Integrative single cell and spatial transcriptomics of colorectal cancer reveals multicellular functional units that support tumor progression

Inbal Avraham-Davidi[1,21]*, Simon Mages[1,2,21], Johanna Klughammer[1,2,21], Noa Moriel[3], Shinya Imada[4], Matan Hofree[1,] Evan Murray[5], Jonathan Chen[5,6,7], Karin Pelka[5,6,8,9], Arnav Mehta[5,6,10], Genevieve M. Boland[5,11], Toni Delorey[1], Leah Caplan[1], Danielle Dionne[1], Itay Tirosh[12], Nir Hacohen[5,6], Fei Chen[5,13], Omer Yilmaz[4,14], Jatin Roper[4,15,16,*], Orit Rozenblatt-Rosen[1,17,*], Mor Nitzan[3,18,19,*], Aviv Regev[1,17,20,22,*]

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[2]Gene Center and Department of Biochemistry, Ludwig-Maximilians-University Munich, Munich, Germany

[3]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

[4]Department of Biology at MIT, Koch Institute for Integrative Cancer Research at MIT, Cambridge, MA, USA

[5]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA

[6]Massachusetts General Hospital (MGH) Cancer Center, Harvard Medical School (HMS), Boston, MA, USA

[7]Department of Pathology, MGH, Boston, MA, USA

[8]Current address: Gladstone-UCSF Institute of Genomic Immunology, Gladstone Institutes, CA, USA

[9]Current address: Department of Microbiology and Immunology, UCSF, CA, USA

[10]Dana Farber Cancer Institute, Boston, MA, USA

[11]Department of Surgery, MGH, Boston, MA, USA

[12]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel

[13]Harvard Stem Cell and Regenerative Biology, Cambridge MA, USA

[14]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

[15] Current address: Department of Medicine, Division of Gastroenterology, Duke University, Durham, NC, USA

[16] Current address: Department of Pharmacology and Cancer Biology, Duke University, Durham, NC, USA

[17]Current address: Genentech, 1 DNA Way, South San Francisco, CA, USA

[18]Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, Israel

[19]Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

[20]Massachusetts Institute of Technology, Cambridge, MA, USA

[21]These authors contributed equally

[22]Lead contact

*Correspondence: inbalavr@gmail.com (I.A.-D.), jatin.roper@duke.edu (J.R.), orit@broadinstitute.org O.R.-R.), mor.nitzan@mail.huji.ac.il (M.N.), aviv.regev.sc@gmail.com (A.R.)

Abstract

While advances in single cell genomics have helped to chart the cellular components of tumor ecosystems, it has been more challenging to characterize their specific spatial organization and functional interactions. Here, we combine single cell RNA-seq and spatial transcriptomics by Slide-seq, to create a detailed spatial map of healthy and dysplastic colon cellular ecosystems and their association with disease progression. We profiled an inducible genetic CRC mouse model that recapitulates key features of human CRC, assigned cell types and epithelial expression programs to spatial tissue locations in tumors, and computationally used them to identify the regional features spanning different cells in the same spatial niche. We find that tumors were organized in cellular neighborhoods, each with a distinct composition of cell subtypes, expression programs, and local cellular interactions. Three cellular neighborhood archetypes were associated with tumor progression, were active at the same time in different spatial parts of the same tumor, involved dysplasia-specific cellular layouts, and relied on distinct mechanisms: (**1**) inflammatory epithelial regions with endothelial cells and monocytes expressing angiogenesis, inflammation and invasion programs; (**2**) epithelial stem-like regions, associated with plasma and B cell activity; and (**3**) epithelial-to-mesenchymal transition (EMT) regions with dysplastic cells expressing a Wnt signaling program. Comparing to scRNA-seq and Slide-seq data from human CRC, we find that both cell composition and layout features were conserved in both species, with mouse archetypal neighborhoods correlated with malignancy and clinical outcome in human patient tumors, highlighting the relevance of our findings to human disease.

INTRODUCTION

The spatial organization of diverse cells in the tumor ecosystem impacts and drives interactions between malignant cells and neighboring immune and stromal cells, either promoting or suppressing tumor growth (McAllister and Weinberg, 2014). Recent studies have shown that systematic understanding of spatial organization in tumors can shed light on disease progression and response to therapy with specific features correlated with tumor subtypes (Hunter et al., 2021; Pelka et al., 2021; Wagner et al., 2019), cancer prognosis (Jackson et al., 2020; Keren et al., 2018; Schürch et al., 2020), or response to treatment (Grünwald et al., 2021; Jerby-Arnon et al., 2018).

Despite the need to study tumors in their spatial context, genome-scale, high-resolution dissection of the spatial organization of tumors and its functional implications remains challenging, largely due to technical limitations. Methods such as fluorescent *in situ* hybridization (FISH) and immunohistochemistry can only measure a handful of pre-selected transcripts or proteins, whereas single cell RNA-seq (scRNA-Seq) does not directly capture spatial relations. Recent advances in spatial genomics and proteomics allow multiplexed or genome-scale measurements *in situ* (Angelo et al., 2014; Giesen et al., 2014; Goltsev et al., 2018; Marx, 2021; Rodriques et al., 2019; Ståhl et al., 2016; Stickels et al., 2021; Waylen et al., 2020), but with a trade-off between genomic scale and spatial resolution (Palla et al., 2022). This leaves open many fundamental questions about tissue organization and collective function, including whether there are canonical functional units in tumors, what may be their organization in the tumor landscape, and what role does each play in tumor progression.

A case in point is colorectal cancer (CRC), where initial lesions, adenomatous polyps, progress over time to carcinoma and eventually to metastatic disease. While the mutations that drive this process were extensively studied (Cancer Genome Atlas Network, 2012; Fearon, 2011; Fearon and Vogelstein, 1990; Kwong and Dove, 2009), and the cellular ecosystem of CRC has now been deeply charted (Becker et al., 2022; Chen et al., 2021; Pelka et al., 2021), the spatial landscape is less well-characterized. In a recent study of human CRC, we statistically associated cell profiles across tumors and showed that they map to different cellular communities (Pelka et al., 2021), reside in different locations in the tumor and reflect different tumor subtypes. However, absent genome wide *in situ* measurements, these statistical inferences do not yet reflect the full spatial organization of the tumor.

Here, we deciphered the spatial and cellular organization of colorectal cancer (CRC) by combining scRNA-seq and spatial transcriptomics by Slide-seq, using a novel computational framework (Mages et al., 2022). We first profiled an inducible genetic mouse model of colorectal cancer that recapitulates key features of human CRC (Roper et al., 2017, 2018), before and at two time points following tumor initiation, integrating the spatial and cellular profiles to create a cellular map of the tumor landscape, revealing dysplastic-specific cellular layout and potential physical interactions. We found that the tumor landscape is organized in archetypal cellular neighborhoods, with distinct epithelial, immune, and stromal cell compositions, each governed by different gene programs. Three of the cellular neighborhood archetypes are associated with tumor progression, each activating different biological pathways but all active simultaneously albeit in different parts of the tumor. We devised a computational framework to compare single cell and spatial features of tumors between species and applied it to scRNA-seq and Slide-seq data from human CRC.

Multiple features were conserved between tumors in the mouse model and the human patients, and the mouse archetypal neighborhoods correlated with malignancy and clinical outcome (progression-free intervals (PFI) and overall survival (OS)) in human patient tumors. Our findings highlight the multicellular functional tissue modules in the CRC tumor ecosystem and provide a general approach that can be applied to other tissues, tumors and disease conditions.

RESULTS

A cellular atlas of genetic models of colorectal cancer initiation and progression

To chart the cellular ecosystem of CRC and how it changes during tumor initiation and progression, we studied two genetic mouse models of CRC, with inactivation of *Apc* (leading to benign adenoma), followed by an oncogenic $Kras^{G12D/+}$ mutation and then inactivation of *Trp53* (associated with the transition to carcinoma) (Golovko et al., 2015; Roper et al., 2017, 2018) (**Fig. 1a**). In the AV/premalignant model, $Apc^{fl/fl}Villin^{creERT2}$ mice are injected with 4-hydroxytamoxifen to the submucosal layer of the colon, inducing the deletion of *Apc* specifically in epithelial cells within the injection site, and resulting in a local lesion, reproducing the pathology of human adenoma, three weeks later (Roper et al., 2017). In the AKPV/malignant model ($Apc^{fl/fl}$; *LSL-*$Kras^{G12D}$; $Trp53^{fl/fl}$;$Rosa^{26LSL\text{-}tdTomato/+}$; $Villin^{CreERT2}$ mice, Methods), 4-hydroxytamoxifen injection also induces an oncogenic $Kras^{G12D/+}$ mutation and then inactivation of *Trp53*, resulting in invasive carcinomas.

We first generated a single-cell atlas of the models consisting of 48,115 high quality scRNA-seq profiles from normal colon, premalignant (AV, 3 weeks after 4-hydroxytamoxifen induction) and

malignant (AKPV, 3 and 9 weeks after induction) tissues. We captured a diverse cellular census (**Fig. 1b,c**, **Methods**), with 35 clusters annotated *post hoc* by the expression of known marker genes (**Fig. 1b**, **Supp. Fig. 1a,b**, **Methods**) across epithelial, immune and stromal cell compartments (**Supp. Fig. 1c**).

Tumorigenesis causes shifts in cell composition and infiltration of stromal and immune cells

In immune and stromal cells, dysplasia induced shifts in proportions of cells pre-existing in normal tissue, as well as infiltration of new cell subsets in the tissue (**Fig. 1c,d** and **Supp. Fig. 1d,e**). This resulted in both increase in cells of existing populations (*e.g.*, γδT cells (TNK05 (GdT_Il17+)) and emergence of new dysplasia-associated cells (*e.g.*, granulocytes (Gran01, Gran02) and monocytes (Mono02, Mono03)) mirroring observations in human CRC (Pelka et al., 2021), breast cancer (Hagerling et al., 2019), and non-small cell lung cancer (Arenberg et al., 2000) (**Fig. 1c,d** and **Supp. Fig. 1d,e** and **2a,b**). Infiltration is likely to underlie many of these changes as many of the increasing cell subsets (granulocytes, monocytes, mast cells) expressed genes, such as Sell and Ccr2, indicating tissue recruitment, and as the cells dramatically increase in proportion despite negligible signals of proliferation programs.

