An inflammatory hub at the luminal surface of primary MMRd and MMRp tumors

(A) Inflammatory hub 3 in MMRd specimens. Nodes respresent gene programs, and the size of each node is proportional to the log ratio of the respective mean program activity in MMRd versus normal. Edge thickness is proportional to co-variation.

(B) Venn diagrams showing the overlap of top weighted genes (left) and predicted transcription factors (right) for inflammatory gene programs in myeloid, stromal, and malignant compartments.

(C) Violin plots showing program activities of pM15 and pM20 across myeloid cell clusters and pS13 activity across stromal cell clusters.

(D) Expression level of all chemokines and cytokines present in the top genes of the depicted NMF-based programs (indicated with black dot on the left) across the specified clusters and malignant cells with high versus low pEpiTd17 program activity. Genes are normalized across all cell clusters in the dataset (not only clusters shown).

(E) Interactions between CXCR1/2 and cognate chemokines. Clusters with high activity for the co-varying or similar inflammatory gene programs are marked in red.

(legend continued on next page)

(Figure 6F). Malignant cells, CAFs, monocytes, and neutrophils thus appear to work in concert to recruit myeloid cells and amplify recruitment of myeloid cells via inflammatory cytokines.

To localize this inflammatory hub within the tumor tissue, we stained MMRd and MMRp specimens for markers of neutrophils, myeloid cells, and malignant epithelial cells along with *IL1B* and *CXCL1* transcripts. 7 of 8 examined specimens showed significant accumulations of neutrophils along with *IL1B*+ and *CXCL1*+ cells at the interface of the malignant cells with the colonic lumen (Figure 6G; Figure S6B), particularly at sites with abundant necrosis. Although *CXCL1* was observed in malignant and myeloid cells, a strong *CXCL1* signal was present in cells that are neither myeloid nor epithelial. Although these cells are likely *MMP3*+ CAFs because they express the highest level of *CXCL1* by scRNA-seq (Figure 6D) and are mostly found at the luminal surface (Figure 3I), further imaging studies are needed to confirm this prediction. Given the localization of cells and molecules in this inflammatory hub (Figure 6G) and stromal remodeling (Figure 3I) at the luminal border, we suggest that damage at the luminal edge of primary CRCs may contribute to positive inflammatory feedback loops that drive a myeloid and neutrophil-rich milieu in these tumors.

A coordinated network of *CXCL13*+ T cells with myeloid and malignant cells

Hub 6 (Figures 5A and 7A) was comprised of ISG/MHC class II gene programs expressed in myeloid and malignant cells (likely induced by interferon γ [IFNγ] and driven by IRF/STAT transcription factors; Table S6; Figure 7B), which co-varied with *IFNG*/MHC-II and *CXCL13*/*PDCD1* T cell programs. These T cell programs include markers of activation and exhaustion (Table S2) that are known to mark chronically stimulated tumor-reactive T cells (Gros et al., 2014; Simoni et al., 2018; Thommen et al., 2018).

Importantly, we did not derive this hub in an MMRp-specific analysis (Figure S5) and observed weaker activities of the core programs and reduced connectivity (e.g., the link between malignant pEpiTd19 and T cell pTNI18 programs is lost) of the network when we projected the network onto MMRp tumors in our dataset and an external scRNA-seq dataset (Figure 7A; Figure S7A), consistent with the weaker immunogenicity of MMRp tumors.

To validate the co-activity of ISG/MHC-II malignant and *CXCL13* T cell programs, we performed spatially indexed transcript profiling (GeoMx digital spatial profiling; STAR Methods) of tissue sections from 3 tumors that showed high *CXCL13* T cell program activity in matching scRNA-seq data. We profiled 45 regions of interest (ROIs) per tumor section and further segmented each region into epithelial versus non-epithelial areas (Figure 7C). We observed a positive correlation between ISG expression in malignant epithelial areas and *CXCL13* expression in adjacent non-epithelial areas across all regions per tumor (Figure 7D), further supporting potential interactions between malignant and T cells in this hub.

In addition to inhibitory receptors expressed by exhausted T cells, the malignant ISG/MHC-II program featured inhibitory molecules, including transcripts encoding the enzymes *IDO1* and *CD38*. *IDO1* and *CD38* expression in the malignant ROIs of 4 individuals was comparable with expression measured by scRNA-seq for the same individuals. Moreover, *IDO1* or *CD38* expression was spatially correlated with ISG scores (Figure 7E) in individuals with high scRNA-seq-derived expression of these two genes and the *CXCL13* T cell program. These results show that negative feedback is part of the hub's function and regulated by patient-specific and region-specific factors in each tumor.

CXCL13+ T cells are localized within foci of *CXCL10/CXCL11*-expressing cells throughout the tumor

Given the spatially correlated expression of ISGs in malignant cells with *CXCL13* in non-malignant regions (Figure 7D), we hypothesized that T cells would be spatially organized around cells expressing T cell attracting chemokines. We examined all chemokines in the hub 6 gene programs and found that myeloid, malignant, and stromal ISG programs included the chemokines *CXCL9*, *CXCL10*, and *CXCL11* (Figure 7F) and that their cognate receptor *CXCR3* was upregulated in activated T cells and certain dendritic cell (DC) subsets (Figure 7F). Using our spatially indexed transcriptomics dataset of three highly T cell-infiltrated samples (individuals C107, C110, and C132), we validated this observation by finding that *CXCL13* expression in non-epithelial cells was associated with *CXCR3* ligand expression in malignant cells of the same ROI (Figure 7G).

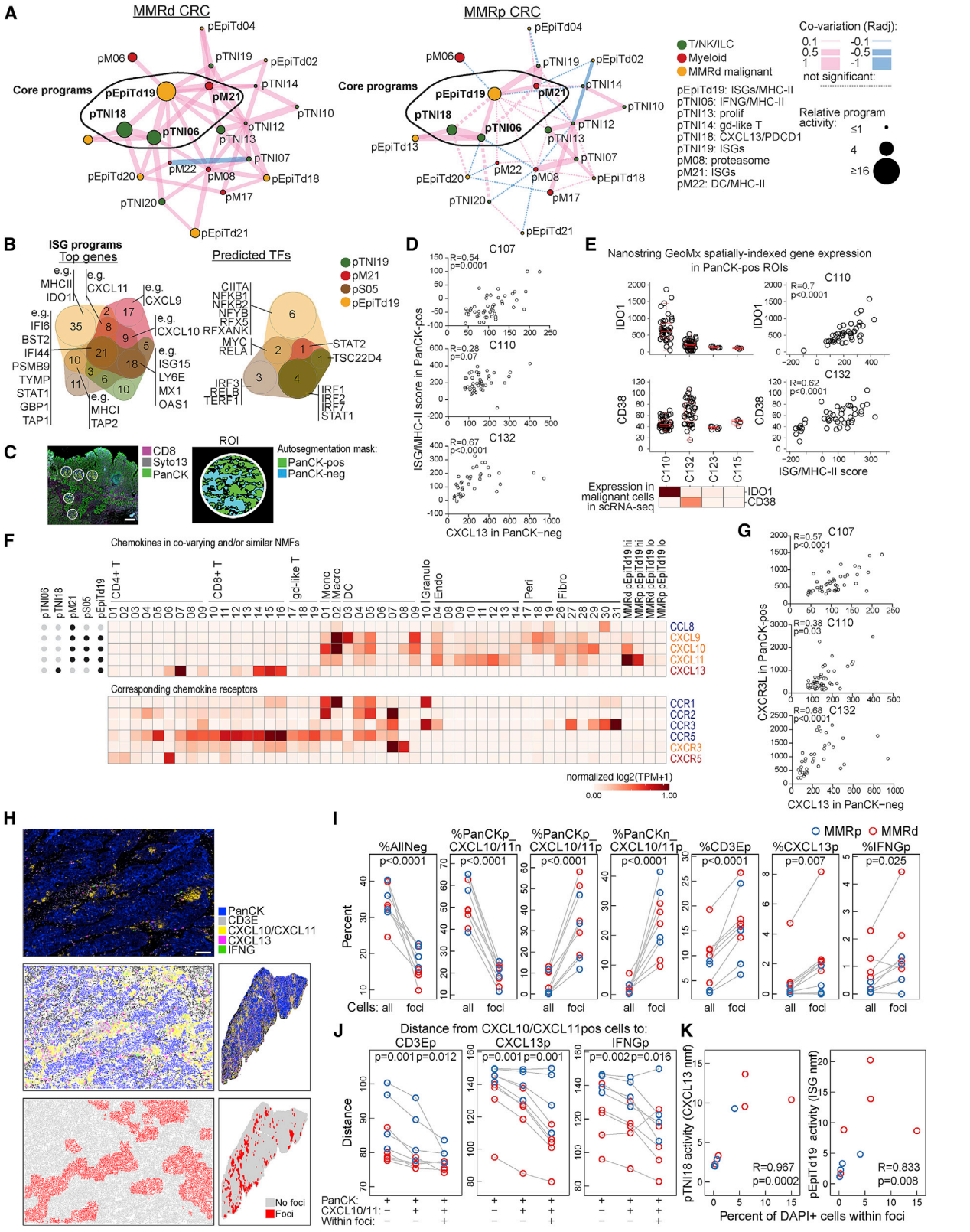
To further validate this spatial association at single-cell resolution, we performed whole-section staining of nine CRC specimens from our scRNA-seq cohort (Figures 7H–7K; Table S7). We found that *CXCL10*/*CXCL11*-positive cells were clustered into large foci enriched for cells expressing *CXCL13* and/or *IFNG* as well as *CD3E*+ T cells (Figures 7H and 7I; Figure S7B; STAR Methods). Interestingly, foci in specimens with high (3 MMRd and 1 MMRp) versus low (2 MMRd and 3 MMRp) *CXCL13*+ T cell program activity tended to show *CXCL10/CXCL11* expression in malignant versus non-epithelial cells, respectively (Figures S7B and S7C), but additional studies are needed to confirm this observation.

Across all samples, *CXCL10*/*CXCL11*+ malignant cells were, on average, closer to *CD3E*+, *CXCL13*+, and *IFNG*+ cells than their *CXCL10*/*CXCL11*-negative counterparts, and these distances were especially small within foci (Figure 7J). Last,

(F) Primary CRC-derived fibroblasts and SNU-407 MMRd CRC cell line were stimulated with 10 ng/ml IL-1A, IL-1B, IL-6, or tumor necrosis factor (TNF) for 14 h or not treated. Transcriptional signatures were determined by RNA-seq. Shown are log fold changes compared with unstimulated cells. Data are representative of 2 independent experiments each.

(G) Representative RNA ISH/IF image shows accumulations of neutrophils (CD66b-IF) and *IL1B* and *CXCL1* ISH signals at the malignant interface (*EPCAM*-ISH) with the colonic lumen. Myeloid cells are marked by *TYROBP*-ISH. Scale bar, 100 μm. Right: quantification of cell phenotypes in 8 CRC specimens (one clinical paraffin block did not contain LM) shows *IL1B*, *CXCL1*, and neutrophil (CD66b) signals enriched in the LM, defined as ≤ 360 μm from the luminal border of the tumor. Paired two-tailed t test. Individual C110 does not show CD66b enrichment at the LM.

See also Figure S6 and Tables S5 and S6.



(legend on next page)

specimens with greater scRNA-seq-derived activity of pTNI18 (*CXCL13* program) and pEpiTd19 (ISG program) had more cells participating in *CXCL10/CXCL11* foci (Figure 7K; Figure S7B). Our findings thus reveal spatially organized foci of activated *IFNG*+ and *CXCL13*+ T cells and *CXCL10/CXCL11*+ myeloid and malignant cells, providing evidence that a positive feedback loop—by which T cell-derived IFNγ induces expression of CXCR3 ligands to attract more T cells—may be critical for formation of these immune cell hotspots within tumors.

DISCUSSION

Tumors are heterogeneous, but the immune cells within tumors are less plastic and exhibit a more limited set of behaviors. Here we identified recurring, spatially organized cell-cell interactions that contribute to a coordinated multi-cellular immune response in MMRd and MMRp tumors.

Our study shows that T cells are organized in structured cell neighborhoods within human tumors. Formation of hotspots likely depends on a positive feedback loop in which T cell-expressed *IFNG* drives induction of *CXCR3* chemokines (as part of the ISG response) that then attract more T cells and other cells. Supporting this notion, recent studies showed that expression of *CXCR3* chemokines in myeloid cells is required for inducing anti-tumor T cell responses following checkpoint inhibitor treatment in mice (Chow et al., 2019; House et al., 2020). Furthermore, several studies have linked the *CXCR3* chemokine system to T cell entry into tissues, including CD8+ T cell recruitment in melanoma (Harlin et al., 2009), viral infection (Nakanishi et al., 2009), and vaccination, in which topical CXCL9 and CXCL10 administration recruited activated T cells into epithelial tissue, even in the absence of antigen (Shin and Iwasaki, 2012). In humans, an IFNγ-induced signature (Ayers et al., 2017; Cristescu et al., 2018), which overlaps with the genes we observed in the programs of hub 6, was associated with favorable response to PD-1 blockade in multiple human tumor types. Furthermore, a recent meta-analysis across 7 tumor types (including CRC) found that clonal TMB and CXCL9/CXCL13 expression were the strongest predictors of a checkpoint inhibitor response (Litchfield et al., 2021). In contrast to the positive feedback loop, persistent ISG hubs in tumors may drive immunosuppression because of negative feedback that upregulates co-inhibitory factors such as *PD1/PDL1*, *Lag3*/MHC-II, *Tim3/LGALS9*, and *IDO1*. Indeed, mechanistic work in the B16 melanoma mouse model suggests that IFNγ can drive a multigenic resistance program (Benci et al., 2016). Whether the positive or negative feedback is dominant at a particular location or time will be important to determine across tumors and treatments.

Another important question is whether these multicellular immune formations are similar to previously observed structures in tissues. TLSs (Sautès-Fridman et al., 2019) are often found below the invasive margin of tumors (Mlecnik et al., 2016), contain germinal center B cells, and have been associated with high T cell activity, favorable prognosis, and effective response to immunotherapy (Cabrita et al., 2020; Coppola et al., 2011; Helmink et al., 2020; Petitprez et al., 2020; Sautès-Fridman et al., 2019). In contrast, hub 6 was found in the tumor center and did not harbor germinal centers, and tumors were depleted of B cells relative to normal colon. A few studies observed aggregates that are not likely to be TLSs. In an early study of melanoma immunity, staining for IFNγ, T cells and PD-L1 showed their spatial proximity in tumors (Taube et al., 2012). Another group observed

Figure 7. A coordinated network of *CXCL13*+ T cells with myeloid and malignant cells expressing ISGs

(A) Hub 6 in MMRd specimens (left) and projected onto MMRp specimens (right). Nodes respresent gene programs, and the size of each node is proportional to the log ratio of the respective mean program activity in MMRd or MMRp versus normal. Edge thickness is proportional to co-variation. Pink lines depict positive and blue lines negative correlations. Non-significant edges are depicted as dotted lines.

(B) Overlap of top weighted genes (left) and predicted transcription factors (right) for ISG programs in T/NK/ILC cell, myeloid, stromal, and malignant compartments.

(C) Image shows a portion of the tissue from individual C110 with regions selected for spatially indexed transcriptomics (GeoMx DSP CTA). ~45 regions of interest (ROIs) per specimen were sampled, and each ROI was auto-segmented into PanCK-positive and -negative regions. Scale bar, 500 μm.

(D) Three CRC specimens with high *CXCL13* activity (C110, C132, and C107) were analyzed by spatially indexed transcriptomics (GeoMx DSP CTA) as described in (C). *CXCL13* signal in PanCK-negative regions was correlated to an ISG/MHC-II signal score (STAR Methods) in the paired PanCK-positive regions (Spearman correlation).

(E) Quantification of the NanoString GeoMx Digital Spatial Profiler (DSP) Cancer Transcriptome Atlas (CTA) assay showing high *IDO1* expression in malignant cells of individual C110 and high *CD38* expression in malignant cells of C132, consistent with scRNA-seq data (heatmap, log2(TP10K+1)). Right: Spearman correlation between *IDO1* (top) or *CD38* (bottom) expression and ISG/MHC-II scores (as calculated in D) in malignant cells of the respective individuals.

(F) All chemokines present in the top genes of the depicted NMF-based programs (indicated by black dot at left) as expressed in the depicted clusters and malignant cells with high versus low pEpiTd19 program activity. Genes are normalized across all cell clusters in the dataset (not only the clusters shown).

(G) GeoMx DSP CTA assay as in (D), showing Spearman correlation of *CXCL13* signal in PanCK-negative regions with CXCR3 ligand expression (i.e., sum of *CXCL9*, *CXCL10*, and *CXCL11*) in the paired PanCK-positive regions.

(H) PanCK-IF, *CD3E*-ISH, *CXCL10/CXCL11*-ISH, *CXCL13*-ISH, and *IFNG*-ISH were performed on 9 tumor tissue slides from different donors (MMRd, n = 5: C110, C123, C132, C139, C144; MMRp, n = 4: C103, C112, C126, C107). Cells were phenotyped using Halo software. An image section from C123 is shown (top), as well as a computational rendering of the same section (center left) and the full slide (center right). Cells were characterized by a 100-μm neighborhood and clustered by their neighborhood features to identify "foci" and "no foci." Scale bar, 100 μm.

(I) Based on the approach in (H), percentages of the indicated phenotype (p, positive; n, negative) among all DAPI+ cells or DAPI+ cells within the foci were calculated. *CXCL10/CXCL11*p, *CD3E*p, *CXCL13*p, and *IFNG*p cells are significantly enriched in foci (paired two-tailed t tests).

(J) Distances were calculated from *CXCL10/CXCL11*-positive cells to the indicated phenotypes (mean distance across 100-μm neighborhoods) outside or inside the foci. If a phenotype was not observed in the 100-μm neighborhood, the distance was set to 150 μm. p values for paired two-tailed t tests are shown.

(K) Percentage of cells within foci (among all DAPI+ cells) was correlated to scRNA-seq-based pTNI18 and pEpiTd19 activities from the respective specimens (Spearman correlation).

See also Figure S7 and Tables S5, S6, and S7.

aggregates of stem-cell-like CD8+ T cells with MHC class II+ cells, which were associated with less progressive kidney cancer in affected individuals (Jansen et al., 2019). A third study showed that vaccination of mice induced an IFNγ/CXCR3-dependent spatial hub of T cells and myeloid cells expressing CXCL10. This hub formed around the vasculature and facilitated entry of circulating T cells into the tissue (Prizant et al., 2021), providing a platform for frequent encounters of T cells with other cells to coordinate immune responses.

The other hub was centered around an inflammatory positive feedback loop between inflammatory CAFs, monocytes, and neutrophils and located at the luminal surface. The luminal surface of colonic tumors has an abnormal epithelial lining, and the tumor mass protrudes into the gut lumen, where it can suffer abrasive injury from colonic contents. Tissue damage could lead to entry of microbial ligands or release of immunostimulatory ligands from dead cells, resulting in inflammation. The inflammatory response may be intertwined with wound-healing responses that can lead to granulation tissue. Interestingly, a recent study in mice showed that damage-induced interleukin-1 (IL-1) can trigger RSPO3 expression in GREM1+ mesenchymal cells (Cox et al., 2021), suggesting that there might be a connection between the inflammatory hub and the transcriptional and spatial remodeling of the stromal cell compartment we observed in human CRC. Indeed, we observed dilated blood vessels at the luminal surface, consistent with previous studies in CRC (Kather et al., 2017), which were surrounded by highly inflammatory fibroblasts expressing matrix metalloproteinases (MMPs), known to contribute to tumor angiogenesis and tissue remodeling (Deryugina and Quigley, 2015). The inflammatory hub furthermore featured the Treg program and a T cell program including IL-17. IL-17 has been shown to promote angiogenesis and tumor expansion in murine models (Charles et al., 2009; Chung et al., 2013; Numasaki et al., 2003), including through CAF activation and recruitment of granulocytes that can support tumor growth (Charles et al., 2009; Chung et al., 2013). Thus, multiple features of the inflammatory hub are implicated in suppression of anti-tumor responses and promotion of tumor growth.

Our study provides a rich dataset of cellular states, gene programs, and their transformations in tumors (such as the profound changes observed in stromal cells) across a relatively large cohort of individuals with CRC. Our predictions of several multicellular hubs based on co-variation of gene programs and subsequent spatial localization of two immune-malignant hubs organizes a large set of cell states and programs into a smaller number of coordinated networks of cells and processes. Understanding the molecular mechanisms underlying these hubs and studying their temporal and spatial regulation upon treatment will be critical for advancing cancer therapy.

LIMITATIONS OF THE STUDY

We prioritized participant safety and tumor purity by not sampling down through the invasive border for scRNA-seq but were able to capture all tumor regions by imaging. Our study was designed to compare the immunologic features of treatment-naive primary human MMRd and MMRp CRC by focusing on cell types and states and cellular interaction networks in these two types of tumors and did not consider tumor genetics or neoantigens. Larger cohorts are needed to cover the heterogeneity of all CRC subtypes. Last, the median follow-up time of our participants is only ~2 years, which limits the possibility of survival analyses.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human tumor specimens
  - Human cell lines
- METHOD DETAILS
  - *In vitro* cytokine stimulation of fibroblasts and CRC cells
  - Tissue processing, CD45 enrichment, and scRNA sequencing
  - RNAscope *in situ* hybridization with co-immunostaining
  - Nanostring GeoMx Digital Spatial Profiling method to measure the expression of ~1500 genes in paired epithelial and non-epithelial regions
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - scRNA-seq pre-processing and quality control filtering
  - Selection of variable genes, dimensionality reduction and clustering
  - Cluster connectivity
  - Cluster assignment by gradient boosting and filtering of potential doublets
  - Classifying malignant cells by gradient boosting
  - Identification of gene expression programs by NMF
  - Identification of shared gene programs in malignant epithelial cells
  - Calculating NMF transcriptional program activity
  - Testing for covarying NMF expression programs
  - Constructing a network of expression program similarity
  - Visualization of single cell profiles
  - Identification of differentially expressed genes
  - Pearson residual calculation in contingency tables
  - Transcription factor target enrichment in gene expression programs
  - Preprocessing of bulk RNA-seq data from fibroblast and cancer cell line stimulation experiment
  - Preprocessing of microarray datasets
  - Preprocessing of bulk RNA-seq from TCGA
  - Calculating gene signature scores in bulk expression datasets
  - Image analysis with HALO
  - Image analysis, neighborhood definition, and clustering

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.08.003.

ACKNOWLEDGMENTS

We thank the Broad Genomics Platform, Broad Flow Cytometry Facility, Pathology and Surgery Departments at MGH and BWH; members of the Villani, Benoist, Brenner, Regev, and Hacohen labs; and Anna Hupalowska. This work was made possible by the generous support of the Evergrande Center for Immunologic Diseases at BWH and HMS (to A.C.A. and A.M.M.), the Klarman Cell Observatory (to O.R.-R. and A. Regev), HHMI (to A. Regev), NIH/NCI R01 CA208756 (to N.H.), and the Arthur, Sandra, and Sarah Irving Fund for Gastrointestinal Immuno-Oncology (to N.H.). The project was also funded in part by federal funds from NCI, NIH Task Order HHSN261100039 under Contract HHSN261201500003I and is part of the HTAPP consortium and NIH HTAN (Human Tumor Atlas Network) consortium paper package. A list of HTAN members is available at: https://humantumoratlas.org/htan-authors/. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. We are also thankful for a research fellowship of the DFG, SU2C Peggy Prescott Early Career Scientist Award PA-6146, SU2C Phillip A. Sharp Award SU2C-AACR-PS-32, BroadIgnite, and NIH/NCI K99CA259511 (to K.P.); NIH/NCI T32CA207021 (to J.H.C.); The Doris Duke Charitable Foundation, The Pancreatic Cancer Action Network, and NIH-NCI K08 CA218420-02, P50 CA127003, and U01 CA224146 (to A.J.A); NIH grant R35 CA197735 (to S.O.); K08CA222663, U54CA225088, a Burroughs Wellcome Fund Career Award for Medical Scientists, the CUMC Louis V. Gerstner, Jr. Scholars Program, and the CUMC Velocity Fellow Program (to B.I.); NIH/NCI R01CA205406, DOD CA160344, and the Project P Fund (to K.N.); U54 CA224068 (to R.B.C. and N.H.); SU2C Colorectal Cancer Dream Team Translational Research Grant SU2C-AACR-DT22-17 (to R.B.C. and N.H., administered by AACR, a scientific partner of SU2C), a Conquer Cancer Foundation of ASCO Career Development Award (to M.G.); and U2C CA233195 (to B.E.J.). N.H. is the David P. Ryan, MD Endowed Chair in Cancer Research, a gift from Arthur, Sandra, and Sarah Irving.

AUTHOR CONTRIBUTIONS

Conceptualization, K.P., M.H., J.H.C., S.S., B.I., A.J.A., A.C.A., O.R.-R., A. Regev., and N.H.; methodology, K.P., M.H., J.H.C., S.S., J.D.P., V.J., M.H., and M.B.; software, M.H., S.S., A.B., K.H.X., C.S.S., C.D.L., and L.T.N.; formal analysis, K.P., M.H., J.H.C., S.S., J.D.P., V.J., A.B., K.H.X., S.X.C., D.R.Z., D.J.L., J.W.R., O.C., and L.T.N.; investigation, K.P., J.H.C., J.D.P., V.J., D.D., W.H.G., D.R.Z., C.A.F., L.T.N., J.W., I.W., M.K., J.W., M.-j.S., J.Y., B.V., A.M.M., and A.K.; resources, D.R.Z., J.W.R., C.A.F., M.L.H., T.E.C., R.B., N.M., J.I., H.K., D.L.B., A.S., J.L.H., S.O., and K.N.; data curation, J.H.C., L.K.D., S.L.D., D.T.F., T.S., K.N., M.G., and A.J.A.; writing – original draft, K.P., M.H., J.H.C., S.S., and N.H.; writing – review & editing, K.P., M.H., J.H.C., S.S., J.D.P., S.X.C., S.O., B.E.J., K.N., M.G., L.T.N., A.J.A., A.C.A., O.R.-R., A. Regev., and N.H.; supervision, K.P., M.H., J.H.C., A. Rotem., S.V., M.G., L.T.N., G.M.B., A.J.A., A.C.A., O.R.-R., A. Regev., and N.H.; project administration, D.D., T.D., T.M., M.N.G.-R., A.S.A., L.K.B., L.K.D., S.L.D., S.T.P., E.K.H., J.K.M., D.T.F., T.S., E.Z.T., and J.J.-V.; funding acquisition, K.P., B.I., B.E.J., R.B.C., A.H.S., V.K.K., K.N., G.M.B., A.J.A., A.C.A., O.R.-R., A. Regev., and N.H.

DECLARATION OF INTERESTS

K.P., M.H., J.H.C., V.K.K., A.J.A., O.R.-R., A. Regev., and N.H. are co-inventors on US Patent Application No. 16/995,425 relating to methods for predicting outcomes and treating colorectal cancer as described in the manuscript. A.J.A. is a Consultant for Oncorus, Arrakis Therapeutics, and Merck and receives research funding from Mirati Therapeutics, Deerfield, and Novo Ventures. R.B.C. receives consulting/speaking fees from Abbvie, Amgen, Array Biopharma/Pfizer, Asana Biosciences, Astex Pharmaceuticals, AstraZeneca, Avidity Biosciences, BMS, C4 Therapeutics, Chugai, Elicio, Fog Pharma, Fount Therapeutics/Kinnate Biopharma, Genentech, Guardant Health, Ipsen, LOXO, Merrimack, Mirati Therapeutics, Natera, N-of-one/QIAGEN, Novartis, nRichDx, Revolution Medicines, Roche, Roivant, Shionogi, Shire, Spectrum Pharmaceuticals, Symphogen, Tango Therapeutics, Taiho, Warp Drive Bio, and Zikani Therapeutics; holds equity in Avidity Biosciences, C4 Therapeutics, Fount Therapeutics/Kinnate Biopharma, nRichDx, and Revolution Medicines; and has received research funding from Asana, AstraZeneca, Lilly, and Sanofi. V.K.K. consults for Pfizer, GSK, Tizona Therapeutics, Celsius Therapeutics, Bicara Therapeutics, Compass Therapeutics, Biocon, and Syngene. G.M.B. has sponsored research agreements with Palleon Pharmaceuticals, Olink Proteomics, and Takeda Oncology; served on SABs for Novartis and Nektar Therapeutics; and received honoraria from Novartis. A.C.A. is a paid consultant for iTeos Therapeutics, and is an SAB member for Tizona Therapeutics, Compass Therapeutics, Zumutor Biologics, and ImmuneOncia, which have interests in cancer immunotherapy. A.C.A.'s interests were reviewed and managed by the BWH and Partners Healthcare in accordance with their conflict of interest policies. M.G. receives research funding from BMS, Merck, and Servier. J.W.R., C.A.F., and M.L.H. are employees of and stockholders for NanoString Technologies Inc. D.R.Z. is a former employee of NanoString Technologies Inc. B.I. is a consultant for Merck and Volastra Therapeutic. R.B. is an UptoDate Author. A. Rotem is an equity holder in Celsius Therapeutics and NucleAI. K.N. has research funding from Janssen, Revolution Medicines, Evergrande Group, Pharmavite; advisory board: Seattle Genetics, BiomX; consulting: X-Biotix Therapeutics; research funding: BMS, Merck, and Servier. B.E.J. is on the SAB for Checkpoint Therapeutics. O.R.-R. is a named inventor on patents and patent applications filed by the Broad Institute in single-cell genomics. From October 2020, O.R.-R. is an employee of Genentech. A. Regev. is a founder of and equity holder in Celsius Therapeutics, an equity holder in Immunitas Therapeutics, and was an SAB member for Thermo Fisher Scientific, Syros Pharmaceuticals, and Neogene Therapeutics until August 1, 2020. From August 1, 2020, A. Regev. is an employee of Genentech. A. Regev. is a named inventor on several patents and patent applications filed by the Broad Institute in single-cell and spatial genomics. N.H. holds equity in BioNTech and is an advisor for Related Sciences/Danger Bio.

Received: December 23, 2020
Revised: May 28, 2021
Accepted: August 3, 2021
Published: August 26, 2021

REFERENCES

André, T., Shiu, K.-K., Kim, T.W., Jensen, B.V., Jensen, L.H., Punt, C., Smith, D., Garcia-Carbonero, R., Benavides, M., Gibbs, P., et al.; KEYNOTE-177 Investigators (2020). Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer. N. Engl. J. Med. *383*, 2207–2218.

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D.R., Albright, A., Cheng, J.D., Kang, S.P., Shankaran, V., et al. (2017). IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. J. Clin. Invest. *127*, 2930–2940.

Benci, J.L., Xu, B., Qiu, Y., Wu, T.J., Dada, H., Twyman-Saint Victor, C., Cucolo, L., Lee, D.S.M., Pauken, K.E., Huang, A.C., et al. (2016). Tumor Interferon Signaling Regulates a Multigenic Resistance Program to Immune Checkpoint Blockade. Cell *167*, 1540–1554.e12.

Bielecki, P., Riesenfeld, S.J., Hütter, J.C., Torlai Triglia, E., Kowalczyk, M.S., Ricardo-Gonzalez, R.R., Lian, M., Amezcua Vesely, M.C., Kroehling, L., Xu, H., et al. (2021). Skin-resident innate lymphoid cells converge on a pathogenic effector state. Nature *592*, 128–132.

Boland, C.R., and Goel, A. (2010). Microsatellite instability in colorectal cancer. Gastroenterology *138*, 2073–2087.e3.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Cabrita, R., Lauss, M., Sanna, A., Donia, M., Skaarup Larsen, M., Mitra, S., Johansson, I., Phung, B., Harbst, K., Vallon-Christersson, J., et al. (2020). Tertiary lymphoid structures improve immunotherapy and survival in melanoma. Nature *577*, 561–565.

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Cardenas, M.A., Prokhnevska, N., and Kissick, H.T. (2021). Organized immune cell interactions within tumors sustain a productive T cell response. Int. Immunol. *33*, 27–37.