Two of four monocyte subsets, Mono02 and Mono03, were unique dysplasia-associated cells (**Supp. Fig. 2d-f**), and were respectively enriched for general inflammatory response genes (FDR=5.7 10^{-30}, Fisher's exact test in GO term enrichment) and interferon beta and gamma response genes (FDR=3.5 10^{-11}, 1.0 10^{-13}). T cell subsets showed the expected diversity across nine subsets (**Supp. Fig. 2g-i**) (Smith and Garrett, 2011), with a significant decrease (out of all T cells) in gamma delta (γδ) and Cd8 T cell (TNK01) in the dysplastic microenvironment and an

increase in IL17-producing γδT cells (TNK05) (**Supp. Fig. 2j**). This is consistent with the T cell composition in tumors from mismatch repair proficient (MMRp) CRC patients (**Supp. Fig. 2k**).

Within the five subsets of stromal cells (including vascular endothelial and lymphatic endothelial cells and three fibroblast subsets, **Supp. Fig. 2l-n** and **Methods**), vascular endothelial cells (Endo01) were enriched in dysplastic lesions compared to normal colon (FDR=1.8 10^{-3}, Welch's t test on CLR transformed compositions; **Supp. Fig. 2o**). This is in line with vascular adaptation to the tumor's growing needs for nutrients and oxygen (Ziyad and Iruela-Arispe, 2011) and with the increased expression of the vascular growth factor *Vegf-A* in both monocytes and macrophages (**Supp. Fig. 2p**).

Cell-intrinsic expression shifts in different sub-lineages in the dysplastic epithelium

Epithelial cells showed dramatic cell-intrinsic changes between normal tissues and either premalignant adenomas or malignant carcinomas, such that the cell profiles of dysplastic epithelial cells in both the premalignant and malignant models were highly distinct from normal epithelial cells (and similar to each other) (**Fig. 1c**, **Supp. Fig. 3a** and **Methods**). In a two-dimensional embedding, normal epithelial cell profiles (41% of cells, from normal mice) separated from those from premalignant and malignant models (59% of cells; **Fig. 2a** and **Supp. Fig. 3a**), suggesting a shift in expression profiles from the normal state common to all dysplastic cells. Notably, 11% of the epithelial cells from premalignant/malignant mice were classified as non-dysplastic healthy cells, indicating that normal, non-dysplastic, cells are present in or adjacent to the lesion microenvironment (**Fig. 2b** and **Supp. Fig. 3a)**. We annotated three clusters as dysplastic – Epi01 (dysplastic enterocyte-like), Epi05 (dysplastic secretory-like) and Epi10 (dysplastic enterocyte-like) – because they expressed high levels of *Apc* target genes (*e.g.*, *Axin2, Ascl2, Myc, Ccnd1,*

Lgr5) and were enriched in tdTomato$^+$ cells from AKPVT (malignant) mice (**Fig. 2c** and **Supp. Fig. 3b).**

Interestingly, dysplastic secretory-like epithelial cells (Epi05 (dysplastic secretory-like)) had distinguishing markers (*e.g.*, *Ccl9, Mmp7, Ifitm3*) from their counterparts in normal tissue (Epi04 (secretory)) (**Fig. 2a-c** and **Supp. Fig. 3a,c**). *Mmp7* and *Ifitm3* are known to promote metastasis in human CRC (Li et al., 2011; Zeng et al., 2002), and *Ccl9* expression by epithelial cells promotes tumor invasion through recruitment of Ccr1$^+$ myeloid cells to the tumor's invasive front in a mouse model of CRC (Kitamura et al., 2007). Notably, *Ccr1* is expressed by newly recruited monocytes, macrophages and granulocytes in our model, suggesting a potential mechanism for tumor infiltration and invasion (**Supp. Fig. 3c**). Thus, dysplastic secretory epithelial cells may perform additional functions in support of tumor progression.

Expression programs for stem-like functions, Wnt signaling, angiogenesis and inflammation are activated in dysplastic epithelial cells

In addition to the major distinctions between the absorptive and secretory lineage in the normal and dysplastic epithelial compartment, both normal and dysplastic epithelial cells varied along a continuum, as expected and previously observed in the ongoing differentiation in the colon epithelium (Biton et al., 2018; Haber et al., 2017; Pelka et al., 2021; Smillie et al., 2019). Using non-negative matrix factorization (iNMF from LIGER (Welch et al., 2019), **Methods**), we recovered 20 expression programs spanning the different epithelial functions, and annotated them by Gene Ontology terms enriched in their top 100 weighted genes (**Fig. 2d,e**, **Supp. Fig. 3d-j**, **Methods**).

The programs enriched in different dysplastic cells highlighted key processes that play a role in tumor promotion, including stem cell programs, Wnt signaling, angiogenesis, and inflammation and innate immunity, including interferon alpha, beta and gamma pathways (**Fig. 2d,e**). In particular, the stem cell program (#16) detected in some cells across all conditions, was enriched in dysplastic samples (FDR=5.6 10^{-10}, Welch's t-test on CLR transformed compositions), reminiscent of a recently described population in human (Becker et al., 2022; Chen et al., 2021). Comparing cells from dysplastic and normal samples that express the stem cell program, the dysplastic cells had distinct expression profiles with induction of negative regulators of Wnt signaling (FDR=4.5 10^{-5}, Fisher's exact test in GO term enrichment, *e.g.*, *Notum*, Wnt inhibitory factor 1 (*Wif1*) and *Nkd1*) and genes that are related to cellular response to interferon-gamma (FDR=1.7 10^{-6}, *e.g.*, *Ccl9, Ccl6*) and immune system process (FDR=6.4 10^{-4}, *e.g.*, Ifitm1 and Ifitm3). This is consistent with recent studies showing that *Apc*-mutant stem cells secrete negative regulators of Wnt signaling to induce the differentiation of the WT stem cells in their proximity, thereby outcompeting them and promoting tumor formation (Flanagan et al., 2021; van Neerven et al., 2021). Thus, stem cells from dysplastic lesions may have non-canonical function and regulation (**Fig. 2e,f**). In addition, the programs for Wnt signaling (expressing both positive and negative regulators; #4, FDR=2.8 10^{-6}), angiogenesis (#14, FDR=1.2 10^{-9}), inflammatory response (#6, FDR=1.4 10^{-6}), and innate immune response and interferon response (#7, FDR=1.2 10^{-2}) were all predominantly expressed or enriched in premalignant/malignant epithelium (all with Welch's t-test on CLR transformed compositions, **Fig. 2e** and **Supp. Fig. 3g-j,l**). These results are consistent with the known role of the Wnt signaling pathway in CRC, and of angiogenesis, response to hypoxia and inflammation in tumor progression (Clevers, 2006; Folkman, 2002; Lasry et al., 2016).

Malignant-like tissue programs and composition are conserved between mouse and human tumors

To evaluate the relevance of our findings to human colorectal cancer, we compared them to a scRNA-seq atlas we recently generated from tumor and adjacent normal tissue from 62 patients with either MMRp or MMRd CRC (Pelka et al., 2021). We compared mouse and human tumors in terms of their epithelial expression programs, cellular composition, and cell associations in multicellular hubs (Pelka et al., 2021).

To assess the similarity between mouse and human programs we controlled for overall cross-species and batch differences by normalizing program-specific expression profiles with species-specific background profiles (**Methods**), and then calculating the Pearson correlation coefficients of these normalized scores between the human and mouse programs. Epithelial cells from human and mouse tumors expressed programs highly correlated between the species (**Fig. 3a** and **Supp. Fig. 4a,b**, **Methods**), including for cell cycle, inflammation, epithelial secretory, angiogenesis, Wnt signaling, stem cell like and normal colon functions.

Co-variation in cell proportions across samples (by scRNA-seq) was also conserved between human and mouse tumors, suggesting broad conservation of tumor composition. For example, in both species the proportion of endothelial cells and fibroblasts correlated across samples, as did T, B and epithelial cell proportions in human MMRd tumors and mouse dysplastic lesions (**Supp. Fig. 4c**). Moreover, when we transferred epithelial program annotations from mouse to human scRNA-seq and calculated their co-variation across samples in each species, programs 11

(proliferation), 14 (angiogenesis) and 16 (stem cells), co-varied both across dysplastic mouse tumors and across human MMRp and MMRd tumors (**Supp. Fig. 4d** and **Methods**), suggesting a conserved dysplastic tissue architecture.

"Tissue hubs" originally defined as expression programs from different subsets of cells (epithelial, T/NK and myeloid cells) that co-vary across human CRC tumors (Pelka et al., 2021) were also conserved in mouse tumors. Specifically, we used TACCO to map human expression programs to cell subsets in mouse scRNA-seq, and then assessed their correlation across mouse tumor samples. We found correlated program activation reminiscent of tissue hubs 2, 3 and 5 from MMRp human tumors and with tissue hub 3 from MMRd tumors (**Fig. 3b** and **Supp. Fig. 4e**, **Methods**) (Pelka et al., 2021).

Integrated spatial and single-cell atlas of mouse CRC tumors

To decipher the distribution of cells and programs in the tumor spatial niche, we next used Slide-seqV2 (Stickels et al., 2021) for genome-wide spatial RNA-seq at 10 μm resolution. We sectioned and profiled frozen tissues from four normal colon and six premalignant (AV) lesions using 10 Slide-seqV2 pucks (Methods), recovering 385,428 high quality beads (**Supp. Fig. 5a-c**, **Methods**).

We then integrated the single cell census and spatial profiles using TACCO (Mages et al., 2022) , a novel framework that allowed us to annotate each bead with compositions of discrete cell types (from epithelial, immune and stromal compartments) and to further annotate the epithelial fraction of each bead with a composition of epithelial program activity (**Fig. 1a** "annotation").

We first used TACCO to annotate every bead in the Slide-seq data with a composition of discrete cell subtypes for every puck separately, using its matching single-cell reference (normal or disease; **Fig. 1a**, **Supp. Fig. 5a-c Methods**). To this end, TACCO iteratively solved optimal transport problems to assign cell subtypes to fractions of reads of the beads (Mages, et al., 2022). TACCO relies on unbalanced optimal transport to allow for shifts in the frequency of cell subtypes in the pucks *vs*. the single-cell dataset, while using the reference cellular frequencies as prior knowledge (**Supp. Fig. 5d,f**). TACCO's cell type mapping recapitulated the muscularis layer in its expected tissue location based on the inferred cellular composition pattern (**Supp. Fig. 5c**). This illustrates how TACCO mapped cells correctly by composition.

Next, we used TACCO to map the epithelial gene programs (defined above), focusing on transcript counts that are inferred as derived from epithelial cells. TACCO partitioned the read count matrices for each puck, assigning counts to epithelial cells based on the mapped per-bead cell subtype annotations (from the first step) and the expression profiles associated with each subtype (Mages et al., 2022) (**Methods**). It then summed all epithelial contributions into an epithelial-only spatial count matrix, followed by optimal transport to assign epithelial *program* contributions to individual beads, based on epithelial-only read signals. As for cell type mapping, the proportional contribution of the programs largely recapitulated their contributions in scRNA-seq (**Supp. Fig. 5e,g**).

Altered and less ordered local cellular organization of dysplastic lesions

We assessed the local cellular architecture in term of the preferential proximity of cells of certain type or expressing particular epithelial programs, within a fixed-sized neighborhood, by adapting

an earlier method (Keren et al., 2018). We defined a z-score as significance of the observed neighborship relations compared to the null (Mages et al., 2022) for neighborhoods of 20, 40 or 60 μm diameter (**Fig. 4a,b** and **Supp. Fig. 6a,b**). This z-score is defined with respect to a population of random cell type annotations generated by random permutations of the cell type annotations between the beads, where in our case we permute fractional cell type contributions.

Cell proximity preferences in the normal colon tissue are consistent with the expected morphology, validating our approach (**Fig. 4a,c**). Epithelial cells were organized such that differentiated enterocytes (Epi02 (Enterocytes)) are excluded from the stem cell niche (**Fig. 4a** and **Supp. Fig. 6c**), and endothelial cells and fibroblasts were also spatially co-located in a focused region (**Fig. 4a**), with T cells in their vicinity (**Fig. 4a**).

While some normal tissue features are preserved in dysplastic samples, including co-location of cells of the same lineage (Goltsev et al., 2018; Keren et al., 2018) (**Fig. 4a-c**), there were notable changes, and more disorder. Cell types were more randomly distributed in premalignant *vs*. normal tissue, reflected in lower z-scores ($p=1.6\ 10^{-37}$, Mann-Whitney U test; **Supp. Fig. 6d**). At short distances, all epithelial cells (normal and dysplastic) were preferentially located close to cells from the same subtype (**Fig. 4b**) and even to cells with similar functions: epithelial cells expressing programs associated with malignant-like function (*e.g.*, program 4 (Wnt signaling), 14 (angiogenesis) and 16 (stem cells)) resided close to each other and were spatially distant from cells expressing programs that are related to normal epithelial functions (*e.g.*, program 5 (basolateral plasma membrane), 8 (apical plasma membrane) and 10 (oxidation-reduction process)), supporting a model where tumor progression is structured and compartmentalized (**Supp. Fig. 6e**). Immune

and stromal cells were excluded from epithelial cell neighborhoods with the exception of TNK08 (proliferating T cells) which were highly represented in the proximity of dysplastic epithelial cells (Epi01 (dysplastic enterocyte-like)) (**Fig. 4b, ***), consistent with increased proliferation of T cell subsets that are in contact with tumor cells in CRC (Golby et al., 2002). Granulocytes aggregated together (self-proximal) (**Fig. 4b**) and were relatively close to endothelial cells and dysplasia-associated monocytes (Mono02, Mono03), consistent with their recruitment from the blood through the vessels (**Fig. 4b** and **Supp. Fig. 6f**).

Epithelial regional analysis recovers canonical structures in normal colon

We next defined spatial tissue regions by first identifying "epithelial program regions" as areas of distinct epithelial program activity, and then finding immune or stromal cells associated with each region (**Fig. 1a** "annotation"). Using TACCO, we defined epithelial program regions by Leiden clustering of the weighted sum of neighborship graphs for spatial bead proximity and epithelial expression program similarity, such that transcriptionally similar epithelial beads on different pucks can be connected (despite "infinite" spatial distance, **Methods**). We then used this single framework for region annotation across all pucks (**Fig. 5a**), to determine the distinctive composition of additional cell types in the same set of spatial regions (**Fig.5b-d** and **Supp. Fig. 7a**).

In the normal colon, the regional analysis robustly recovered the expected spatial organization of the healthy colon across five regions and their cellular composition and sublayers (**Fig. 1a**), from luminal/apical to basal. Four regions recovered by TACCO corresponded to different layers of the mucosa (**Fig. 5a** and **5e,f**): a luminal layer with reads found beyond the cellular layer and likely

representing cellular debris trapped in the mucus; three apical layers expressing programs related to normal epithelial function (transmembrane transport, oxidation-reduction process) with gradual transition from apical to basal features; and a basal-most layer, enriched for the deep crypt, proliferation (G1/S,G2/M), MHCII and basolateral plasma membrane programs, all common features of the deep crypt area (Biton et al., 2018; Haber et al., 2017; Sasaki et al., 2016). Finally, region 2, enriched with fibroblasts, myofibroblasts and endothelial cells, and located in the most basal side of the tissue, captured the submucosal and muscularis propria layers, which are predominantly comprised of fibroblasts and muscle, respectively, alongside blood and lymphatic vessels, nerves and immune cells. Overall, TACCO recovered the known organization of the colon, showing the power of our unsupervised mapping approach and shedding light on expression programs that are required for the maintenance of normal colon homeostasis.

Dysplastic lesions maintain some of the programs of the corresponding regions in healthy tissue

Premalignant lesions did not maintain the robust organization of normal tissues, and reflected the expected histopathology of high grade dysplasia, when dysplastic cells are confined to the mucosal layer and do not invade the submucosa (Fleming et al., 2012) (**Fig. 1a** and **5a**) . Specifically, the submucosal and muscularis propria layers from both normal and premalignant tissues were assigned to region 2 (**Fig. 5a**).

Despite the altered morphology, some of the disrupted regions also expressed programs characteristic of their normal healthy function, suggesting that tumor progression is spatially structured and compartmentalized. For example, the region above the submucosa, captured as

region 1 in premalignant lesions and region 5 in normal colon (**Fig. 5a**), had similar features in both premalignant and normal samples. Thus, although the overall spatial organization of the lesion was disrupted, region 1 in premalignant tissue expressed programs that are reminiscent of the normal deep crypt, and was enriched for deep crypt cells and programs that are related to proliferation and MHC II (Biton et al., 2018) (**Fig. 5a,c**).

Other regions in the premalignant tissue also contained some epithelial cells with normal profiles, expressing programs that should allow them to maintain their capacity to perform normal tasks. For example, region 3 expressed apical plasma membrane functions and region 10 was enriched with oxidation-reduction functions (**Fig. 5c**).

To learn about the spatial distribution of the premalignant regions, we measured their distance from region 2 (muscularis) which is a stable landmark in the lesions. Remnants of the layered structure of the healthy tissue were still observed in the premalignant tissue, especially at relatively low distances from the muscularis. For example, healthy region 5 – characteristically located at distances of about 150-200μm from the muscularis – is replaced by dysplastic region 1, peaking at 200μm. All malignant-like regions were spatially associated at ~300-700μm from the muscularis (**Supp. Fig. 7b**), located ~100-400μm apart from each other (**Supp. Fig. 7c**).

Three spatially and functionally distinct tumor region archetypes associated with tumor progression in premalignant lesions

Three regions – 6, 8, and 11 – had epithelial composition and programs that suggested advanced malignant-like characteristics, each highlighting a potentially different mechanism for tumor progression (**Fig. 5g,h**). These three 'malignant-like regions', were enriched (*vs*. all other regions) in stem cell, Wnt signaling and angiogenesis programs (#16, #4 and #14; FDR=6.2 10^{-19}, 5.4 10^{-14}, 1.8 10^{-8}, Welch's t-test on CLR transformed compositions) and depleted of normal epithelial programs (#5, #8, and #10; FDR=7.9 10^{-13}, 2.1 10^{-13}, 4.5 10^{-10}, **Fig. 5i**). Furthermore, the malignant-like regions were enriched in immune cells, including monocytes-macrophages (FDR<=2.4 10^{-5}; excluding Lyve1^{+} macrophages (Mac02)), T cells (FDR<=6.6 10^{-5}; excluding TNK01, γδ/Cd8a T cells; Welch's t-test on CLR transformed compositions), infiltrating granulocytes (FDR<=1.7 10^{-5}), and mast cells (FDR=1.3 10^{-7}), suggesting an ongoing immune response (**Fig. 5j**). However, each one of the three regions had a different epithelial program composition, suggesting that in each type of region there is a different dominant pathway/feature that may drive tumor progression (**Fig. 5c** and **5h**).