Charles, K.A., Kulbe, H., Soper, R., Escorcio-Correia, M., Lawrence, T., Schultheis, A., Chakravarty, P., Thompson, R.G., Kollias, G., Smyth, J.F., et al. (2009). The tumor-promoting actions of TNF-alpha involve TNFR1 and IL-17 in ovarian cancer in mice and humans. J. Clin. Invest. *119*, 3011–3023.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv, arXiv:1603.02754. https://arxiv.org/abs/1603.02754.

Chow, M.T., Ozga, A.J., Servis, R.L., Frederick, D.T., Lo, J.A., Fisher, D.E., Freeman, G.J., Boland, G.M., and Luster, A.D. (2019). Intratumoral Activity of the CXCR3 Chemokine System Is Required for the Efficacy of Anti-PD-1 Therapy. Immunity *50*, 1498–1512.e5.

Chung, A.S., Wu, X., Zhuang, G., Ngu, H., Kasman, I., Zhang, J., Vernes, J.-M., Jiang, Z., Meng, Y.G., Peale, F.V., et al. (2013). An interleukin-17-mediated paracrine network promotes tumor resistance to anti-angiogenic therapy. Nat. Med. *19*, 1114–1123.

Coppola, D., Nebozhyn, M., Khalil, F., Dai, H., Yeatman, T., Loboda, A., and Mulé, J.J. (2011). Unique ectopic lymph node-like structures present in human primary colorectal carcinoma are identified by immune gene array profiling. Am. J. Pathol. *179*, 37–45.

Cox, C.B., Storm, E.E., Kapoor, V.N., Chavarria-Smith, J., Lin, D.L., Wang, L., Li, Y., Kljavin, N., Ota, N., Bainbridge, T.W., et al. (2021). IL-1R1-dependent signaling coordinates epithelial regeneration in response to intestinal damage. Sci. Immunol. *6*, eabe8856.

Cristescu, R., Mogg, R., Ayers, M., Albright, A., Murphy, E., Yearley, J., Sher, X., Liu, X.Q., Lu, H., Nebozhyn, M., et al. (2018). Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science *362*, eaar3593.

Cyster, J.G., Ansel, K.M., Reif, K., Ekland, E.H., Hyman, P.L., Tang, H.L., Luther, S.A., and Ngo, V.N. (2000). Follicular stromal cells and lymphocyte homing to follicles. Immunol. Rev. *176*, 181–193.

Davis, H., Irshad, S., Bansal, M., Rafferty, H., Boitsova, T., Bardella, C., Jaeger, E., Lewis, A., Freeman-Mills, L., Giner, F.C., et al. (2015). Aberrant epithelial GREM1 expression initiates colonic tumorigenesis from cells outside the stem cell niche. Nat. Med. *21*, 62–70.

Delvenne, J.-C., Yaliraki, S.N., and Barahona, M. (2010). Stability of graph communities across time scales. Proc. Natl. Acad. Sci. USA *107*, 12755–12760.

Deryugina, E.I., and Quigley, J.P. (2015). Tumor angiogenesis: MMP-mediated induction of intravasation- and metastasis-sustaining neovasculature. Matrix Biol. *44-46*, 94–112.

Dominguez, C.X., Muller, S., Keerthivasan, S., Koeppen, H., Hung, J., Gierke, S., Breart, B., Foreman, O., Bainbridge, T.W., Castiglioni, A., et al. (2019). Single-cell RNA sequencing reveals stromal evolution into LRRC15+ myofibroblasts as a determinant of patient response to cancer immunotherapy. Cancer Discov. *10*, 232–253.

Elmentaite, R., Ross, A.D.B., Roberts, K., James, K.R., Ortmann, D., Gomes, T., Nayak, K., Tuck, L., Pritchard, S., Bayraktar, O.A., et al. (2020). Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn's Disease. Dev. Cell *55*, 771–783.e5.

Elyada, E., Bolisetty, M., Laise, P., Flynn, W.F., Courtois, E.T., Burkhart, R.A., Teinor, J.A., Belleau, P., Biffi, G., Lucito, M.S., et al. (2019). Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. Cancer Discov. *9*, 1102–1123.

Esmailian, P., and Jalili, M. (2015). Community Detection in Signed Networks: the Role of Negative ties in Different Scales. Sci. Rep. *5*, 14339.

Griffiths, J.A., Scialdone, A., and Marioni, J.C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. Mol. Syst. Biol. *14*, e8046.

Gros, A., Robbins, P.F., Yao, X., Li, Y.F., Turcotte, S., Tran, E., Wunderlich, J.R., Mixon, A., Farid, S., Dudley, M.E., et al. (2014). PD-1 identifies the patient-specific $CD8^+$ tumor-reactive repertoire infiltrating human tumors. J. Clin. Invest. *124*, 2246–2259.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356.

Haberman, Y., Tickle, T.L., Dexheimer, P.J., Kim, M.-O., Tang, D., Karns, R., Baldassano, R.N., Noe, J.D., Rosh, J., Markowitz, J., et al. (2014). Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J. Clin. Invest. *124*, 3617–3633.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat. Methods *14*, 955–958.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., Kim, E., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. *46* (D1), D380–D386.

Harlin, H., Meng, Y., Peterson, A.C., Zha, Y., Tretiakova, M., Slingluff, C., McKee, M., and Gajewski, T.F. (2009). Chemokine expression in melanoma metastases associated with CD8+ T-cell recruitment. Cancer Res. *69*, 3077–3085.

Harnack, C., Berger, H., Antanaviciute, A., Vidal, R., Sauer, S., Simmons, A., Meyer, T.F., and Sigal, M. (2019). R-spondin 3 promotes stem cell recovery and epithelial regeneration in the colon. Nat. Commun. *10*, 4368.

Helmink, B.A., Reddy, S.M., Gao, J., Zhang, S., Basar, R., Thakur, R., Yizhak, K., Sade-Feldman, M., Blando, J., Han, G., et al. (2020). B cells and tertiary lymphoid structures promote immunotherapy response. Nature *577*, 549–555.

Hilkens, J., Timmer, N.C., Boer, M., Ikink, G.J., Schewe, M., Sacchetti, A., Koppens, M.A.J., Song, J.-Y., and Bakker, E.R.M. (2017). RSPO3 expands intestinal stem cell and niche compartments and drives tumorigenesis. Gut *66*, 1095–1105.

House, I.G., Savas, P., Lai, J., Chen, A.X.Y., Oliver, A.J., Teo, Z.L., Todd, K.L., Henderson, M.A., Giuffrida, L., Petley, E.V., et al. (2020). Macrophage-Derived CXCL9 and CXCL10 Are Required for Antitumor Immune Responses Following Immune Checkpoint Blockade. Clin. Cancer Res. *26*, 487–504.

Huang, B., Chen, Z., Geng, L., Wang, J., Liang, H., Cao, Y., Chen, H., Huang, W., Su, M., Wang, H., et al. (2019). Mucosal Profiling of Pediatric-Onset Colitis and IBD Reveals Common Pathogenics and Therapeutic Pathways. Cell *179*, 1160–1176.e24.

Jansen, C.S., Prokhnevska, N., Master, V.A., Sanda, M.G., Carlisle, J.W., Bilen, M.A., Cardenas, M., Wilkinson, S., Lake, R., Sowalsky, A.G., et al. (2019). An intra-tumoral niche maintains and differentiates stem-like CD8 T cells. Nature *576*, 465–470.

Jorissen, R.N., Lipton, L., Gibbs, P., Chapman, M., Desai, J., Jones, I.T., Yeatman, T.J., East, P., Tomlinson, I.P.M., Verspaget, H.W., et al. (2008). DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. Clin. Cancer Res. *14*, 8061–8069.

Karpus, O.N., Westendorp, B.F., Vermeulen, J.L.M., Meisner, S., Koster, J., Muncan, V., Wildenberg, M.E., and van den Brink, G.R. (2019). Colonic CD90+ Crypt Fibroblasts Secrete Semaphorins to Support Epithelial Growth. Cell Rep. *26*, 3698–3708.e5.

Kather, J.N., Zöllner, F.G., Schad, L.R., Melchers, S.M., Sinn, H.-P., Marx, A., Gaiser, T., and Weis, C.-A. (2017). Identification of a characteristic vascular belt zone in human colorectal cancer. PLoS ONE *12*, e0171378.

Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., and Sabeti, P.C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. eLife *8*, e43803.

Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., et al. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N. Engl. J. Med. *372*, 2509–2520.

Le, D.T., Durham, J.N., Smith, K.N., Wang, H., Bartlett, B.R., Aulakh, L.K., Lu, S., Kemberling, H., Wilt, C., Luber, B.S., et al. (2017). Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science *357*, 409–413.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791.

Lee, H.-O., Hong, Y., Etlioglu, H.E., Cho, Y.B., Pomella, V., Van den Bosch, B., Vanhecke, J., Verbandt, S., Hong, H., Min, J.-W., et al. (2020). Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat. Genet. *52*, 594–603.

Li, S.K.H., and Martin, A. (2016). Mismatch Repair and Colon Cancer: Mechanisms and Therapies Explored. Trends Mol. Med. *22*, 274–289.

Li, Y., and Ngom, A. (2013). The non-negative matrix factorization toolbox for biological data mining. Source Code Biol. Med. *8*, 10.

Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon, L.K., Tan, W.S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. *49*, 708–718.

Li, H., van der Leun, A.M., Yofe, I., Lubling, Y., Gelbard-Solodkin, D., van Akkooi, A.C.J., van den Braber, M., Rozeman, E.A., Haanen, J.B.A.G., Blank, C.U., et al. (2019). Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. Cell *176*, 775–789.e18.

Li, B., Gould, J., Yang, Y., Sarkizova, S., Tabaka, M., Ashenberg, O., Rosen, Y., Slyper, M., Kowalczyk, M.S., Villani, A.-C., et al. (2020). Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. Nat. Methods *17*, 793–798.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Litchfield, K., Reading, J.L., Puttick, C., Thakkar, K., Abbosh, C., Bentham, R., Watkins, T.B.K., Rosenthal, R., Biswas, D., Rowan, A., et al. (2021). Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell *184*, 596–614.e14.

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database (Oxford) *2015*, bav095.

Llosa, N.J., Luber, B., Tam, A.J., Smith, K.N., Siegel, N., Awan, A.H., Fan, H., Oke, T., Zhang, J., Domingue, J., et al. (2019). Intratumoral Adaptive Immunosuppression and Type 17 Immunity in Mismatch Repair Proficient Colorectal Tumors. Clin. Cancer Res. *25*, 5250–5259.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., and Marioni, J.C.; participants in the 1st Human Cell Atlas Jamboree (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., et al. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. *10*, e1001453.

McCarthy, N., Manieri, E., Storm, E.E., Saadatpour, A., Luoma, A.M., Kapoor, V.N., Madha, S., Gaynor, L.T., Cox, C., Keerthivasan, S., et al. (2020). Distinct Mesenchymal Cell Populations Generate the Essential Intestinal BMP Signaling Gradient. Cell Stem Cell *26*, 391–402.e5.

Mlecnik, B., Bindea, G., Angell, H.K., Maby, P., Angelova, M., Tougeron, D., Church, S.E., Lafontaine, L., Fischer, M., Fredriksen, T., et al. (2016). Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability. Immunity *44*, 698–711.

Nakanishi, Y., Lu, B., Gerard, C., and Iwasaki, A. (2009). CD8(+) T lymphocyte mobilization to virus-infected tissue requires CD4(+) T-cell help. Nature *462*, 510–513.

Numasaki, M., Fukushi, J., Ono, M., Narula, S.K., Zavodny, P.J., Kudo, T., Robbins, P.D., Tahara, H., and Lotze, M.T. (2003). Interleukin-17 promotes angiogenesis and tumor growth. Blood *101*, 2620–2627.

Öhlund, D., Handly-Santana, A., Biffi, G., Elyada, E., Almeida, A.S., Ponz-Sarvise, M., Corbo, V., Oni, T.E., Hearn, S.A., Lee, E.J., et al. (2017). Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. J. Exp. Med. *214*, 579–596.

Oliveira, G., Stromhaug, K., Klaeger, S., Kula, T., Frederick, D.T., Le, P.M., Forman, J., Huang, T., Li, S., Zhang, W., et al. (2021). Phenotype, specificity and avidity of antitumour CD8 $^{+}$ T cells in melanoma. Nature *596*, 119–125.

Olsen, J., Gerds, T.A., Seidelin, J.B., Csillag, C., Bjerrum, J.T., Troelsen, J.T., and Nielsen, O.H. (2009). Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. Inflamm. Bowel Dis. *15*, 1032–1038.

Overman, M.J., Lonardi, S., Wong, K.Y.M., Lenz, H.-J., Gelsomino, F., Aglietta, M., Morse, M.A., Van Cutsem, E., McDermott, R., Hill, A., et al. (2018). Durable Clinical Benefit With Nivolumab Plus Ipilimumab in DNA Mismatch Repair-Deficient/Microsatellite Instability-High Metastatic Colorectal Cancer. J. Clin. Oncol. *36*, 773–779.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Petitprez, F., de Reyniès, A., Keung, E.Z., Chen, T.W.-W., Sun, C.-M., Calderaro, J., Jeng, Y.-M., Hsiao, L.-P., Lacroix, L., Bougoüin, A., et al. (2020). B cells are associated with survival and immunotherapy response in sarcoma. Nature *577*, 556–560.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098.

Posch, F., Silina, K., Leibl, S., Mündlein, A., Moch, H., Siebenhüner, A., Samaras, P., Riedl, J., Stotz, M., Szkandera, J., et al. (2017). Maturation of tertiary lymphoid structures and recurrence of stage II and III colorectal cancer. OncoImmunology *7*, e1378844.

Prizant, H., Patil, N., Negatu, S., Bala, N., McGurk, A., Leddon, S.A., Hughson, A., McRae, T.D., Gao, Y.-R., Livingstone, A.M., et al. (2021). CXCL10+ peripheral activation niches couple preferred sites of Th1 entry with optimal APC encounter. Cell Rep. *36*, 109523.

Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell *171*, 1611–1624.e24.

Rahman, M., Jackson, L.K., Johnson, W.E., Li, D.Y., Bild, A.H., and Piccolo, S.R. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. Bioinformatics *31*, 3666–3672.

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., et al.; Cancer Genome Atlas Research Network (2018). Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Rep. *23*, 181–193.e7.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Sato, T., Stange, D.E., Ferrante, M., Vries, R.G.J., Van Es, J.H., Van den Brink, S., Van Houdt, W.J., Pronk, A., Van Gorp, J., Siersema, P.D., and Clevers, H. (2011). Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology *141*, 1762–1772.

Sautès-Fridman, C., Petitprez, F., Calderaro, J., and Fridman, W.H. (2019). Tertiary lymphoid structures in the era of cancer immunotherapy. Nat. Rev. Cancer *19*, 307–325.

Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D.R., et al. (2020). Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. Cell *182*, 1341–1359.e19.

Seshagiri, S., Stawiski, E.W., Durinck, S., Modrusan, Z., Storm, E.E., Conboy, C.B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B.S., et al. (2012). Recurrent R-spondin fusions in colon cancer. Nature *488*, 660–664.

Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., et al. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell *166*, 1308–1323.e30.

Shin, H., and Iwasaki, A. (2012). A vaccine strategy that protects against genital herpes by establishing local memory T cells. Nature *491*, 463–467.

Simoni, Y., Becht, E., Fehlings, M., Loh, C.Y., Koo, S.-L., Teng, K.W.W., Yeong, J.P.S., Nahar, R., Zhang, T., Kared, H., et al. (2018). Bystander $CD8^+$ T cells are abundant and phenotypically distinct in human tumour infiltrates. Nature *557*, 575–579.

Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., et al. (2019). Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell *178*, 714–730.e22.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

Stzepourginski, I., Nigro, G., Jacob, J.-M., Dulauroy, S., Sansonetti, P.J., Eberl, G., and Peduto, L. (2017). CD34+ mesenchymal cells are a major component of the intestinal stem cells niche at homeostasis and after injury. Proc. Natl. Acad. Sci. USA *114*, E506–E513.

Taube, J.M., Anders, R.A., Young, G.D., Xu, H., Sharma, R., McMiller, T.L., Chen, S., Klein, A.P., Pardoll, D.M., Topalian, S.L., and Chen, L. (2012). Colocalization of inflammatory response with B7-h1 expression in human melanocytic lesions supports an adaptive resistance mechanism of immune escape. Sci. Transl. Med. *4*, 127ra37.

Thommen, D.S., Koelzer, V.H., Herzig, P., Roller, A., Trefny, M., Dimeloe, S., Kiialainen, A., Hanhart, J., Schill, C., Hess, C., et al. (2018). A transcriptionally and functionally distinct $PD\text{-}1^+$ $CD8^+$ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. Nat. Med. *24*, 994–1004.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al.; Cancer Genome Atlas Research Network (2018). The Immune Landscape of Cancer. Immunity *48*, 812–830.e14.

Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. *15*, 3221–3245.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59.

Worthley, D.L., Churchill, M., Compton, J.T., Tailor, Y., Rao, M., Si, Y., Levin, D., Schwartz, M.G., Uygur, A., Hayakawa, Y., et al. (2015). Gremlin 1 identifies a skeletal stem cell with bone, cartilage, and reticular stromal potential. Cell *160*, 269–284.

Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J.Y., et al. (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature *564*, 268–272.

Zhang, L., Li, Z., Skrzypczynska, K.M., Fang, Q., Zhang, W., O'Brien, S.A., He, Y., Wang, L., Zhang, Q., Kim, A., et al. (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell *181*, 442–459.e29.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
IF: mouse anti-CD66b (G10F5)	BioLegend	Cat#305102; RRID:AB_314494
IF: mouse anti-PanCK (AE1/AE3)	Agilent	Cat#M3515; RRID:AB_2132885
IF: Opal Polymer HRP Ms + Rb	Akoya Biosciences	Cat#ARH1001EA; RRID: N/A
NanoString: mouse anti-CD8a-AlexaFluor 647	BioLegend	Cat#372906; RRID:AB_2650712
Biological samples		
Human colorectal cancer specimens from surgical resections	Prospective Collection at Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH)	Table S1
Human adjacent normal colon specimens from surgical resections	Prospective Collection at Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH)	Table S1
Chemicals, peptides, and recombinant proteins		
FISH: RNAscope 2.5 LS Protease III	Advanced Cell Diagnostics	Cat#322102; RRID: N/A
FISH: RNAscope 2.5 LS Hydrogen Peroxide	Advanced Cell Diagnostics	Cat#322101; RRID: N/A
FISH: RNAscope 2.5 LS Rinse	Advanced Cell Diagnostics	Cat#322103; RRID: N/A
FISH: RNAscope LS Multiplex AMP 1	Advanced Cell Diagnostics	Cat#322801; RRID: N/A
FISH: RNAscope LS Multiplex AMP 2	Advanced Cell Diagnostics	Cat#322802; RRID: N/A
FISH: RNAscope LS Multiplex AMP 3	Advanced Cell Diagnostics	Cat#322803; RRID: N/A
FISH: RNAscope LS Multiplex HRP C1	Advanced Cell Diagnostics	Cat#322804; RRID: N/A
FISH: RNAscope LS Multiplex HRP C2	Advanced Cell Diagnostics	Cat#322805; RRID: N/A
FISH: RNAscope LS Multiplex HRP C3	Advanced Cell Diagnostics	Cat#322806; RRID: N/A
FISH: RNAscope LS Multiplex HRP Blocker	Advanced Cell Diagnostics	Cat#322807; RRID: N/A
FISH: RNAscope Multiplex TSA Buffer	Advanced Cell Diagnostics	Cat#322809; RRID: N/A
FISH/IF: DAPI	Sigma Aldrich	Cat#D9542-10MG; RRID: N/A
FISH/IF: BOND Epitope Retrieval Solution 2-1L (RTU)	Leica Biosystems	Cat#AR9640; RRID: N/A
FISH/IF: BOND Dewax Solution – 1L (RTU)	Leica Biosystems	Cat#AR9222; RRID: N/A
FISH/IF: BOND Wash Solution 10X Concentrate – 1L	Leica Biosystems	Cat#AR9590; RRID: N/A
FISH/IF: Thermo Scientific Reagent Grade Deionized Water	ThermoFisher	Cat#23-751628; RRID: N/A
IF: Antibody Diluent / Block	Akoya Biosciences	Cat#ARD1001EA; RRID: N/A
IF: Plus Automation Amplification Diluent	Akoya Biosciences	Cat#FP1609; RRID: N/A
NanoString: RNase AWAY Surface Decontaminant	ThermoFisher	Cat#7000TS1; RRID: N/A
NanoString: Water, Milli-Q, DEPC-Treated	Broad Institute SQM	Cat#DEPCH2O20L; RRID: N/A
NanoString: Formalin 10% Prefill/Label	Patterson Veterinary Supply	Cat#07-831-8994; RRID: N/A
NanoString: Formamide (Deionized)	ThermoFisher	Cat#AM9342; RRID: N/A
NanoString: UltraPure SSC, 20X	ThermoFisher	Cat#15557044; RRID: N/A
NanoString: Proteinase K Solution (20 mg/mL)	ThermoFisher	Cat#AM2548; RRID: N/A
NanoString: eBioscience IHC Antigen Retrieval Solution - High pH (10X)	ThermoFisher	Cat#00-4956-58; RRID: N/A
NanoString: Tris base	Sigma Aldrich	Cat#10708976001; RRID: N/A
NanoString: Glycine	Sigma Aldrich	Cat#G7126; RRID: N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
NanoString: (R)-(+)-Limonene	Sigma Aldrich	Cat#183164; RRID: N/A
NanoString: Tween-20 solution, 10%	Teknova Inc	Cat#100216-360; RRID: N/A
NanoString: Buffer S	NanoString Technologies	Cat#N/A; RRID: N/A
NanoString: Buffer W	NanoString Technologies	Cat#N/A; RRID: N/A
NanoString: Buffer R	NanoString Technologies	Cat#N/A; RRID: N/A
Tissue Processing: Human Serum	Sigma Aldrich	Cat#H3667; RRID: N/A
Tissue Processing: RPMI 1640 Medium, low HEPES, low bicarbonate, no glutamine	ThermoFisher	Cat#42402016; RRID: N/A
Tissue Processing: PBS, pH 7.4	ThermoFisher	Cat#10010023; RRID: N/A
Tissue Processing: BSA	Cell Signaling Technology	Cat#9998S; RRID: N/A
Tissue Processing: Premium Grade Fetal Bovine Serum (FBS)	VWR	Cat#89510-194; RRID: N/A
Tissue Processing: 2-Mercaptoethanol	ThermoFisher	Cat#21985023; RRID: N/A
Tissue Processing: eBioscience 10X RBC Lysis Buffer (Multi-species)	ThermoFisher	Cat#00-4300-54; RRID: N/A
Cell Stimulation: L-Glutamine	ThermoFisher	Cat#25030149; RRID: N/A
Cell Stimulation: Penicillin: Streptomycin solution	VWR	Cat#45000-652; RRID: N/A
Cell Stimulation: Recombinant Human IL-6	PeproTech	Cat#200-06; RRID: N/A
Cell Stimulation: Recombinant Human TNF-α	PeproTech	Cat#300-01A; RRID: N/A
Cell Stimulation: Recombinant Human IL-1β	PeproTech	Cat#200-01B; RRID: N/A
Cell Stimulation: Recombinant Human IL-1α	PeproTech	Cat#200-01A; RRID: N/A
Cell Stimulation: Buffer TCL	QIAGEN	Cat#1031576; RRID: N/A
Critical commercial assays		
FISH: RNAscope LS Multiplex Fluorescent Reagent Kit	Advanced Cell Diagnostics	Cat#322800; RRID: N/A
FISH: RNAscope LS 4-Plex Ancillary Kit Multiplex Reagent Kit	Advanced Cell Diagnostics	Cat#322830; RRID: N/A
NanoString: GeoMx Solid Tumor TME Morphology Kit	NanoString Technologies	Cat#N/A; RRID: N/A
NanoString: GeoMx Nuclear Stain Morphology Kit	NanoString Technologies	Cat#N/A; RRID: N/A
CRC Sample Processing: Human Tumor Dissociation Kit	Miltenyi Biotec	Cat#130-095-929; RRID: N/A
FISH/IF: Opal 480 Reagent Pack	Akoya Biosciences	Cat#FP1500001KT; RRID: N/A
FISH/IF: Opal 520 Reagent Pack	Akoya Biosciences	Cat#FP1487001KT; RRID: N/A
FISH/IF: Opal 570 Reagent Pack	Akoya Biosciences	Cat#FP1488001KT; RRID: N/A
FISH/IF: Opal 620 Reagent Pack	Akoya Biosciences	Cat#FP1495001KT; RRID: N/A
FISH/IF: Opal 690 Reagent Pack	Akoya Biosciences	Cat#FP1497001KT; RRID: N/A
FISH/IF: Opal 780 Reagent Pack	Akoya Biosciences	Cat#FP1501001KT; RRID: N/A
Tissue Processing: CD45 MicroBeads, human	Miltenyi Biotec	Cat#130-045-801; RRID: N/A
Tissue Processing: LS Columns	Miltenyi Biotec	Cat#130-042-401; RRID: N/A
Sequencing: NextSeq 500/550 High Output Kit v2.5	Illumina	Cat#20024907; RRID: N/A
Sequencing: Chromium Single Cell 3′ Library & Gel Bead Kit v2	10X Genomics	Cat#PN-120237; RRID: N/A
Sequencing: Chromium Single Cell 3′ Library & Gel Bead Kit v3	10X Genomics	Cat#PN-1000075; RRID: N/A
Experimental models: Cell lines		
SNU-407	CCLE	RRID: CVCL_5058
Primary CRC-derived fibroblast cell line	This study	RRID: N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
10x Single cell RNA-seq data	GEO	GEO: GSE178341
Raw RNA-seq sequencing reads	dbGaP	dbGaP: phs002407.v1.p1
Interactive web pages for exploration of data	Broad Institute	https://portals.broadinstitute.org/crc-immune-hubs; https://broad.io/crchubs
Software and algorithms		
R (> v3.6.1)	CRAN	https://www.r-project.org/
xgboost (v0.90.0.2)	Chen and Guestrin 2016	https://xgboost.readthedocs.io/en/latest/
R dropletUtils v1.7.1	Lun et al., 2019 (Bioconductor)	https://bioconductor.org/packages/release/bioc/html/DropletUtils.html
Python (Anaconda)	Ancaonda	https://www.anaconda.com/
Scanpy/Paga v1.2	Wolf et al., 2019	https://github.com/theislab/scanpy
Correlation consensus NMF (ccNMF)	This paper	https://github.com/matanhofree/crc-immune-hubs
umiSaturationQC	Habib et al., 2017	https://github.com/matanhofree/umiSaturation
PartitionStability Graph Community detection (clustering)	(Delvenne et al., 2010)	https://github.com/michaelschaub/PartitionStability
Signed-community-detection	Esmailian and Jalili 2015	https://github.com/pouyaesm/signed-community-detection
Multicore-tSNE	Github	https://github.com/DmitryUlyanov/Multicore-TSNE
NeNMF	Github	https://github.com/hiroyuki-kasai/NMFLibrary
NMF toolbox (v1.4)	Li and Ngom 2013	https://sites.google.com/site/nmftool/
MATLAB (R2017a, R2019a, R2020a)	The Mathworks	https://www.mathworks.com/
Pegasus (v0.17.0)	Li et al., 2020	https://pegasus.readthedocs.io/en/stable/
CellRanger 3.1.0	10X Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest
GraphPad Prism	Graphpad Software	https://www.graphpad.com
HALO	Indica Labs	https://indicalab.com/halo/
Analysis code	This study	https://github.com/matanhofree/crc-immune-hubs
Other		
FISH: RNAscope LS 2.5 Probe- Hs-CXCL13	Advanced Cell Diagnostics	Cat#311328; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-CXCL14	Advanced Cell Diagnostics	Cat#425298; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-IL1B	Advanced Cell Diagnostics	Cat#310368; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-RSPO3-O2	Advanced Cell Diagnostics	Cat#490588; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-MMP3-C2	Advanced Cell Diagnostics	Cat#403428-C2; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-VWF-C2	Advanced Cell Diagnostics	Cat#560468-C2; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-CXCL10-C2	Advanced Cell Diagnostics	Cat#311858-C2; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-CXCL11-C2	Advanced Cell Diagnostics	Cat#312708-C2; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-CXCL1-C2	Advanced Cell Diagnostics	Cat#427158-C2; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-INHBA-C2	Advanced Cell Diagnostics	Cat#415118-C2; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-EPCAM-C3	Advanced Cell Diagnostics	Cat#310288-C3; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-GREM1-C3	Advanced Cell Diagnostics	Cat#312838-C3; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-IFNG-C3	Advanced Cell Diagnostics	Cat#310508-C3; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-MYH11-C3	Advanced Cell Diagnostics	Cat#444158-C3; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-TYROBP-C3	Advanced Cell Diagnostics	Cat#457458-C3; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-CD3E-C4	Advanced Cell Diagnostics	Cat#553978-C4; RRID: N/A
FISH: RNAscope LS 2.5 Probe- Hs-COL1A1-C4	Advanced Cell Diagnostics	Cat#401898-C4; RRID: N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FISH: RNAscope LS 2.5 Probe- Hs-COL1A2-C4	Advanced Cell Diagnostics	Cat#432728-C4; RRID: N/A
FISH/IF: Bond Research Detection System	Leica Biosystems	Cat#DS9455; RRID: N/A
FISH/IF: BOND Open Containers 30 mL	Leica Biosystems	Cat#Op309700; RRID: N/A
FISH/IF: BOND Universal Covertiles 100 pack	Leica Biosystems	Cat#S21.2001; RRID: N/A
FISH/IF: ProLong Diamond Antifade Mountant	Fisher Scientific	Cat#P36961; RRID: N/A
FISH/IF: Fisherbrand Superfrost Plus Microscope Slides	Fisher Scientific	Cat#12-550-15; RRID: N/A
FISH/IF: Microscope Cover Glass 24 × 40 - 1.5	Fisher Scientific	Cat#12-544C; RRID: N/A
FISH/IF: Globe Scientific Non-graduated Plastic Test Tube	Fisher Scientific	Cat#22-010-094; RRID: N/A
NanoString: HybriSlip Hybridization Covers	Grace Bio-Labs	Cat#714022; RRID: N/A
FISH/IF: BOND RX Fully Automated Research Stainer	Leica Biosystems	N/A
FISH/IF: Vectra Polaris featuring MOTiF	Akoya Biosciences	N/A
NanoString: GeoMx Digital Spatial Profiler	NanoString Technologies	N/A
Tissue Processing: Precision Balances ML203T/00	Mettler Toledo	Cat#ML203T/00; RRID: N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Nir Hacohen (nhacohen@broadinstitute.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Sequencing data of de-identified human subject specimens have been deposited at dbGaP: phs002407.v1.p1; expression transcript count matrices at GEO: GSE178341. Additional resources for exploring the data are available at our supplemental web page (https://broad.io/crchubs) and the Broad Institute's Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP1162). Accession numbers and links to web pages are also listed in the Key resources table.

The principal analysis code used to analyze data and generate the results presented here has been deposited at github (https://github.com/matanhofree/crc-immune-hubs). The github link is also listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human tumor specimens

Institutional Review Boards at MGH and BWH approved protocols for tissue collection used for sequencing. Informed consent was obtained from all subjects prior to collection. Age and sex of subjects can be found in Table S1. Only patients with primary treatment-naive colorectal cancer were included in this study. Two patients were excluded after collection when it was discovered that they had concurrent hematologic neoplasms: myelofibrosis/AML (C108) and multiple myeloma (C117). Patient H&E slides were from the pathology department archives.

Human cell lines

A primary fibroblast culture was derived from a human CRC organoid culture established from an MMRd specimen from a 64 yo female patient. Initiation and culture of the MMRd CRC specimen was performed as described previously (Sato et al., 2011). Fibroblasts grew out the matrigel, adhered to the bottom of the plate and were separated from the CRC organoid culture during passaging. Upon separation from CRC organoids, fibroblasts were further expanded in DMEM supplemented with 10% FBS, 2 mM L-glutamine, and PenStrep at 37°C, 5% CO2, and frozen down in 90% FBS + 10% DMSO.