Region 6 was characterized by an inflammatory and angiogenic multicellular community, with epithelial and immune cells expressing inflammatory programs, endothelial cells and monocytes connected in a pro-angiogenic circuit, and pro-invasive genes expressed by both endothelial and immune cells (**Fig. 5b,c**). Specifically, region 6 was distinctly enriched for proliferation (programs 3 and 11; FDR=4.1 10^{-32}, 2.0 10^{-25}, Welch's t-test on CLR transformed compositions) and inflammatory epithelial programs (programs 6 and 7; FDR=1.3 10^{-9}, 6.6 10^{-12}), and its non-epithelial compartment was correspondingly enriched for genes from inflammatory pathways, including the response to TNF, IL-1 and IFN γ (FDR=4.4 10^{-3} , 4.3 10^{-3}, 1.1 10^{-5}, Fisher's exact test in GO term enrichment), and chemotaxis of monocytes, neutrophils, and lymphocytes

(FDR=8.6 10^{-5}, 3.4 10^{-11}, 1.3 10^{-2}), suggesting recruitment of inflammatory cells from the circulation or other parts of the tissue. Region 6 was also enriched for collagen binding genes and collagen-containing extracellular matrix (ECM) genes (FDR=1.5 10^{-2}, 2.4 10^{-5}, Fisher's exact test), which are important for migration and invasiveness (Winkler et al., 2020). These include *Sparc*, expressed mainly by endothelial cells and fibroblasts in our data, known to promote colorectal cancer invasion (Drev et al., 2019); and *Ctss*, a peptidase expressed by T cells and monocytes-macrophages that promotes CRC neovascularization and tumor growth (Burden et al., 2009). Finally, gene expression patterns in endothelial cells and monocytes in region 6 suggested active angiogenesis through a multi-cellular feedback loop, with enriched numbers of vascular and lymphatic endothelial cells expressing immune-attracting chemokines (*Cxcl9*) and adhesion molecules (*e.g. Chd5, Mcam*) (**Fig. 5b** and **Supp. Fig. 7d**), monocytes expressing proangiogenic factors that induce proliferation of endothelial cells (*e.g.*, *Mmp12*), and monocytes and macrophages expressing *Ctsd*, which increases tumorigenesis in CRC models (Basu et al., 2019) (**Supp. Fig. 7d**).

Region 8 was enriched for deep crypt cells (program 13; FDR=9.0 10^{-7}, Welch's t-test on CLR transformed compositions), reminiscent of the normal stem cell niche in normal colon, an epithelial innate immune program (program 1; FDR=1.2 10^{-50}) expressed by secretory cells in premalignant and malignant lesions, and plasma and B cell activity. Unlike the canonical (normal) deep crypt region (region 5), which is enriched for MHCII expression (program 18; FDR=5.7 10^{-27}), this region was depleted for the program's expression (FDR=1.2 10^{-33}), suggesting a possible evasion mechanism (**Fig. 5b,c**). The region's non-epithelial compartment was enriched for B cell activation and BCR signaling genes (FDR=4.8 10^{-4}, 2.3 10^{-4}, Fisher's exact test in GO term enrichment).

This may be related to B cell function in protection from lumen antigens (Spencer and Sollid, 2016) or to tertiary lymphoid structures (TLS), which is correlated with clinical benefits in cancer patients (Sautès-Fridman et al., 2019).

Region 11 was heavily populated by cells expressing the Wnt signaling pathway program (#4, FDR=5.5 10^{-23}, Welch's t-test on CLR transformed compositions), with several lines of evidence supporting an active epithelial to mesenchymal transition (EMT) in this region. Epithelial cells in region 11 were enriched for the expression of mesenchymal genes, including Vimentin (Mendez et al., 2010) (*Vim*, FDR=1.2 10^{-242}, Fisher's exact test), *Prox1* (Lu et al., 2012) (FDR=1.4 10^{-155}), and *Sox11* (Oliemuller et al., 2020) (FDR=3.5 10^{-233}) (**Supp. Fig. 7e**), as well as for EMT signatures from a mouse model of lung adenocarcinoma (Marjanovic et al., 2020) (FDR=4.0 10^{-137}, Mann-Whitney U test) and from human head and neck squamous cell carcinoma tumors (Puram et al., 2017) (FDR=1.4 10^{-51}, Mann-Whitney U test). This is consistent with the role of Wnt signaling in promoting EMT and a mesenchymal phenotype in CRC, breast cancer and other epithelial tumors (DiMeo et al., 2009; Schwab et al., 2018). Region 11 non-epithelial cells also expressed genes encoding MHC-I binding proteins (FDR=4.5 10^{-2}, Fisher's exact test in GO term enrichment) and actin cytoskeleton filament and binding proteins (FDR=9.0 10^{-6}, 1.9 10^{-2},1.2 10^{-8}). Organization of the cytoskeleton affects migration, adherence, and interaction of lymphocytes with antigen presenting cells (Penninger and Crabtree, 1999). Notably, region 11 also peaked at a more distal part of the tissue at ~900μm from the muscularis suggesting an outgrowth of the tissue towards the lumen (**Supp. Fig. 7b**).

Non-epithelial cells formed two cellular hubs in the malignant-like regions (6,8,11) (**Supp. Fig. 7f**): An endothelial-fibroblast hub, detected in all three regions, and an immune hub with B cells, TNK cells, monocytes, and macrophages, which was prominent in inflammatory region 6, weaker (less spatially correlated) in region 8, and not correlated in region 11. Thus, activation of an immune response is reflected by close proximity between immune cells.

Overall, our analysis identified three regional archetypes of multicellular communities associated with tumor progression: (**1**) inflammatory epithelial regions with endothelial cells and monocytes expressing angiogenesis, inflammation and invasion programs; (**2**) epithelial stem-like regions, associated with plasma and B cell activity; and (**3**) regions with epithelial to mesenchymal transition (EMT) and Wnt signaling dysplastic cells. Each region archetype highlights different processes that modulate tumorigenesis or invasion, and the three archetypes co-exist in the same tumor at different spatial locations.

Conservation of the spatial organization of human and mouse tumors

The overall spatial distribution of cell types and epithelial profiles was conserved between mouse and human tumors, when comparing scRNA-seq (Pelka et al., 2021) with Slide-Seq data from two MMRd patients and mouse pre-malignant tumors. First, we focused on areas of the pucks that were at least 75μm away from fibroblast-enriched areas and measured the distribution of each cell type relative to endothelial cells, as a proxy to blood vessels. Cell type composition at short distances from endothelial cells was consistent in both species, suggesting a close correspondence of spatial distributions (**Supp. Fig. 8a**). Next, we examined mouse-defined regions in human tumors, using

TACCO to map the expression profiles associated with the epithelial, immune and stroma compartment in each of the TACCO-identified mouse regions to scRNA-seq profiles from human CRC, and probabilistically annotated region-specific expression profiles for each scRNA-seq profile from the human samples. This identified two main "meta compartments", with epithelial, stromal and immune profiles from human MMRp and MMRd tumors associated with archetypal malignant regions 6, 8 and 11 (as well as 0 and 2), while those from normal human tissue were associated with normal regions (*e.g.*, 5, 10, 12) (**Fig. 6a** and **Supp. Fig. 8b**, **Methods**). Finally, transferring human cell type and mouse region annotations to the human pucks showed that one tumor (C110, high grade MMRd with immune cell infiltration), expressed mostly profiles associated with inflammatory region 6 and epithelial stem cell like region 8, whereas the other (G4209T) expressed mostly region 11 profiles, with patches of inflammatory region 6 (**Fig. 6b** and **Supp. Fig. 8c,d**, **Methods**).

Region archetypes are associated with tumor progression in human colorectal tumors

We assessed if the regional epithelial programs that we spatially identified in mouse and are conserved in human are relevant to definition of human disease. To this end, we constructed pseudo-bulk profiles from epithelial cells for our mouse samples and for recently published human samples profiled along different stages of malignant transformation, from normal tissue to polyp to CRC (Becker et al., 2022; Che et al., 2021; Chen et al., 2021; Joanito et al., 2022; Khaliq et al., 2022a; Pelka et al., 2021; Zheng et al., 2022). We scored each epithelial pseudo-bulk profile with the differentially expressed genes between the epithelial parts of the regions and computed the principal components of these scores across all human and mouse samples (**Methods**). The first

principal component captured features that are related to malignancy, with higher values for human tumors *vs*. polyps (**Fig. 6c** and **Supp. Fig. 8e**). In addition, malignant-like region (6/8/11) scores were higher in dysplastic *vs*. normal samples (**Fig. 6d**). Thus, the spatial region profiles defined in mice capture features that correlate with malignant transformation in human.

Moreover, expression of the malignant like regions (6/8/11) in tumors is associated with clinical outcome data. We scored each tumor based on genes that were differentially expressed between the full expression profile of malignant-like regions (6/8/11) and compared the progression-free interval (PFI) and overall survival (OS) for patients in TCGA whose RNA-seq profiles were in the top and bottom quartile of malignant-like region scores (**Methods**). High scores for malignant-like region 11 (EMT) were correlated with shorter PFI and shorter overall survival (OS) (although OS was less robust to changes in the number of differentially expressed genes used to construct the score), while those for malignant-like region 6 (inflammation) correlate with longer PFI and longer OS (**Fig. 6e,f**). This suggests that region 11 is associated with pro-tumorigenic properties in human patients, while region 6 might be associated with tumor controlling properties. This highlights the importance of multicellular functional tissue modules in the CRC tumor ecosystem.

DISCUSSION

Here, we systematically charted the spatial organization of cellular expression in dysplastic tissue of colorectal adenoma, to help identify putative functional units in the tumor. We used TACCO (Mages. et al., 2022) to integrate scRNA-seq and Slide-seq data, not only by mapping cell types to their positions, but also distinguishing different cell programs, the regions that they dominate,

and their characteristic microenvironments. This allowed us to overcome technical limitations, such as lack of spatial context in scRNA-seq and sparse readout in Slide-seq, and to generate a high-resolution spatial map of the dysplastic landscape transcending beyond the mapping of individual cells to spatial positions. We used this map to show correlation with clinical outcome in human patient tumors.

Our scRNA-seq analysis revealed profound enrichment of a stem cell program in dysplastic tissues. The profiles of dysplastic cells expressing this program are distinct from normal stem cells and enriched with expression of negative regulators of the WNT signaling pathway and inflammation, suggesting a non-canonical function. The abundance of these cells with stemness potential across all our malignant-like regions, points to a dynamic population that can affect the cells in its proximity, by secretion of negative regulators of the WNT signaling and inflammatory function but may also adopt various functions depending on the environmental cues and dysplasia-associated cells in its proximity. A similar population, designated “high-plasticity cell state”, was previously described in a mouse model of lung adenocarcinoma and in human patients, where it was correlated with resistance to chemotherapy (Marjanovic et al., 2020). Whether these cells can be manipulated to take on specific phenotypes or even to differentiate into normal-like enterocytes given the appropriate signal from the microenvironment, remain as open questions.