SNU-407 MMRd CRC cell line (Depmap ID: ACH-000955, Cosmic ID: 1660034, Sanger ID: 1907, Cellosaurus RRID: CVCL_5058) was derived from a male patient as part of the Cancer Cell Line Encyclopedia (CCLE) project at the Broad and was fingerprinted at the

Broad Genomic Platform to make sure SNPs match the original line, and tested for mycoplasma. SNU-407 were cultured at 37°C, 5% CO2 in RPMI containing 2 mM L-glutamine, 1 mM HEPES, 1 mM sodium pyruvate, 4500 mg/L glucose, and 1500 mg/L sodium bicarbonate and supplemented with 10% FBS and PenStrep, and frozen down in 90% FBS + 10% DMSO.

METHOD DETAILS

In vitro cytokine stimulation of fibroblasts and CRC cells

Primary fibroblasts derived from CRC specimen and SNU-407 MMRd CRC cell line were seeded in a 96 well plate (20K cells/well fibroblasts, 50K cells/well CRC cells), rested overnight and then stimulated for 14h with 10 ng/ml IL-6, TNF, IL-1α, or IL-1β, or left untreated. Upon stimulation, cells were lysed in TCL with 1% BME (50 ul per well). Smart-seq2 protocol was used as previously described (Picelli et al., 2013) to generate mini-bulk RNA-seq libraries (with 500 cells starting material per condition). Libraries were sequenced on Illumina NextSeq500 Sequencer. Results are representative of two independent experiments. SNU-407 cell line was fingerprinted at the Broad Genomic Platform to make sure SNPs match the original line, and tested for mycoplasma.

Tissue processing, CD45 enrichment, and scRNA sequencing

Samples were cut by pathology assistants at MGH and BWH hospitals. To preserve the invasive border for clinical pathological evaluation, the pathology assistants did not sample tumor down to the invasive border. The tissue was transported in ice cold RPMI with 5% human serum prior to processing. Tissue was transferred into a Petri dish on ice. Fat, necrotic, and fibrous areas were removed. Residual blood and stool were washed off the tissue with cold RPMI. Tissue allocated for dissociation was minced into small pieces (~1 mm3) using a scalpel prior to enzymatic dissociation. Thereafter tissue was transferred into 1.5 mL Eppendorf tubes, each containing 1 mL of enzymatic digestion mix (Miltenyi, Human Tumor Dissociation kit). 1 mL of digestion mix was used per 50 mg of tissue. The Eppendorf tubes were then transferred to a rotation shaker set to 37°C and 550 rpm and shaken for 20 min. The digestion mix was subsequently filtered through a 70 μm cell strainer sitting on a 50 mL falcon tube on ice and mechanically dissociated once more with the plunger of a 1ml syringe against the screen. The filter and enzymatic mixture were washed with RPMI containing 2% human serum as needed until the suspension passed through the filter. The cell suspension was spun at 500 g for 7 min in a 4 C pre-cooled centrifuge to pellet the cells. The pellet was lysed in 4ml ammonium-chloride-potassium (ACK) buffer for 2 minutes and then stopped with RPMI containing 2% human serum. The cell suspension was then centrifuged at 500 g for 7 min at 4 C. The resulting cell pellet was resuspended in loading buffer (PBS containing 0.04% m/v BSA) and filtered through the cell strainer snap cap (Corning 352235) into a 1.5 mL Eppendorf tube. The cell suspension was centrifuged at 500 g for 2 min at 4 C. The pellet was resuspended in cold loading buffer (PBS containing 0.04% m/w BSA) and counted by trypan exclusion. The suspension was then diluted to 1000 cells/μl. 8000 cells were loaded into each channel of the 10x Chromium controller, following the manufacturer-supplied protocol for the 3′ kit. Additionally, a CD45-enriched sample was run for each specimen. To this end, dissociated and ACK-lysed cells were resuspended in cold PBS with 2 mM EDTA and 0.5% FCS and CD45+ cells were enriched using CD45 MicroBeads, human (Miltenyi) following the manufacturer's instructions. Cells were resuspended in loading buffer and loaded with 8000 cells per channel as described above. 10x libraries were constructed using the 10x supplied protocol and sequenced at the Broad Institute Genomics Platform. We note, that our tissue dissociation protocol was optimized to recover both malignant epithelial and immune cells in high quality, which required a mild dissociation procedure that is not ideal to extract stromal cells.

RNAscope *in situ* hybridization with co-immunostaining

Patient cohort for RNAscope analysis was: C103 (MMRp), C107 (MMRp), C110 (MMRd), C112 (MMRp), C123 (MMRd), C126 (MMRp), C132 (MMRd), C139 (MMRd), C144 (MMRd). 5 μm sections were cut from formalin-fixed paraffin-embedded blocks onto SuperFrost plus slides and baked at 65°C for 2 hours before use. Mixed RNAscope (Advanced Cell Diagnostics)/antibody antigen retrieval and staining with Opal (Akoya Biosciences) fluorophores was performed on a Leica Bond Rx instrument following the RNAscope LS Multiplex Fluorescent v2 Assay combined with Immunofluorescence protocol (322818-TN). The only two variations from the written protocol were (1) an open wash dispense after the peroxide step and (2) DAPI (Sigma D9542) was dispensed twice at the end of the protocol at a concentration of 1 μg/mL. Slides were rinsed in water (Fisher 23-751628) prior to coverslipping (Fisher 12-544C) with mountant (Life Technologies P36961). Stained slides were imaged using a Vectra Polaris microscope.

Nanostring GeoMx Digital Spatial Profiling method to measure the expression of ~1500 genes in paired epithelial and non-epithelial regions

5 μm formalin-fixed paraffin-embedded tissue sections were baked at 65°C for 1 hour and manually prepared using the manufacturer supplied V1.4 protocol (MAN-10087-03). Per protocol, the slides were washed thrice for 5 minutes in CitriSolv, and then twice for 5 minutes in each of 100% ethanol, 95% ethanol, and then water. Antigen was retrieved by placing slides in a staining jar containing 1x Tris EDTA (pH 9) and incubated at low pressure at 100°C for 20 minutes. This was followed by a 5 minute wash in PBS. Thereafter, slides were placed in a staining jar with 1μg/mL proteinase K and incubated at 37°C for 15 minutes. After proteinase digestion, slides were washed in 10% neutral buffered formalin for 10 minutes. This step was followed with two washes in a stop buffer containing tris and glycine and one wash in 1x PBS. The RNA probe mix (pre-commercial version of Cancer Transcriptome Atlas probeset) was

diluted in buffer R and this hybridization solution was pipetted over the tissue, covered with a hybrislip coverslip, and incubated overnight at 37°C. The following morning, the coverslips were removed and slides washed twice with a stringent wash containing SSC and formamide at 37°C and then twice with SSC. The slides were then stained with fluorescently labeled morphology markers (CD45, Pan-cytokeratin, CD8, and Syto13) for 1 hour and then washed twice in SSC.

Slides were loaded on the GeoMx® microscope for imaging and barcode acquisition, following the manufacturer supplied protocol (MAN-10102-01). An overview scan at 20x was acquired. 45 circular regions of interest measuring 500 μm in diameter were placed on slides. ROIs were segmented into PanCK-positive and -negative areas of interest. The digital mirrored display was then employed to direct the UV laser to collect barcodes according to the specified collection masks.

Library preparation was performed according to manufacturer instructions (Nanostring DSP-Genomics Library Preparation Protocol 01/2019). Per protocol, a PCR mastermix and well-specific indices were employed to index and amplify the collected wells in a thermocycler. Thereafter, amplified barcodes were pooled and purified using AMPure XP beads and ethanol washes. A Bioanalyzer DNA high sensitivity trace was used to assess library quality. Samples were sequenced on the NextSeq2000 platform.

QUANTIFICATION AND STATISTICAL ANALYSIS

scRNA-seq pre-processing and quality control filtering

For droplet-based scRNA-seq, CellRanger v3.1 was used to align reads to the GRCh38 liftover (37 liftover, v28; https://www.gencodegenes.org/human/release_28lift37.html) human genome reference. The output was processed using the dropletUtils R package (version 1.7.1), to exclude any chimeric reads that had less than 80% assignment to a cell barcode, and identify and exclude empty cell droplets (Griffiths et al., 2018; Lun et al., 2019), by testing against a background generated from barcodes with 1,000 to 10 UMIs, with cutoffs determined dynamically based on channel-specific characteristics. UMI and gene saturation was estimated in individual cells by sub-sampling reads without replacement in each cell barcode, in incremental fractions of 2%, with 20 repeats. A saturation function of the form $y = \frac{ax}{x+b} + c$ was fit based on the number of UMIs observed while sampling reads at different depths. Cell barcodes were excluded if they satisfied any one of the following criteria: (**1**) Fewer than 200 genes; (**2**) Fewer than 1,000 reads; (**3**) Fewer than 500 UMIs; (**4**) More than 50% of UMIs mapping to the mitochondrial genome; (**5**) Non-empty droplet with false discovery rate (FDR) less than 0.1; or (**6**) Over 5% of reads estimated as coming from swapped barcodes/chimeric reads (available at the supplemental website, see Data and code availability). The filtered data was clustered and cells were manually assigned to immune/stromal/epithelial groups based on expressed markers. Using outlier exclusion separately for each channel and each channel cell type combination, cells that deviated by > 2 interquartile ranges (IQR) from the median were then flagged based on the following criteria: (**1**) $\log_{10}$(total transcript UMI), (**2**) Fraction of barcode swaps, (**3**) Gene saturation estimate, (**4**) UMI saturation estimate, (**5**) Fraction of UMI supported by > 1 read (Habib et al., 2017). Cells were further flagged if they substantially deviated from the fit based on the following relationships: (**1**) Total reads versus total UMI, (**2**) Total UMI versus log likelihood of being empty (Lun et al., 2019); (**3**) Total UMI versus total number of genes. A cell was excluded if it was flagged by at least two of these criteria for epithelial and immune cells, or at least three criteria for stromal cells.

Selection of variable genes, dimensionality reduction and clustering

After filtering and exclusion, scRNA-seq profiles were clustered across all patients using a non-negative matrix factorization (NMF) (Li and Ngom, 2013) and a graph clustering-based approach. Transcriptionally over-dispersed genes were identified within each experimental batch (i.e., 10x channel) by the difference of the coefficient of variation (CV) from the median CV for other genes with a similar mean expression (Satija et al., 2015). A robust set of 1,000 to 8,000 genes was retained based on an elbow-based criterion, applied to the median of over-dispersed difference statistics based on 200 samples of 75% of cells. In all subsequent analysis of single cell data we used $\log_2$(TP10K+1) values, calculated for the i^{th} gene in the j^{th} cell as $g_{ij} = \log_2\left(\frac{c_{ij}}{\sum_j c_{ij}} 10^4 + 1\right)$, unless indicated otherwise. Next, 80% of genes and samples were sub-sampled between 50 to 200 times, and NMF was used to reduce the dimensionality of the full dataset to between 15 and 40 dimensions as the product of two non-negative matrices (Lee and Seung, 1999). The loading matrices (i.e., activations) of these NMF components were used to calculate the *k*-nearest neighbors (*k*-NN) graph ($k = 21$) based on a cosine similarity distance. This graph was clustered using stability optimizing graph clustering (http://michaelschaub.github.io/PartitionStability/; Delvenne et al., 2010; Shekhar et al., 2016), to identify 7 top level cell type clusters (epithelial, stromal, mast, B, plasma, myeloid, and T cells). To minimize differences across samples due to technical reasons (e.g., 10x v2 versus 10x v3), gene expression measurements of individual genes were quantile normalized, separately among cells of each top-level cellular compartment, such that the expression CDFs for each gene matched across all batches. Next, the same dimensionality reduction by NMF and graph-clustering procedure was applied iteratively to the transcriptomes of each top-level cell type separately, resulting in a total of 88 cell clusters spanning distinct types or states (Table S1). Of note, PCA-based louvain clustering leads to qualitatively similar cell subset definition (data not shown). However, since we de-novo discover gene expression programs by NMF, we decided to consistently use NMF instead of PCA also for the cluster definition.

Cluster connectivity

To identify relationships between clusters ('cluster connectivity') we used Partition-based Graph Abstraction (PAGA) with connectivity model v1.2 on the NMF based *k*-NN graph above (Wolf et al., 2019). PAGA edge thresholds were selected by using the minimum edge weight of the corresponding minimum spanning tree for each *k*-NN graph (Figures 3D and 3F).

Cluster assignment by gradient boosting and filtering of potential doublets

In order to exclude potential doublets and low confidence assignments by clustering we used a classifier for final assignment of cells to clusters. Gradient boosting (R 3.6.1, xgboost v0.90.0.2 (Chen and Guestrin, 2016)) was first applied to build a cell to cluster classifier for each of the top-level seven cluster types and subsequently to each of the 88 low-level clusters. During training, we included only high quality cells: (1) we excluded potential doublets, defined as cells appearing by manual examination between major high-level cell type regions with expression features from both cell types; (2) cells with possible quality concerns that were not substantial enough for removal during QC; (3) cells with elevated potential ambient RNA contamination, retaining 314,524 cells (85%) for final classifier training.

For each of the seven top-level cell types, a separate classifier was trained to predict each cell type separately (one-versus-all), in a 5-fold cross-validation scheme. Next, using the probability scores of the held-out test-set we identified an optimal cutoff for each class based on an ROC analysis comparing the true positive rate (TPR = true positives divided by all positive predictions) to the false positive rate (FPR = true negative divided by all negatives) and selecting the point at which the ROC curve intersects with the diagonal. Cells that were ambiguously assigned in this way to more than one cluster were removed as potential doublets.

Next, a similar classification training scheme was applied separately to cells from each top-level cell type (epithelial, stromal, mast, B, plasma, myeloid, and T cells). We used 5-fold cross-validation and ROC analysis to select thresholds. In cases where a cell was assigned to more than one subtype, we used the assignment with the higher predictive score. Cells that could not be assigned confidently by any classifier were excluded from further analysis.

Classifying malignant cells by gradient boosting

Adjacent normal tissue, which was sampled distantly from the tumor (e.g., ~10cm apart), is expected to be tumor-cell free. We used gradient boosting to train a classifier predicting malignant from non-malignant epithelial cells based on the source channel type (tumor versus adjacent normal), in a 5-fold cross validation scheme. We separately trained two classifiers, one predicting isTumor and another predicting isNormal, and used the geometric mean of the resulting probabilities as the final statistic. In subsequent analyses, we considered epithelial cells from tumor channels with a predicted score greater than 0.75 to be malignant, and cells from normal channels with a predicted score < 0.25 to be normal epithelial cells. Overall, by this measure ~95% of tumor channel epithelial cells were predicted to be malignant, and 98% of normal channel epithelial cells were predicted to be non-malignant cells. The classifier predictions were highly concordant with those made by inferred copy number alterations with only ~11% of likely malignant cells showing no substantial copy number differences from normal (8% for MMRp, and 15% for MMRd), and 2% of likely normal cells showing copy number differences (data not shown). Copy number alterations were only determined for epithelial cells.

Identification of gene expression programs by NMF

To identify robust transcriptional programs, we adapted a consensus NMF procedure (Kotliar et al., 2019). We used as input the weight components matrices (W matrices) from an NMF procedure that was run on 50-200 subsampled gene x cell subsets, as described above (see section on Selection of variable genes, dimensionality reduction and clustering). We excluded outlier components by sorting components by their cosine distance to the 20th nearest neighbor and excluding components with unusually high distance by an elbow-based criterion. Next, we constructed a *k*-NN graph (k = 30), and identified clusters of highly similar components in this graph using stability optimizing graph clustering (Delvenne et al., 2010), with an exponentially varied scale parameter (0.1 to 10). The components in each cluster were median-averaged into a single component, resulting in a shortlist of "consensus NMF" components. These were used as the initialization component matrix for a second round of NMF of all cells and highly variable genes (as described in Selection of variable genes, dimensionality reduction and clustering). The above procedure was applied separately to each top-level cell population and to epithelial cells from normal channels. For each cell type, this resulted in eight solutions, of between 8-48 clusters corresponding to different choices of the resolution parameter. For each cell type, a single solution was selected based on examination of the mean cluster silhouette, inflection of residual error graph, and by manual examination of the top genes in the output programs.

To characterize the expression programs identified with this procedure, we used the top 150 genes in each of the components, ranked by the following weighting scheme: For the i^{th} gene and j^{th} component we define the scaled weight as follows: $WS_{ij} = W_{ij} \max_{k \neq j} \log \frac{W_{ij}+1}{Wik+1}$ where W_{ik} is the largest weight for gene i in the rest of the components, i.e., $k \neq j$. This weighting scheme prioritizes for high weight (highly expressed; first term in WS_{ij} formula) and unique genes in each component (second term in WS_{ij} formula). For the visualization of relative gene weights of each gene within a program as circle (as in Figures 2C, 2F, 3D, 3F, and 4D), weights were scaled to [0,1] range.

Identification of shared gene programs in malignant epithelial cells

To identify expression programs shared across malignant epithelial cells, from multiple individual patients, the above consensus NMF procedure was first applied separately to malignant cells from each patient (cells from tumor channels and classified as malignant as described above). For each patient a separate consensus NMF expression program set (*W* matrix) was generated, with the number of programs chosen automatically based on the residual error graph. Next, a similar consensus approach was applied to the combined list of all per-patient consensus NMF program sets (all W matrices, one per patient) as well as a set of 17 normal epithelial programs (identified as described above - Identification of gene expression programs by NMF), in order to capture malignant and normal epithelial programs in a single combined NMF solution. After this consensus clustering procedure had completed, NMF clusters including one or more normal epithelial programs were excluded and the corresponding normal NMF programs were used in their place. This was done for all specimens (resulting in 43 pEpi programs), and separately for MMRd and MMRp tumors (resulting in 29 pEpiTd and 32 pEpiTp programs, respectively; Table S4).

Calculating NMF transcriptional program activity

In order to calculate the NMF program activity matrix (*H*), we used non-negative least-squares (NNLS), solving the following equation for the matrix *H*, $H = argmin_{H>0}|X - WH|_F$, given *X* and *W*, where *H* is the 'program activity' matrix, *k* is the by cell matrix; *X* is the gene by cell expression matrix, and *W* is the gene by *k* NMF expression program matrix. *W* was restricted to at most top 100 weighted genes per NMF component (selected as described above - Identification of gene expression programs by NMF). In this way we can calculate the activity values for any cell including cells not part of the original NMF procedure used to discover the program "dictionary" (e.g., pEpiTd* in MMRp cells or in data from Lee et al. (2020).

Testing for covarying NMF expression programs

We calculated the co-variation of two programs A, B as the correlation (see below) between the vectors of their program activity across the patients, where program activity is calculated by the cell type in which the program was initially defined (e.g., pTNI* programs in T/NK/ILC cells). We restricted this analysis to include only patients where at least 1,000 cells were captured and did not consider stromal cells due to their low number per patient (stromal cells account for < 5% of all profiled cells). In order to capture relationships between expression programs that are active in only a small number of cells, we calculated for each patient, cell type, and expression program, the program activity values in this cell type at five quantiles (0.25 0.5, 0.75, 0.95, 0.99). We then calculated the Pearson correlation coefficient, *R,* for every pair of NMF programs, separately for each quantile across patients. The correlation for each quantile was Fisher transformed (i.e., arctanh(R)) and the mean of the five values was used as a test statistic and compared against a null distribution of mean Fisher transformed R values generated by permuting the patient ID assignment (and keeping cell type, and overall NMF value distribution unchanged). A p value was calculated by counting how often the permuted R is above the true observed R (p = (# R > R')/(# permutations), and separately how often the permuted R is below the observed R. The minimum of these (scaled by two) was taken as the outcome empirical p value statistic and reported at a Benjamini-Hochberg FDR of 10%. We report the raw correlation at the 0.75 quantile and the adjusted R, calculated as the difference of mean true R values, and the mean of permuted R values across 10,000 permutations. We constructed a signed weighted network from the pairwise R values retaining only significant edges (FDR < 0.1).

Next, we discovered modules ('hubs') in the resulting network using a module detection algorithm for signed graphs (i.e., having both negative and positive edges; Esmailian and Jalili, 2015). This method explores a space of solutions set by a resolution parameter in the range 0.001 to 0.2, and a random-walk parameter (tau = 0.2), and outputs the optimal solution based on the Constant Potts Model of graph modularity. We applied this method iteratively, and split modules if they were larger than 3 nodes and improved the signed weighted modularity of the solution.

Constructing a network of expression program similarity

A network of expression program similarities was constructed for pTNI*, pS*, pM*, and pEpi* programs by calculating for every pair of program genes a pairwise Jaccard similarity (i.e., for sets A and B J = |A intersect B|/|A union B|) of the top 50 program genes (selected as described above - Identification of gene expression programs by NMF). The resulting similarity matrix was used to construct a Gaussian kernel matrix (as in constructing a tSNE, with perplexity of 30 and a tolerance of 10^{-5}). The kernel matrix was filtered to retain the top 4% of value pairs to construct the final network, and visualized using a force-directed layout algorithm.

Visualization of single cell profiles

We generated tSNE plots per compartment from NMF loading matrices, with a perplexity value of 30 and the Barnes-Hut approximation method (Van Der Maaten, 2014). A global tSNE of all cells was generated using Pegasus with the default parameters and using SVD for the preliminary embedding (v0.17.0; Li et al., 2020).

Identification of differentially expressed genes

Differentially expressed genes (DEGs) were identified using a two-step procedure applied to the log(TP10K+1) values, first using a Mann-Whitney-Wilcoxon Ranksum test, and then sorting genes by Wilcoxon statistic, and testing each of the top 1,000 genes for

differential expression using a generalized linear mixed model using a normal distribution, with terms for the total UMI and the total number of genes, and a fixed effect intercept term for each patient. We report the likelihood ratio Wald-test p value comparing this model to one also including a categorical class term.

Genes were identified as differentially expressed in a particular set of cells if they met all of the following criteria: (**1**) Ranksum test with a Benjamini-Hochberg FDR < 0.1; (**2**) Minimum expression in at least 5% of cells; (**3**) Area Under a Receiver Operating Curve (AUROC) > 0.55, (**4**) 1.25 log fold change versus all other cells; and **(5)** Wald-test with a Benjamini-Hochberg FDR < 0.1. We included tables for the top 100 significant genes (sorted by AUCROC), for immune, stromal and epithelial cells (Tables S2, S3, and S4).

Pearson residual calculation in contingency tables

Enrichment/depletion of particular cell clusters compared to adjacent normal colon tissue (as shown in Figures 2A, 3B, and S2B) were determined using the Pearson residual. The Pearson residual is a measure of relative enrichment for cells in a contingency table. It is calculated here as: $R = \frac{obs-exp}{\sqrt{exp}}$, where the expected value is calculated as the product of row and column marginal probabilities by the total count.

Transcription factor target enrichment in gene expression programs

Transcription factor target gene predictions are aggregated from the following database: (1) Trrust (v2; https://www.grnpedia.org/trrust/, retrieved April 2018; Han et al., 2018), (2) MsigDB (v.7.1; http://www.gsea-msigdb.org/gsea/msigdb/, retrieved March 2020; Liberzon et al., 2015), (3) RegNetwork web (http://regnetworkweb.org/, retrieved Jan 2019; Liu et al., 2015). TF target sets were tested for statistical enrichment within the top genes of each program using the Fisher-exact test. A TF was considered a putative regulator of an NMF program if it showed significant enrichment (FDR < 0.1), had an overlap of at least 3 genes between the top NMF program genes and TF targets, and if the TF gene expression showed a positive correlation with the respective NMF activity.

Preprocessing of bulk RNA-seq data from fibroblast and cancer cell line stimulation experiment

Reads were extracted from image files using bcl2fastq2 (v2.20.00). 2x67nt paired-end reads were mapped to the human genome (GRCh37liftOver) using STAR v2.7.3a and TPM (transcripts per million) was calculated with RSEM v1.3.1. The resulting matrix was $\log_2(x+1)$ transformed for downstream analysis.

Preprocessing of microarray datasets

Microarray datasets were downloaded from GEO: GSE39582; Marisa et al., 2013; GEO: GSE13294; Jorissen et al., 2008) and preprocessed in R to match probe IDs to gene symbols according to the specified microarray chip platform "[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array" with chip definition table GEO: GPL570. For genes represented by multiple probes, the mean value of all probes was taken.

Preprocessing of bulk RNA-seq from TCGA

Standardized RNA-seq expression data for TCGA-COADREAD (CRC) samples were downloaded from GEO along with clinical annotation tables (GSE62944; Rahman et al., 2015). We used log(TPM) values for downstream analysis.

Calculating gene signature scores in bulk expression datasets

We calculated gene signature scores to assess NMF program activities and fibroblast clusters in external bulk RNA-seq cohorts and ISG/MHC-II scores in NanoString GeoMx data (Figure 7). We used the AddModuleScore function of the *Seurat* v3 R package (Butler et al., 2018; Stuart et al., 2019). For each sample, this calculates the average expression of genes in the module subtracted by the average expression of a randomly selected set of control genes with similar expression across the samples. As input to the function, we used normalized expression as described above, and in each case, we used 200 random control genes.

For the NMF program scores, we used the top 150 weighted genes in each program (see Tables S2, S3, and S4). Gene signatures for fibroblast clusters (Figure 3K) were:

- cS27 (*CXCL14+* CAF): *CXCL14*, *AGT*, *NSG1*, *MEST*, *EMID1*, *CST1*, *BMP4*, *WNT4*, *INHBA*
- cS28 (*GREM1+* CAF): *COL10A1*, *GAS1*, *RSPO3*, *COL11A1*, *FAP*, *INHBA*
- cS29 (*MMP3+* CAF): *MMP10*, *CCL20*, *IL1B*, *CSF2*, *STC1*, *INHBA*
- Fibro all: *C1S, LUM*, *DCN*, *RARRES2*, *COL1A2*, *C1R*, *COL6A2*, *COL3A1*, *MMP2*, *FBLN1*, *SERPINF1*, *COL6A1*, *COL6A3*, *COL1A1*, *CTSK*, *TMEM176B*, *MFAP4*, *SPON2*, *PDGFRA*, *TMEM176A*, *PCOLCE*, *CFD*, *VCAN*, *TIMP1*, *AEBP1*, *LGALS3BP*, *EMILIN1*, *LRP1*, *NUPR1*, *OLFML3*, *MEG3*, *FTL*, *CCDC80*, *NBL1*, *FTH1*, *CD63*, *LTBP4*, *IGFBP6*, *TIMP2*, *CLEC11A*, *CST3*, *ECM1*, *IGFBP5*, *MRC2*, *SDC2*, *PLTP*, *CXCL14*, *EFEMP2*, *RHOBTB3*, *RP3-412A9.11.*

Gene signature for MHC-II/ISG was (Figures 7D and 7E):

- ISGscore nanostring: *HLA-DMA*, *HLA-DMB*, *HLA-DPA1*, *HLA-DQB1*, *PSMB10*, *PSMB8*, *PSMB9*, *TAP1*, *TAP2*, *TYMP*, *STAT1*, *CXCL10*, *CXCL11*, *GBP1*, *GBP2*, *GBP4*.

Image analysis with HALO

Raw Vectra Polaris images for each slide were unmixed with inForm software (Akoya Biosciences), using an algorithm built on a library of fluorescence spectra measured using single fluorophore labeled control slides. The unmixed multi-layer image TIFFs from single fields of view were then stitched together fused into a single multi-layer pyramidal TIFF in Halo software (Indica Labs). Tumor regions were manually annotated in Halo. The luminal margin was defined as the region 360 μm radially out into the tumor from the line of outermost growth toward the lumen, and any tissue radially into the lumen was included in the luminal margin. Areas of low tissue quality such as folds, tears, bubbles, edge artifacts, and necrotic tissue were excluded. The FISH-IF v1.2.2 Halo module was used for cell segmentation and phenotyping. The resulting object dataframe was used for calculating phenotypic composition and for further neighborhood and cluster analysis (described in Image analysis, neighborhood definition, and clustering). With the exception of very highly expressed genes, ISH fluorescence was dot-like. The minimum unit dot area and intensity to define a copy were empirically determined by a pathologist (JHC). Copies were recorded as a semiquantitative measure of expression in the output dataframe. Copies were also binned into categories in accordance with recommendations from Advanced Cell Diagnostics: 0, 1+ (1-3 copies/cell), 2+ (4-9 copies/cell), 3+ (10-15 copies/cell), and 4+ (> 15 copies/cell). All ISH probes were called positive if they were category 1+ or above, with the exception of the secreted factors *CXCL1*, *IL1B*, and *RSPO3*, which were categorized as positive if they were 4+.

Image analysis, neighborhood definition, and clustering

For each full slide microscope image, the object data generated with HALO was used to extract a neighborhood for each cell. The neighborhood was defined as all cells within 100 μm and was characterized by: 1) the total number of cells in the neighborhood; 2) the number of cells in the neighborhood from each of the following phenotypes: *PanCK*+, *CXCL10/CXCL11*+, *CXCL13*+, *IFNG*+, *CD3E*+, *CD3E*+*IFNG*+, *CD3E*+*CXCL13*+, *PanCK*+*CXCL10/CXCL11*+, *PanCK*+*CXCL10/CXCL11*-, AllNeg; 3) the mean and median distances to each of the cellular phenotypes, where the distance was set to 150 μm if no cells of a given phenotype were found in the neighborhood; 4) the sum and max of the 'Copies' feature for each ISH stain: *CXCL10/CXCL11*+, *CXCL13*+, *IFNG*+, *CD3E*+.

To identify 'immune-foci' versus 'non-foci' areas we used k-means clustering to cluster cells into two clusters (kmeans() functions from R *stats* package v4.0.1 with parameters: k = 2, nstart = 10, iter.max = 10), where each cell was represented by the sum and max 'Copies' features of its neighborhood. To ensure that clustering results are comparable across all 9 MMRp and MMRd images, the data from all images was concatenated and clustered simultaneously. The cluster with fewer cells was labeled as the foci-cluster, which was validated by manual examination in all 9 images. We also performed *k*-means clustering after shuffling the cell ID-to-neighborhood mapping and ensured that the percent of cells assigned to cluster 2 (i.e., considered foci) for the 9 images was significantly lower (p = 0.003906, Wilcoxon signed rank exact test):

C132	C123	C110	C144	C139	C107	C126	C112	C103
3.25	8.83	3.22	0.47	0.15	2.12	0.15	0.06	0.33
0.06	2.40	0.03	0.00	0.00	0.02	0.00	0.00	0.00

The total number of cells per image and numbers of cells within or outside of foci are recorded in Table S7.