Within the premalignant lesions, alongside malignant-like regions, we found regions with normal features (Regions 3,4,9,10), comparable to regions found in the normal colon, most likely representing compartments driven by clones that were not affected by the genetic perturbation. One of these regions, region 4, contained mainly goblet cells with normal expression profiles.

Whether this neighborhood represents normal cells that reside alongside malignant cells or a cancer transition state, it may modify tumor progression, by recruiting immune cells or by secreting factors that affect epithelial proliferation in adjacent regions. For example, region 4 in premalignant lesions is enriched with chemokine activity genes relative to region 4 in normal colon suggesting a possible role in recruitment of immune cells to the dysplastic landscape. Further work is required to understand the role of these regions (expressing normal features) in tumor progression.

Malignant-like regions activated one of three archetypal regional programs, with coordinated features across epithelial, immune and stromal cells, demonstrating how tumor progression occurs across space. Although the regions are spatially distinct, they reside near each other, and as such may still affect each other by signaling or by using branches of the same main vessels. For example, *Osm* is expressed by cells in region 6, whereas its receptor is expressed on fibroblasts and endothelial cells enriched in region 11. *OSMR* was previously shown to be expressed by inflammatory fibroblasts (Smillie et al., 2019; West et al., 2017) and in CAFs, endothelial cells and pericytes in human CRC (Pelka et al., 2021), and its activation in malignant cells promotes EMT in breast cancer and pancreatic cancer (Smigiel et al., 2017; West et al., 2013) and a mesenchymal state in glioblastoma (Hara et al., 2021). Future studies can help determine if these regions are functionally inter-dependent and if they evolved from the same clones and can inter-convert, or whether they developed independently.

We developed several approaches to allow cross species comparison of tumors at the single cell and spatial level, despite the high level of both intra- and inter-individual variation within each

species. Comparing to human CRC, our analysis suggests that the CRC landscape is organized in similar multicellular functional tissue modules between human and mouse, across disease stages (*e.g.*, mouse adenomas and human carcinomas) and disease subtypes (*e.g.*, MMRp and MMRd). Future studies applying our approaches to patient cohorts could help understand whether the expression of different tissue modules may contribute to the partial response to immunotherapy reported for MMRd patients (André et al., 2020), and to define specific tissue modules predictive of response to therapy. Notably, while our study focused on tumor initiation and progression, its findings may be relevant for tissue response to other challenges (*e.g.*, inflammation, fibrosis, wound healing), which involve activation of similar functional tissue modules, a result of collective function of parenchymal, immune and stromal cells.

Taken together, our integrative approach facilitates spatial analysis with high resolution, constructing regional neighborhoods and their spatial layout at both high cellular resolution and genomic scale. Our work is an important step toward a systematic understanding of the organization of dysplastic tissue with the potential to contribute to improved patient stratification by the multicellular functional units in the tumor landscape.

ACKNOWLEDGEMENTS

We thank N. Friedman, I. Benhar, N. Habib, M. Biton, K. Geiger-Schuller, J.C. Hütter, B. Dumitrascu, E. Baker and A. Greenwald for helpful discussions. We thank C. McCabe, O. Kuksenko and I. Barrera for technical assistance. We thank P. Yadollahpour, E. Dhaval, G. Smith-Rosario, S. Vickovic, D. Schapiro, S. Farhi, D. Abbondanza, A. Segerstolpe and T. Biancalani for their important contribution to this project. We thank L. Gaffney and A. Hupalowska for help with figure preparation. S.M. was supported by a DFG research fellowship (MA 9108/1-1), J.K. was supported by a HFSP long term fellowship (LT000452/2019-L), A.R. was a Howard Hughes Medical Institute (HHMI) Investigator when conducting this work. Work was supported by the Klarman Cell Observatory, a CEGS grant (5RM1HG006193-09) from the NHGRI, the NIH/NIAID (grants 1U24 CA180922, 1U19 MH114821, 1RC2 DK114784), the MIT Ludwig Center, the Manton Family Foundation, and HHMI (A.R.); Azrieli Foundation Early Career Faculty Fellowship, and an ISF Research Grant (1079/21) (M.N.), the Center for Interdisciplinary Data Science Research at the Hebrew University of Jerusalem (N.M. and M.N.), SU2C Peggy Prescott Early Career Scientist Award PA-6146, SU2C Phillip A. Sharp Award SU2C-AACR-PS-32 and NIH/NCI R00CA259511 (K.P.), NIH/NCI R01 CA208756; Arthur, Sandra, and Sarah Irving Fund for Gastrointestinal Immuno-Oncology (N.H.), NIH/NCI R01CA257523, MIT Stem Cell Initiative (Foundation MIT) (O.Y.), and NIH R37CA259363, R21CA256414, R21DK125911, R41EB032693, R01CA254108, R01CA256530, and R01CA244359; DOD W81XWH-20-1-0203; and a Duke-NC State Translational Research Grant (J.R.)

AUTHOR CONTRIBUTIONS

I.A-D. and A.R. conceived the study and designed experiments. I.A-D. J.R., S.Y., E.M., T.D., L.C. and D.D. preformed experiments, with guidance from O.Y. and A.R. S.M., J.K., and N.M. developed computational approaches and analyzed data with I.A-D., M.H., and E.M. and with guidance from M.N. and A.R. M.H., E.M., J.C., K.P., A.M. and G.M.B. generated and analyzed human data, with guidance from N.H.. I.T., N.H., F.C., O.Y, O.R-R, and A.R. M.N and A.R provided supervision and acquired funding. I.A-D., S.M., M.N. and A.R. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

A.M. has served a consultant/advisory role for Third Rock Ventures, Asher Biotherapeutics, Abata Therapeutics, Flare Therapeutics, venBio Partners, BioNTech, Rheos Medicines and Checkmate Pharmaceuticals, is an equity holder in Asher Biotherapeutics and Abata Therapeutics, and has a sponsored research agreement with Bristol-Myers Squibb and Olink Proteomics. G.M.B. has sponsored research agreements with InterVenn Biosciences, Palleon Pharmaceuticals, Olink Proteomics, and Teiko Bio. G.M.B. is a consultant for Ankyra Therapeutics and InterVenn Bio. G.M.B. has been on scientific advisory boards for Merck, Iovance, Nektar Therapeutics, Instil Bio, and Ankyra Therapeutics. G.M.B. holds equity in Ankyra Therapeutics. N.H. holds equity in BioNTech and is a founder of Related Sciences/DangerBio. F.C. is a founder and holds equity in Curio Biosciences. O.Y. holds equity and is a SAB member of AVA Lifesciences. A.R. and O.R.-R. are co-inventors on patent applications filed by the Broad Institute for inventions related to single cell genomics. O.R.-R. has given numerous lectures on the subject of single cell genomics to a wide variety of audiences and in some cases, has received remuneration to cover time and

costs. O.R.-R. is an employee of Genentech since October 19, 2020 and has equity in Roche. A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until July 31, 2020. From August 1, 2020, A.R. is an employee of Genentech and has equity in Roche.

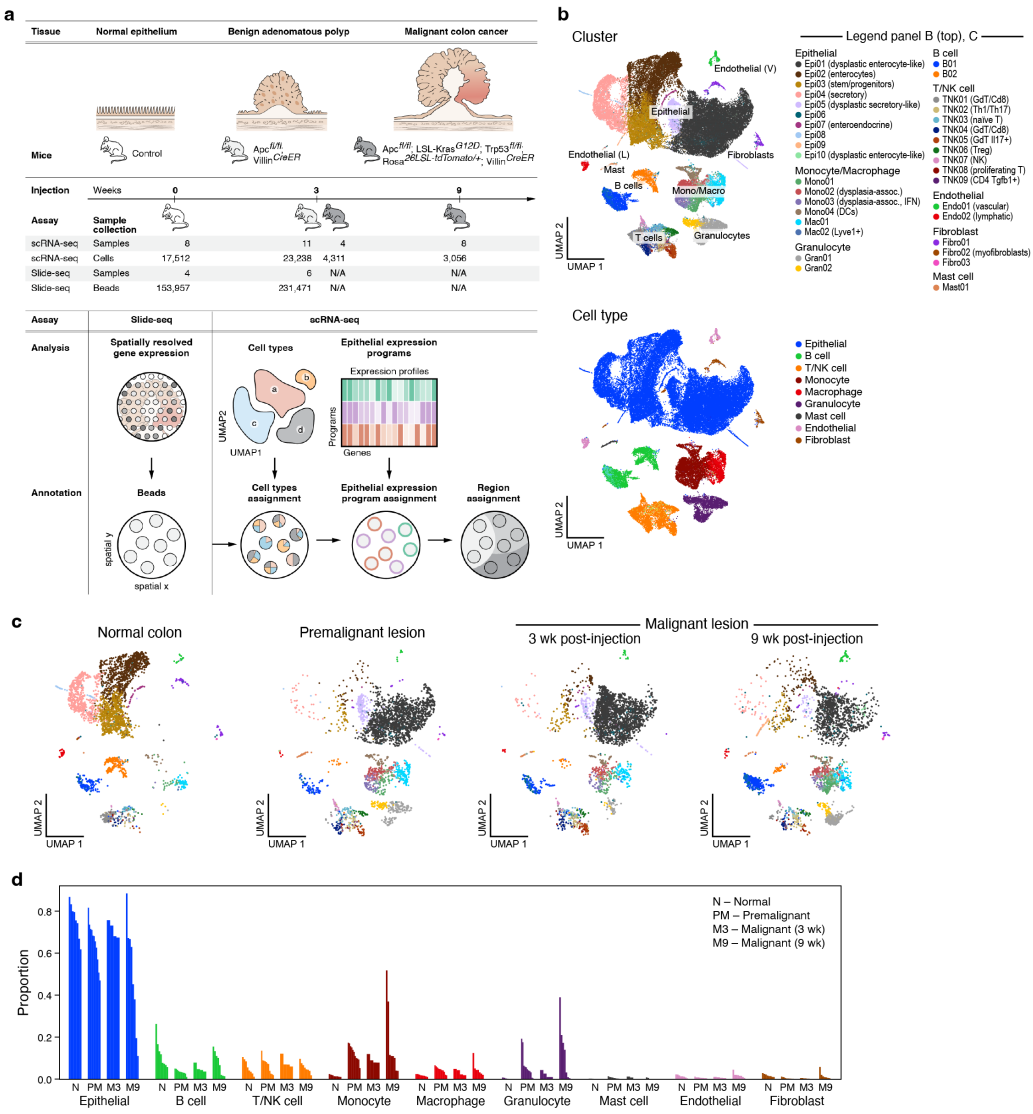


Figure 1. A single cell atlas of healthy colon and dysplastic lesions in mouse.

a. Study overview. **b.** Major cell subsets of healthy colon and dysplastic lesions. 2D embedding of 48,115 single cell profiles colored by cluster (top, legend) or annotated cell type (bottom, legend). **c,d.** Changes in cell composition in dysplastic tissues. c. 2D embedding of single cell profiles, showing only the cells in each condition state, subsampled to equal numbers of cells per condition state, colored by cluster (same legend as in b). d. Proportion of cells (y axis) of each cell type in each sample (x axis).