CellPress
OPEN ACCESS

Supplemental figures

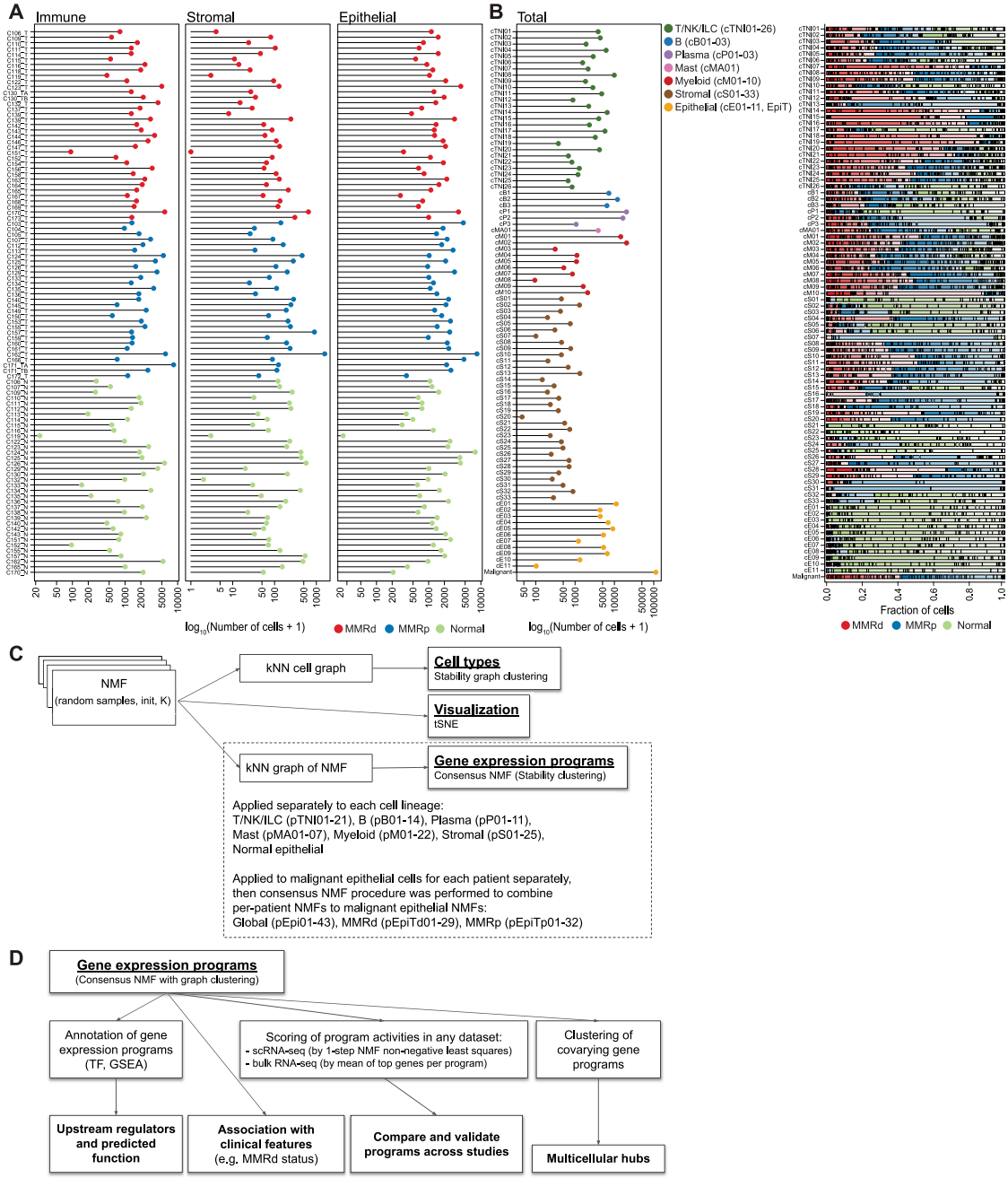


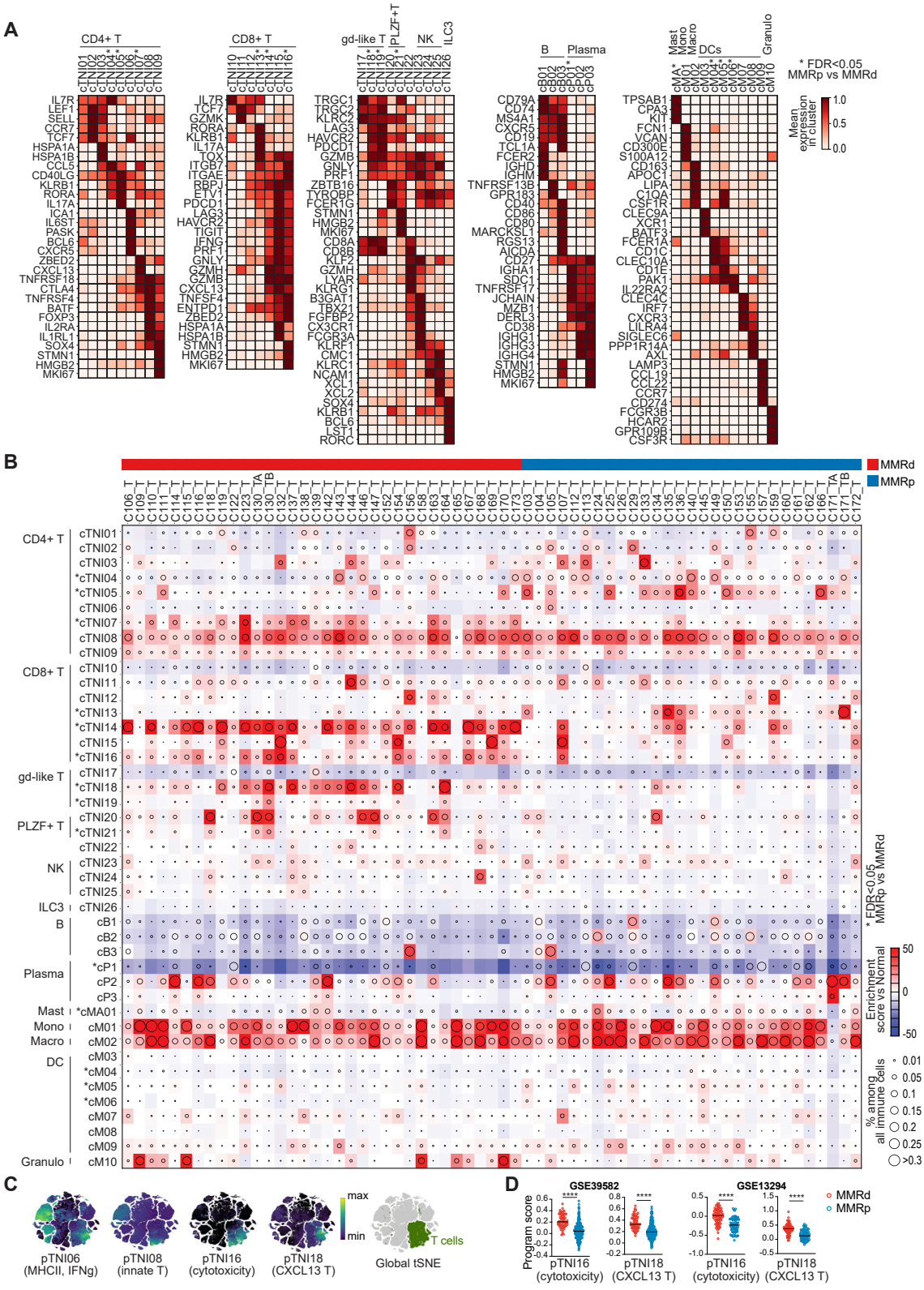
Figure S1. Cellular composition of the scRNA-seq dataset and use of NMF and gene expression programs in our analysis pipeline, related to Figure 1 and Table S1

(A) Number of cells in immune (T/NK/ILC, B, Plasma, Mast, Myeloid), stromal (Endothelial cells, Pericytes, Fibroblasts, Smooth Muscle cells, Schwann cells), and epithelial (malignant in tumor and non-malignant in normal specimens) compartment per specimen.

(B) Number of cells per cluster (left) and fraction of cells from MMRd, MMRp, and normal specimens (right) within each cluster. Each specimen is indicated by a different color shade and separated by a vertical black line.

(C) NMF is underlying cell type clusters, tSNE visualization, and the gene expression programs.

(D) Gene expression programs can be further analyzed in the indicated ways to predict upstream regulators, infer function, or associate the program with clinical features. By calculating gene program activities in other scRNA-seq or bulk RNA-seq datasets, programs can be compared across studies. Clustering of co-varying gene programs enables the prediction of multicellular interaction networks.



(legend on next page)

Figure S2. The immune compartment in MMRd and MMRp CRC, related to Figure 2 and Table S2

(A) Heatmaps showing selected unbiased and well-established marker genes for immune clusters as mean expression in normalized log2(TP10K+1). Clusters with differences in frequency between MMRp and MMRd tumors with Kruskal-Wallis false discovery rate (FDR) < 0.05 are marked with *. A comprehensive list of DEGs for each cluster can be found in Table S2.

(B) Changes in immune cell clusters in MMRp and MMRd tumors relative to adjacent normal tissue, showing frequency of immune cells (dot size) and enrichment/depletion (Pearson residual, colored squares). Clusters with differences in frequency between MMRp and MMRd tumors with Kruskal-Wallis false discovery rate (FDR) < 0.05 are marked with *.

(C) tSNEs showing pTNI06, pTNI08, pTNI16, and pTNI18 program activities on the global tSNE. The location of T cells is indicated in green (right).

(D) Gene signature score for pTNI16 and pTNI18 in MMRd and MMRp CRC in bulk RNA-seq from GSE39582 and GSE13294 patient specimens (Mann–Whitney–Wilcoxon test with **** for $p \leq 0.0001$, *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, ns for >0.05).

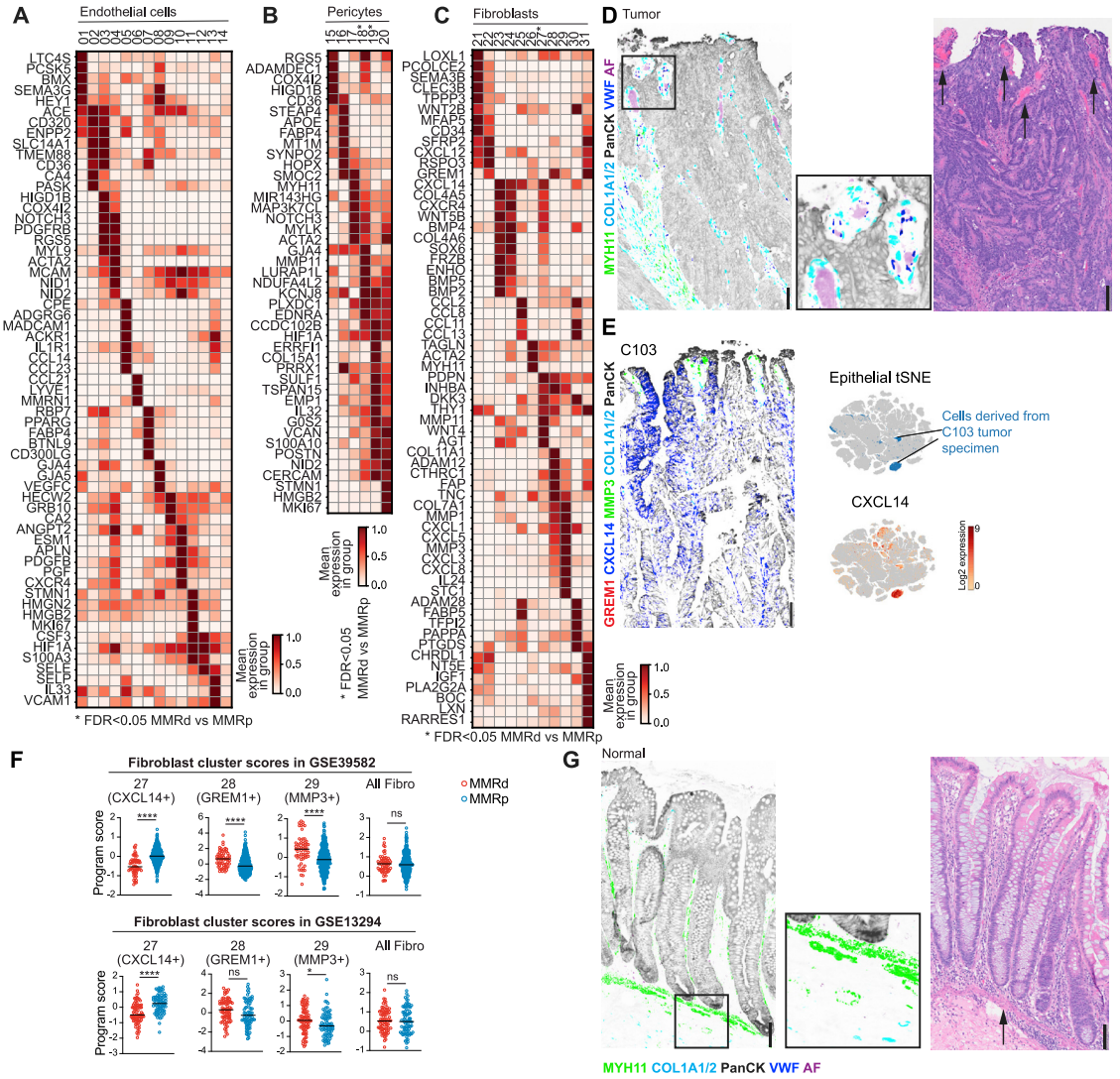


Figure S3. Stromal remodeling in MMRd and MMRp CRC, related to Figure 3 and Table S3

(A) Heatmap showing selected literature based marker genes and differentially expressed genes for endothelial cell clusters as mean expression in normalized log2(TP10K+1). Clusters with differences in frequency between MMRp and MMRd tumors with Kruskal-Wallis false discovery rate (FDR) < 0.05 are marked with *. A comprehensive list of DEGs for each cluster can be found in Table S3.

(B) As (A) for pericyte clusters.

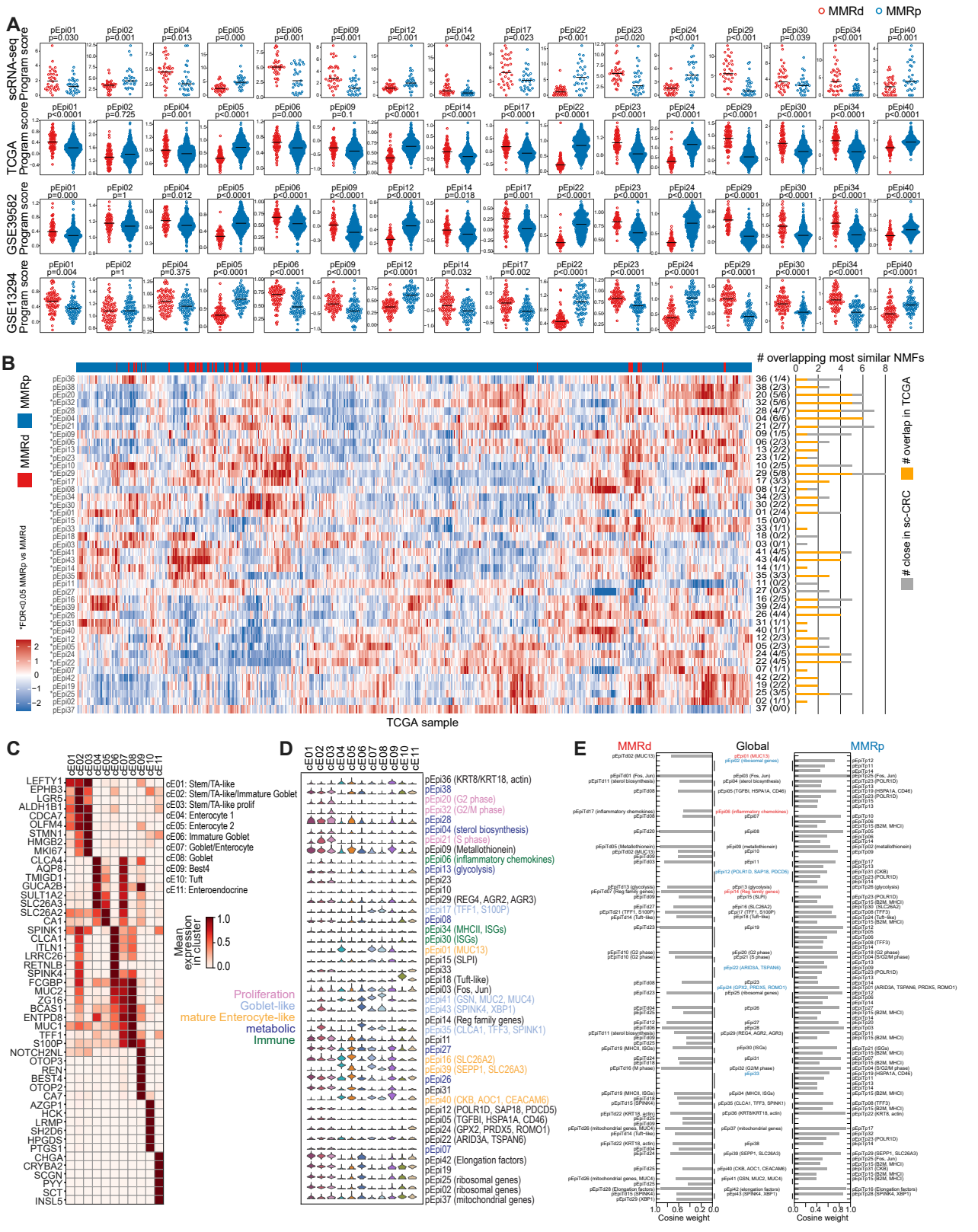
(C) As (A) for fibroblast clusters.