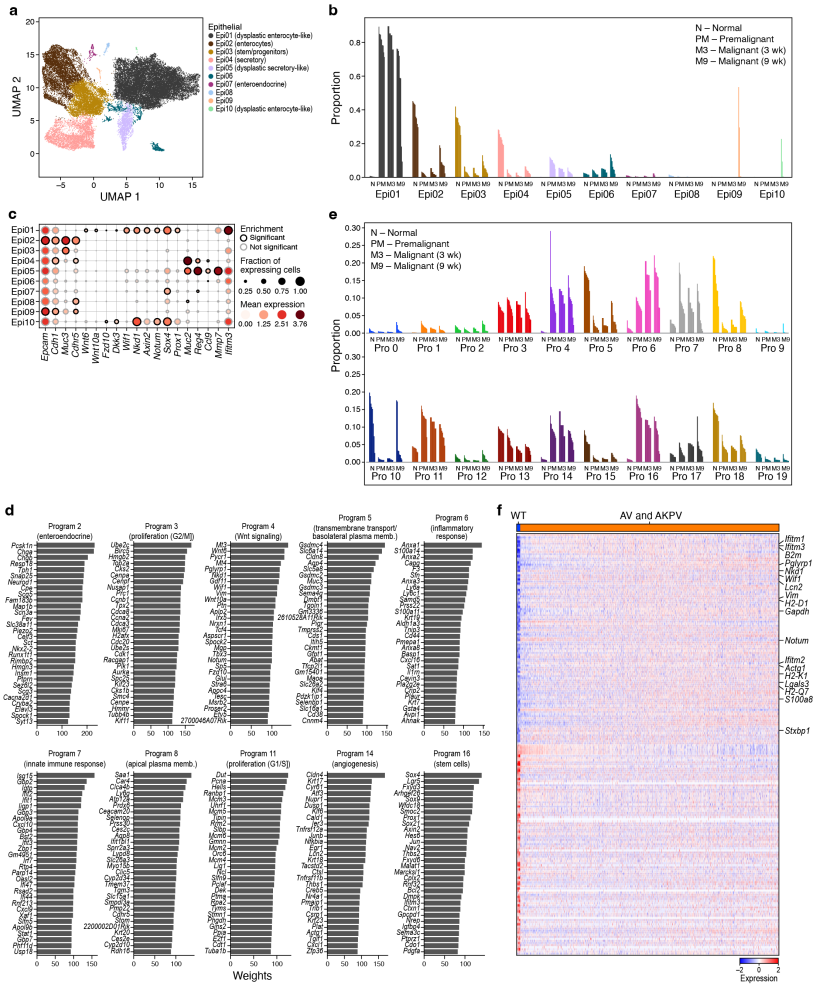


Figure 2. Composition and cell intrinsic expression program changes in dysplastic epithelial cells.

a-c. Compositional changes in epithelial cells in dysplastic tissue. a. 2D embedding of epithelial cell profiles colored by clusters (legend). Cluster Epi06: doublets (not called by Scrublet (Wolock et al., 2019)). b. Proportion of cells out of all epithelial cells (y axis) of each epithelial cell subset in each sample (x axis). c. Fraction of expressing cells (dot size) and mean expression in expressing cells (dot color) of marker genes (columns) for each cluster (rows). **d-e.** d. Use of epithelial cell programs changes in dysplastic tissue. d. Weights (x axis) of each of the 20 top ranked genes (y axis) for each program. e. Proportion of program weights summed over all epithelial cells (y axis)

in each sample (x axis). **f.** Stem cell program 16 is induced in epithelial cells in dysplastic tissue. Scaled log-normalized expression (color bar) of the top 100 genes differentially expressed between cells from normal colon and from dysplastic (premalignant and malignant) across the 10,812 cells that accounted for 90% of program 16's expression across all epithelial cells (columns). Selected program genes are marked.

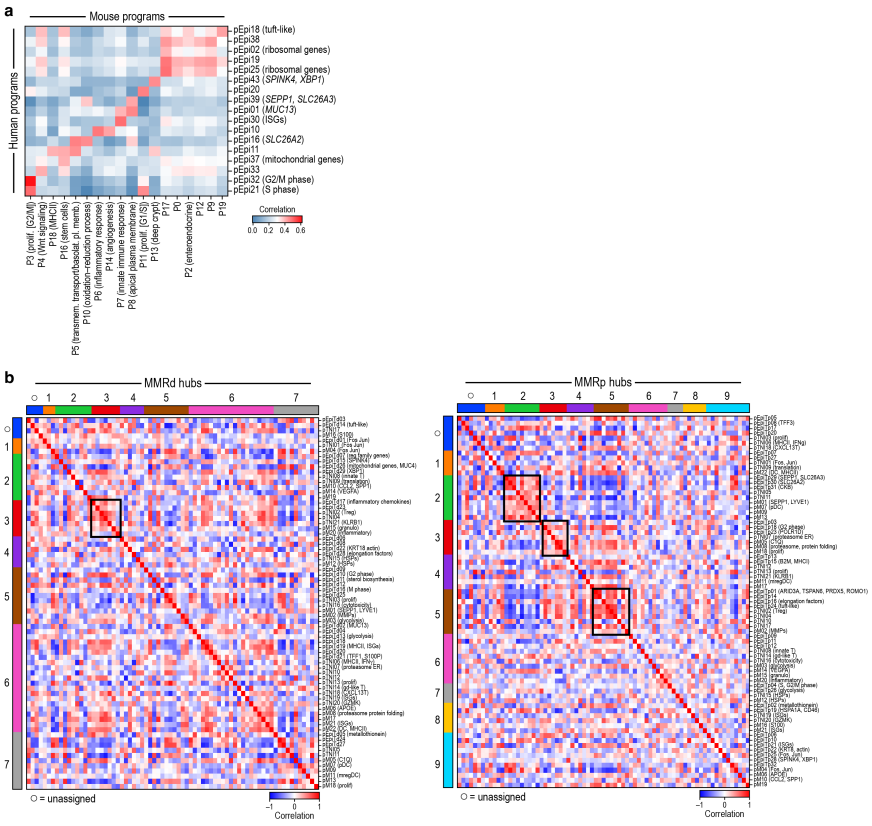


Figure 3. Conservation of malignant-like programs and multicellular hubs between mouse and human CRC.

a. Conservation of epithelial programs in human and mouse. Pearson correlation coefficients (color) between program-specific expression profiles of human (rows) and mouse (columns) programs (Methods). **b.** Key human multicellular hubs are conserved in mouse tumors. Pearson correlation coefficients (color) between the per sample composition profiles of each human expression program (rows, columns) in mouse samples. Boxes: MMRd (left) and MMRp (right) conserved multicellular hubs defined in human tumors (Pelka et al., 2021).