(D) Serial section of the same area as in Figures 3I and 3L stained by multiplex RNA ISH/IF for smooth muscle marker *MYH11*-ISH, fibroblast marker *COL1A1/2*-ISH, epithelial marker PanCK-IF, and endothelial marker *VWF*-ISH (left image). The *MMP3*+ CAFs surround *VWF*+ endothelial cells enclosing autofluorescent (AF) red blood cells. H&E image (right) with dilated blood vessels (bright pink, marked with arrows). Scale bars: 100um.

(E) Representative multiplex RNA ISH/IF image of patient C103 showing *CXCL14*-ISH expression by both epithelial lining fibroblasts and the malignant epithelial cells. Scale bar: 100um. tSNE shows *CXCL14* expression in the malignant cells of patient C103 by scRNA-seq.

(F) Gene signature scores of cell type-specific DEGs for *CXCL14*+ CAFs, *GREM1*+ CAFs, *MMP3*+ CAFs, and all fibroblasts in MMRd and MMRp bulk RNA-seq of patient specimens in GSE39582 and GSE13294. Mann–Whitney–Wilcoxon test **** for $p \leq 0.0001$, *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, ns for > 0.05.

(G) Representative multiplex RNA ISH/IF (as in D) image of *MYH11*+ *COL1A/2*-negative muscularis mucosa below the base of the crypt in non-neoplastic colon (left). H&E image (right) of the same region with arrow pointing to muscularis mucosa. Scale bars: 100μm.



(legend on next page)

Figure S4. The epithelial compartment in adjacent normal colon tissue and MMRd and MMRp CRC, related to Figure 4 and Table S4

(A) Epithelial programs with significantly differential activities between MMRd and MMRp tumors in the scRNA-seq dataset (GLME FDR < 0.05 and > 1.5-fold difference between means) scored in bulk RNA-seq from three external cohorts.

(B) Gene signature for the 43 epithelial programs in bulk RNA-seq from TCGA-CRC (COADREAD) patient specimens. Rows are ordered as in Figure 4B, columns are clustered. Significant MMRd versus MMRp differences are marked with * (Wilcoxon, two-sided with family-wise error rate corrected $p \leq 0.05$). Bar to the right of the heatmap shows the number of most closely correlated programs ($\geq$90th percentile of correlations) based on program activities within scRNA-seq data (yellow+gray) and number of those most closely correlated programs that are preserved in TCGA (yellow).

(C) Heatmap shows selected unbiased and well-established marker genes for normal epithelial cell clusters. A comprehensive list of DEGs for each cluster can be found in Table S4.

(D) Transcriptional activities of epithelial programs within normal epithelial cell clusters.

(E) Similarity between epithelial gene programs and MMRd- and MMRp-derived gene programs based on cosine weight. Programs that only had close matches in MMRd are marked in red, programs that only had close matches in MMRp are marked in blue. See also Table S4.

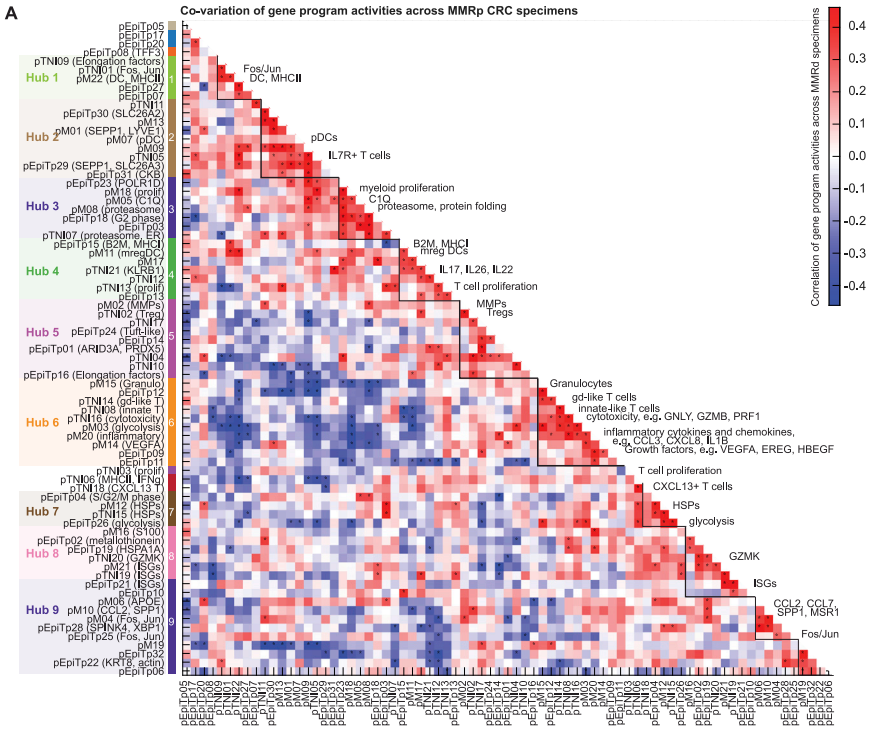


Figure S5. Discovery of multicellular interaction networks in MMRp CRC, related to Figure 5 and Table S5

(A) Heatmap showing pairwise adjusted correlation of gene program activities ('co-variation score') across MMRp specimens (STAR Methods) using specimen-level activities in T/NK/ILC, myeloid, and malignant compartments. Significance is determined using permutation of patient IDs and is indicated with * (FDR < 0.1). Densely connected modules ('hubs') are identified based on graph clustering of the significantly correlated edges.

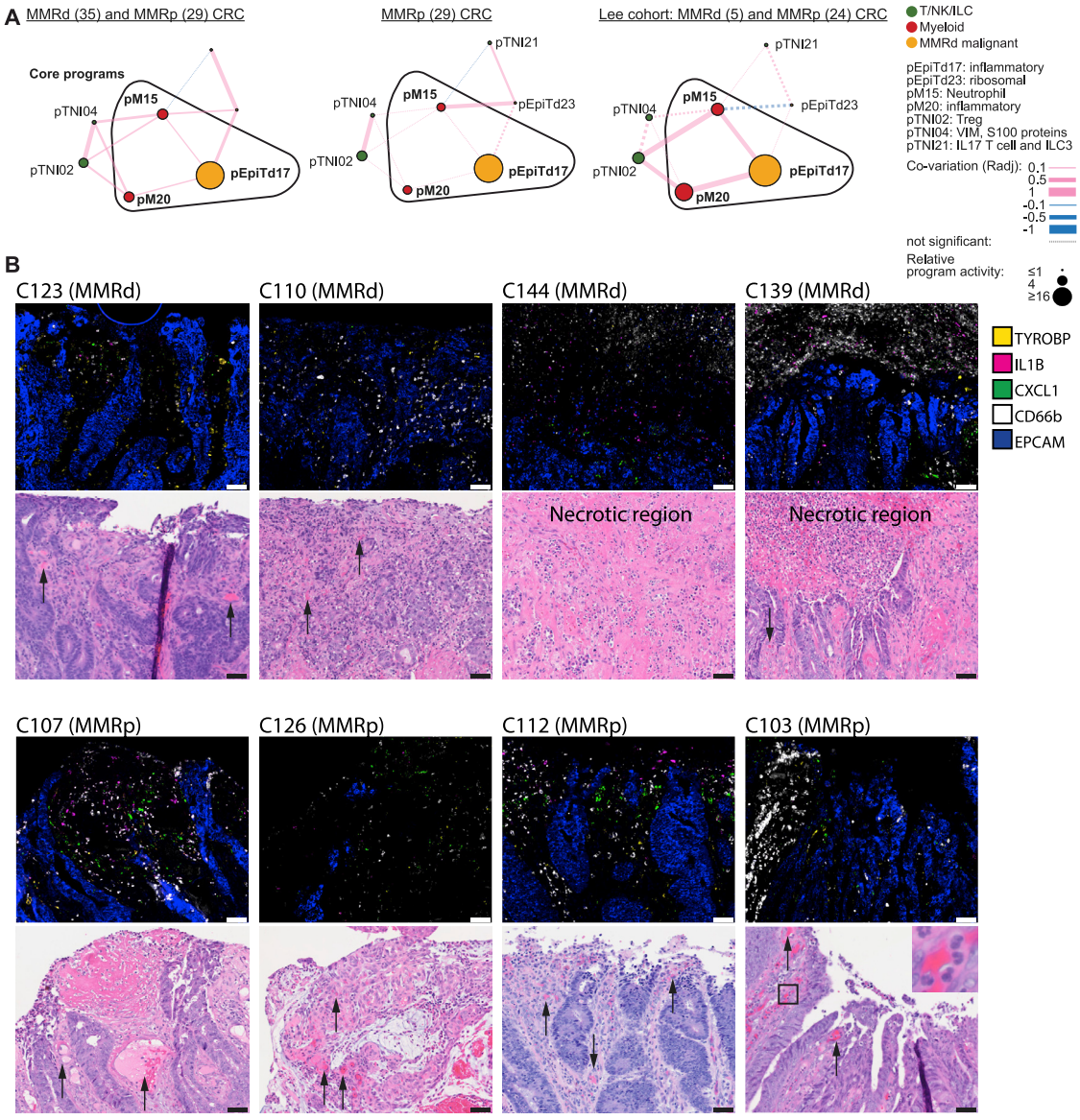


Figure S6. An inflammatory hub at the luminal surface of primary MMRd and MMRp tumors, related to Figure 6 and Tables S5 and S6

(A) Inflammatory hub 3, as discovered in MMRd, projected onto all MMRd and MMRp specimens from our scRNA-seq cohort (left; n = 35 MMRd, n = 29 MMRp), MMRp specimens (middle) or (Lee et al., 2020) (right; n = 5 MMRd, n = 24 MMRp). Nodes respresent gene programs and the size of each node is proportional to the log ratio of the respective mean program activity in MMRd or MMRp versus normal. Edge thickness is proportional to co-variation scores. Pink lines depict positive, blue lines negative correlations. Non-significant edges are depicted as dotted lines.

(B) Multiplex RNA ISH/IF staining for neutrophil marker CD66b-IF, epithelial marker *EPCAM*-ISH, myeloid *TYROBP*-ISH, *IL1B*-ISH, and *CXCL1*-ISH and corresponding H&E images. Representative images of indicated CRC specimens (n = 4 MMRd, n = 4 MMRp) showing accumulations of neutrophils, *IL1B* and *CXCL1* signals at the malignant interface with the colonic lumen, often nearby dilated vessels (marked with arrows) or in necrotic regions (as indicated). Note also that neutrophils are sometimes observed directly within vessels (e.g., C103, inset). Scale bars: 50μm.

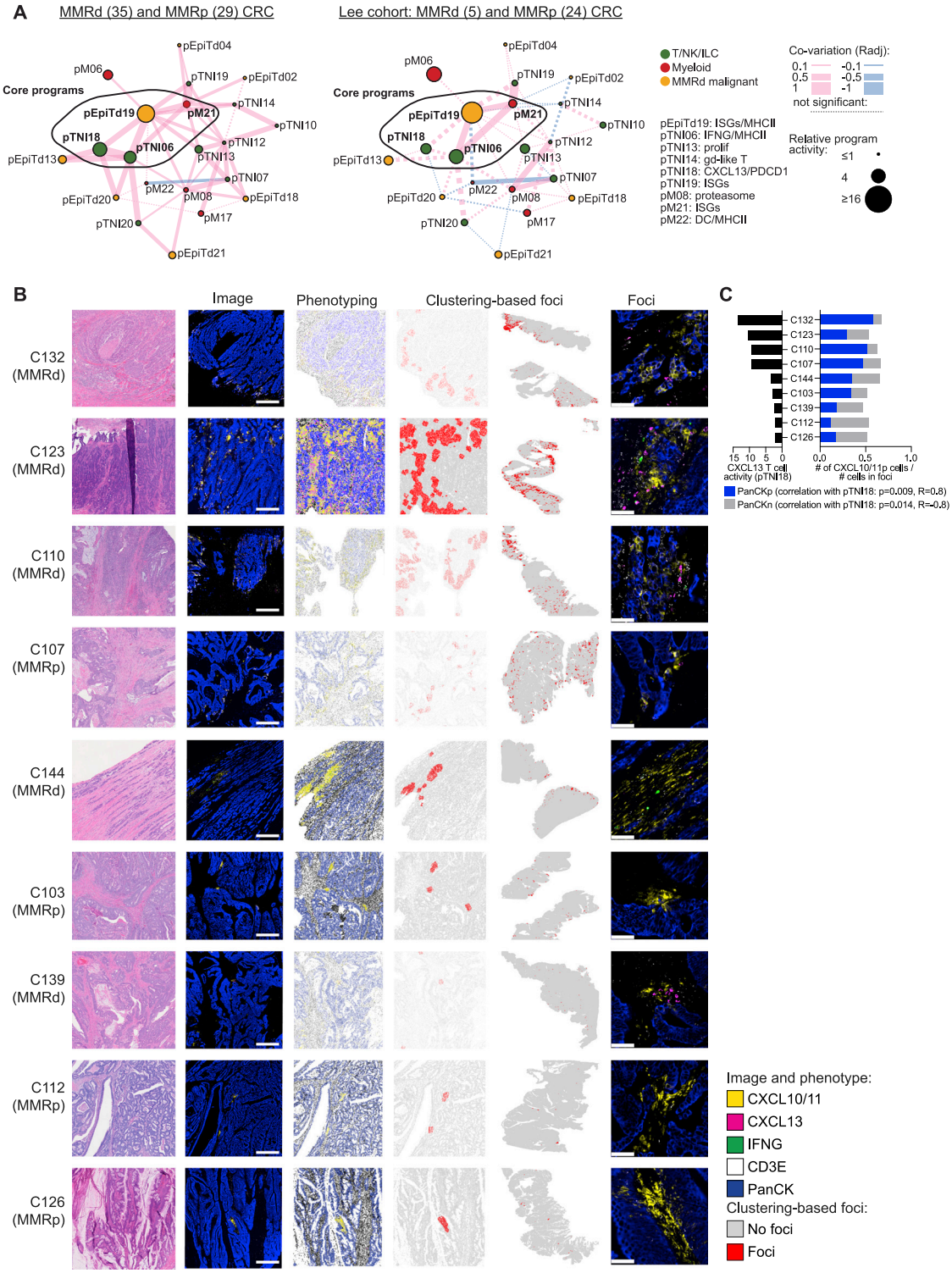


Figure S7. A coordinated network of *CXCL13*+ T, myeloid, and malignant cells expressing ISGs, related to Figure 7 and Tables S5 and S7

(A) ISG/*CXCL13* hub, as discovered in MMRd, projected onto all MMRd and MMRp specimens from our scRNA-seq cohort (left; n = 35 MMRd, n = 29 MMRp) or Lee et al. (2020) (right; n = 5 MMRd, n = 24 MMRp). Nodes respresent gene programs and the size of each node is proportional to the log ratio of the respective mean program activity in MMRd or MMRp versus normal. Edge thickness is proportional to co-variation scores. Pink lines depict positive, blue lines negative correlations. Non-significant edges are depicted as dotted lines.

(legend continued on next page)

(B) Multiplex RNA ISH/IF staining for epithelial marker PanCK-IF, T cell marker *CD3E*-ISH, *CXCL10*/*CXCL11*-ISH, *CXCL13*-ISH, and *IFNG*-ISH on 9 different patient sections (MMRd n = 5: C110, C123, C132, C139, C144; MMRp n = 4: C103, C112, C126, C107). Cells were phenotyped using Halo software and clustered by their neighborhoods (defined as 100 μm) into cells that are part of the foci or not (red and gray, respectively). Shown from left to right for each patient specimen are an H&E section, fluorescent image, a computational rendering of the same section, the assignment to foci in the same section, the assignment of foci in the whole slide scan and magnified fluorescent images of foci. Scale bars: 500 μm for second column, 50 μm for rightmost column.

(C) For each specimen (ordered by their scRNA-seq-based *CXCL13* T cell activity) the fractions of *CXCL10*/*CXCL11*-positive PanCK-positive and *CXCL10*/*CXCL11*-positive PanCK-negative cells within foci are shown. High *CXCL13* T cell activity correlates with higher fractions of *CXCL10*/*CXCL11*-positive PanCK-positive cells (Spearman correlation).