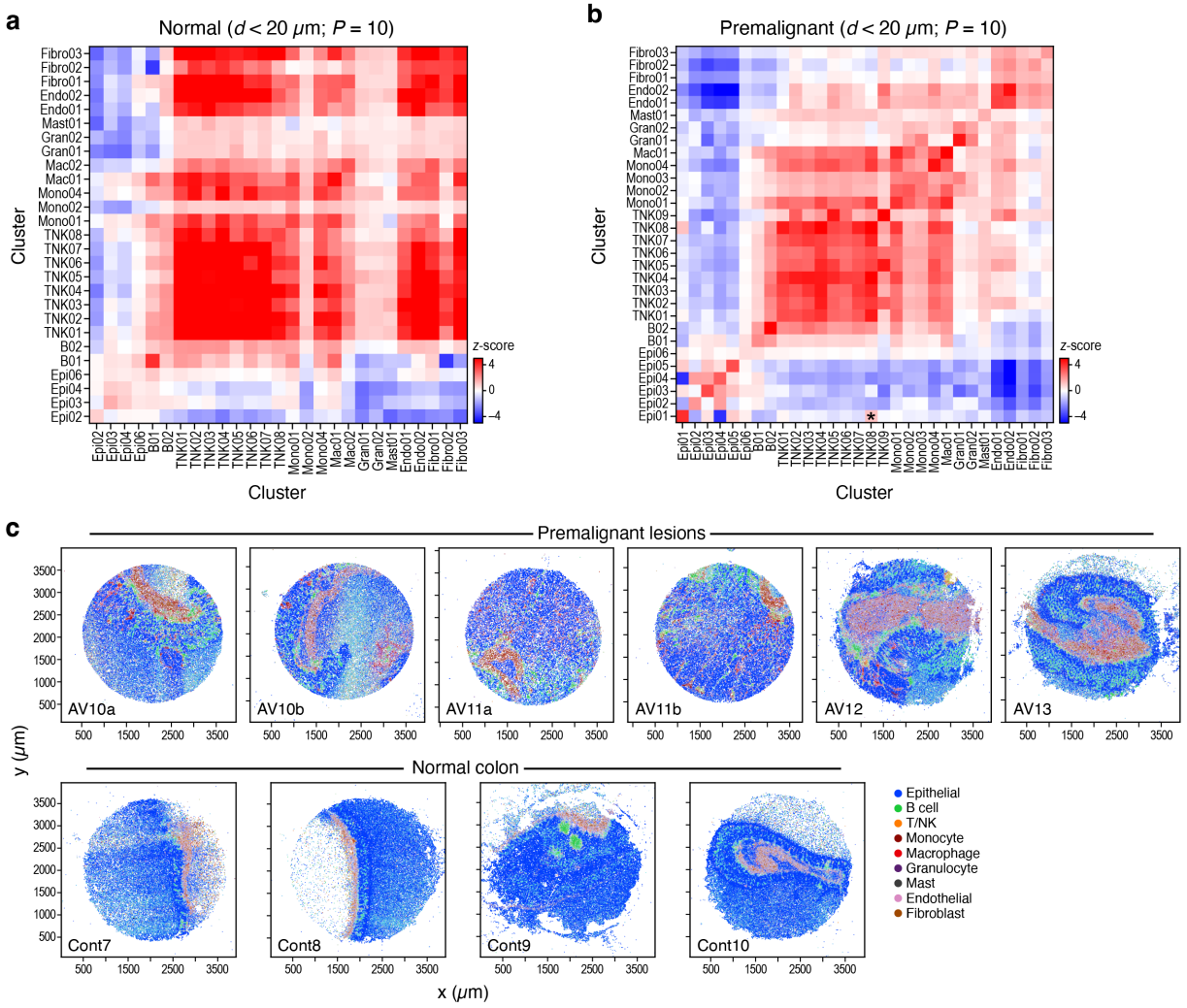


Figure 4. Altered cell type neighborship in CRC.

a,b. Cell type neighborships in normal and dysplastic colon tissue. Short-range (up to 20μm) neighborship enrichment (Z score, color bar) vs. a background of spatially random annotation assignments for each pair of cell annotations (rows, columns) in normal (a) and dysplastic (b) tissue. Asterisks: interactions mentioned in the text. **c.** Cell type distributions *in situ*. Slide-seq pucks of dysplastic (top) and normal (bottom) tissue colored by TACCO assignment of cell labels (legend) (x and y axis: spatial coordinates in μm).

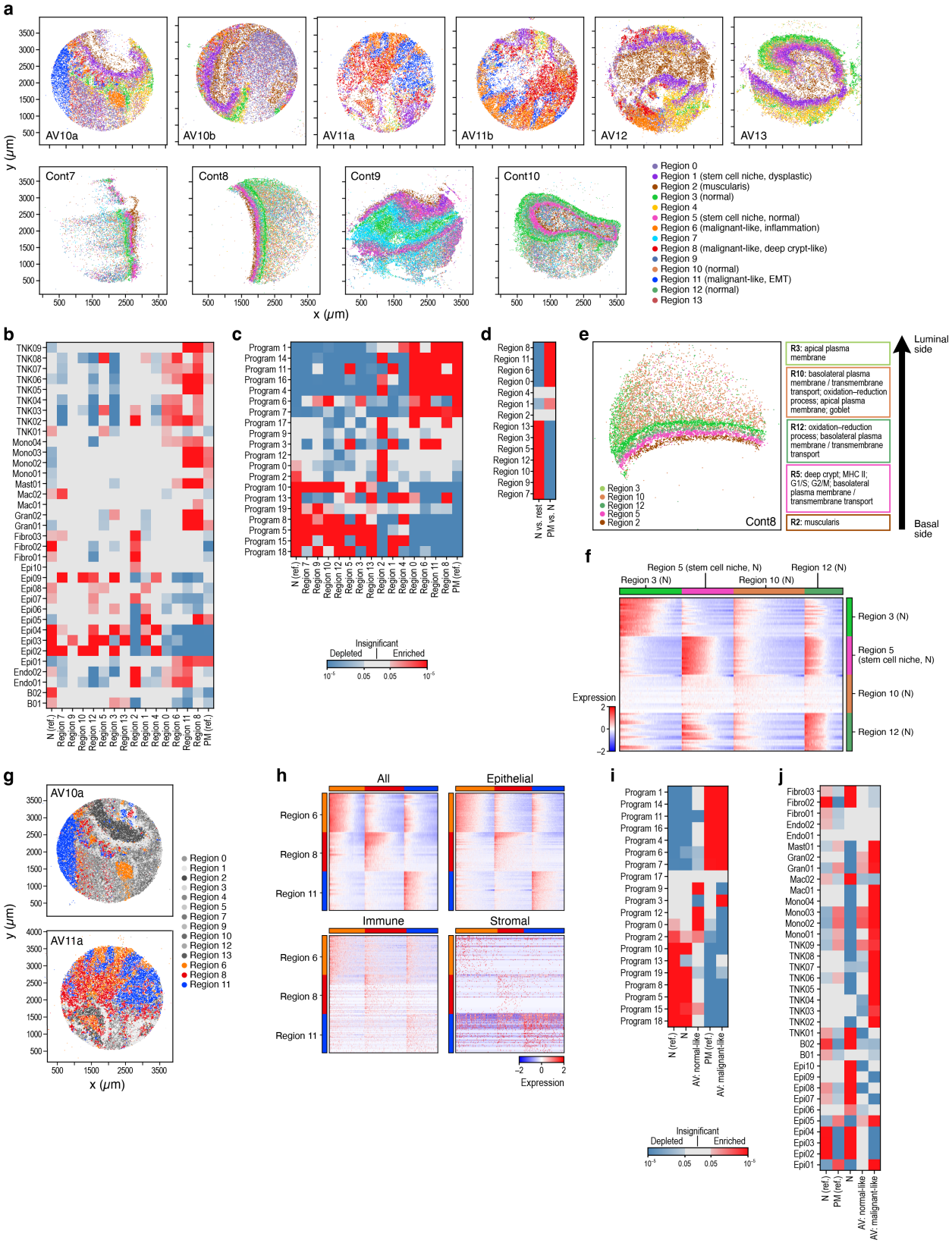


Figure 5. **Three cellular neighborhood archetypes associated with tumor progression.**

a. Spatial regions. Slide-seq pucks of premalignant (top) and normal (bottom) mouse colon colored by TACCO regions (legend) (x and y axis are spatial coordinates in μm). **b,c.** Enrichment and depletion of cell subsets and epithelial programs across different regions. Significance (P value, color bar, Welch's t-test on CLR-transformed compositions) of enrichment (red) or depletion (blue) of specific cell subsets (rows, b) or epithelial cell programs (rows, c) in the different regions defined by TACCO (columns) as well as all normal ("N (ref.)", leftmost column) and premalignant ("PM (ref.)", leftmost column) samples. **d.** TACCO defined regions preferentially relate to normal or premalignant tissue. Significance (P value, color bar, Welch's t-test on CLR-transformed compositions) of enrichment (red) or depletion (blue) of each TACCO defined region (rows) in normal ("N vs. rest") and premalignant ("PM vs. N") samples (columns). **e.** TACCO reveals normal colon architecture. Left: Slide-seq puck of normal mouse colon colored by TACCO region annotations (legend) (x and y axis: spatial coordinates (μm)). Right: Main epithelial expression programs enriched in each region (FDR<6.3 10^{-4}, Welch's t test on CLR-transformed compositions) except region 2 (muscularis), which is characterized by non-epithelial (stromal) cell types. **f.** Expression signatures of cells in normal regions 3,5,10 and 12. Scaled log-normalized expression of the top 20 differentially expressed genes (rows) for each bead (columns) in the region. **g,h**. Archetypal malignant-like regions. g. Slide-seq pucks of two premalignant lesions colored by TACCO annotations of malignant-like regions 6, 8 and 11. h. Scaled log-normalized expression of the top 20 differentially expressed genes (rows) of each bead (top, columns) in the region; or epithelial (top right), immune (bottom left) or stromal (bottom right) fractions of beads (columns) in regions 6, 8 and 11 in dysplastic lesions. **i,j.** Epithelial cell subsets and programs associated with "malignant-like", "normal-like" and normal tissues. Significance (P value, color

bar, Welch's t-test on CLR-transformed compositions) of enrichment (red) or depletion (blue) of epithelial cell programs (i, rows) or epithelial, immune and stromal cell subsets (j, rows) in different tissue types (columns) based on Slide-seq or scRNA-seq ("ref.") samples

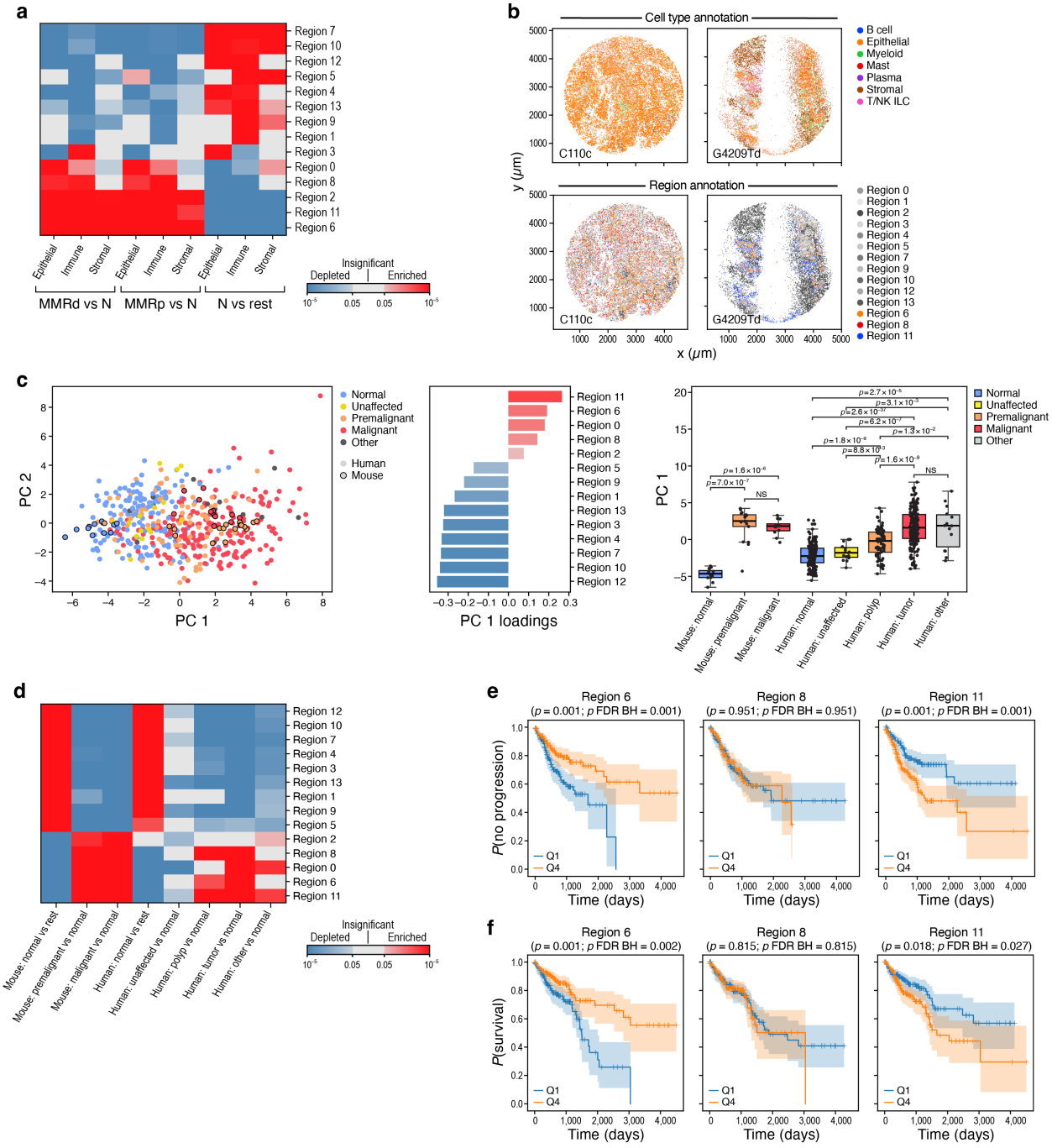
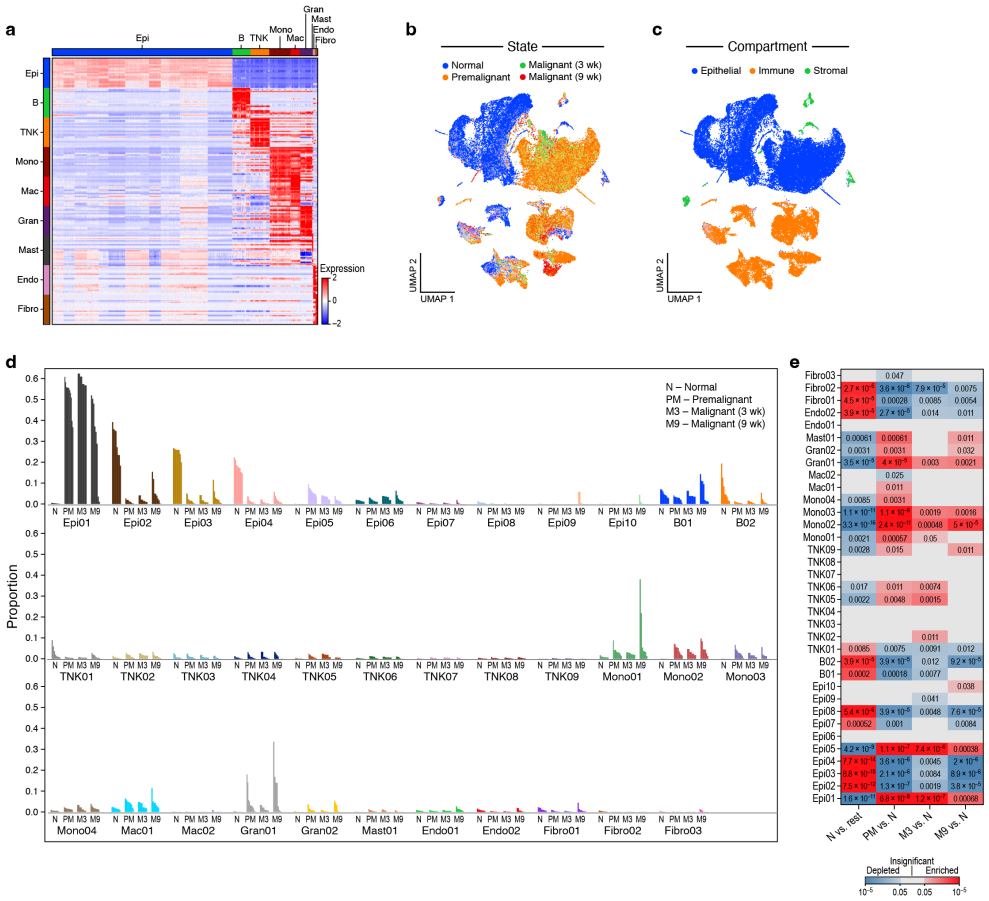


Figure 6. Tumor region archetypes associated with tumor progression in human colorectal tumors.

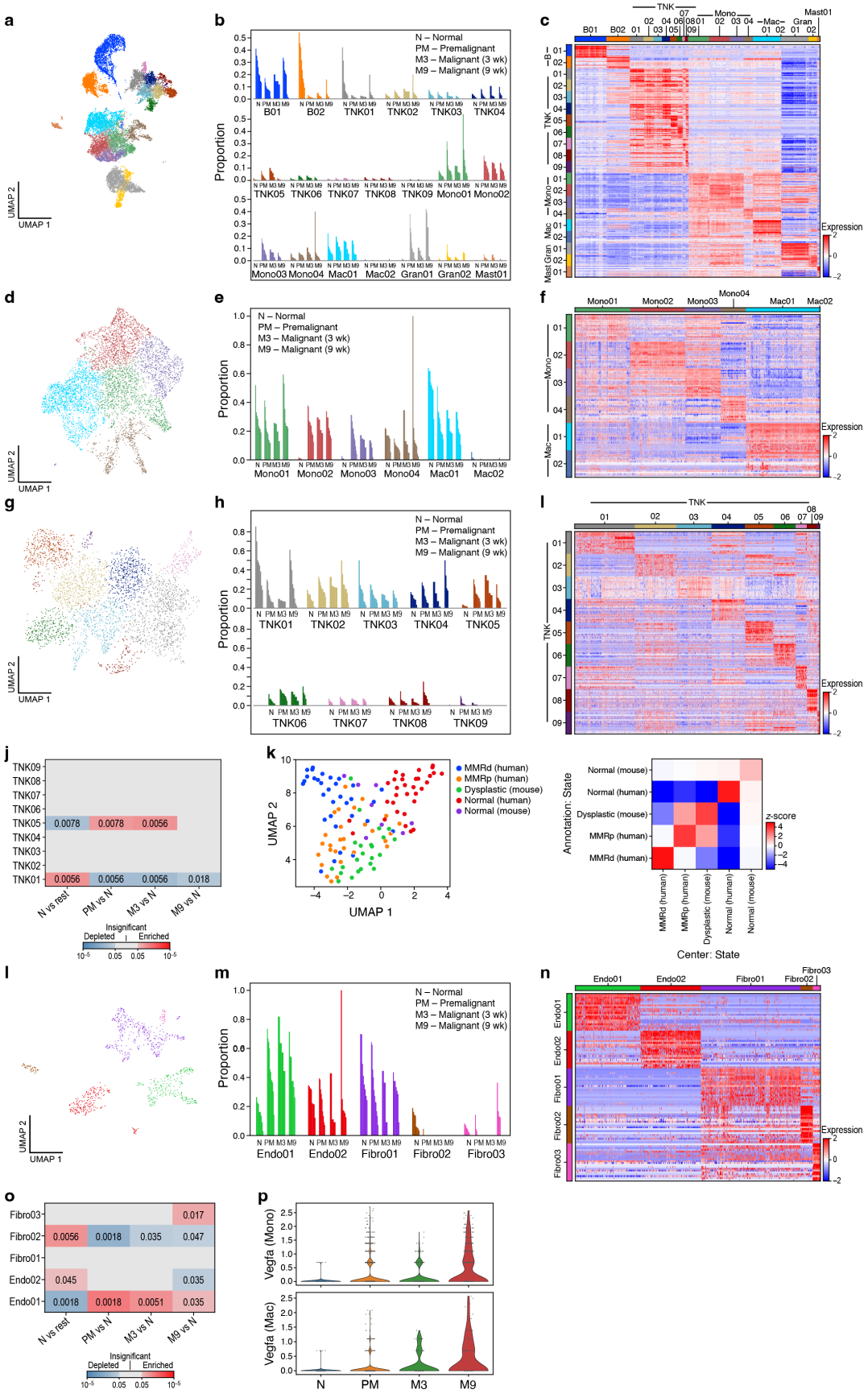
a. Expression profiles characterizing mouse regions are recapitulated in human tumors. Enrichment (red) or depletion (blue) of region-associated epithelial, immune or stromal profiles

(rows) compared between normal, MMRp, or MMRd samples (columns). **b.** Mouse defined regions are discernable in human tumor spatial data. Human cell types (top, color) or mouse regions (bottom, color: malignant-like regions, grayscale: normal-like regions) mapped to Slide-seq pucks of two MMRd tumors (x and y axis are spatial coordinates in μm). **c.** Mouse regions capture malignant features in human tumors. Left: First (PC1, x axis) and second (PC2, y axis) principal components of mouse region scores of mouse and human epithelial pseudo-bulk samples. Middle: PC1 loadings (x axis) of each mouse region score (y axis). Right: PC1 loadings (box plots show mean, quartiles, and whiskers for the full data distribution except for outliers outside 1.5 times the interquartile range (IQR)) for each type of mouse or human sample (x axis). **d.** Enrichment (red) or depletion (blue) of region-associated profile scores (rows) between normal and dysplastic samples (columns) in human or mouse. **e,f.** Expression of malignant like regions 6 and 11 in tumors are associated with PFI and OS in human patients. Kaplan-Meier PFI (e, n = 170 (Liu et al., 2018)) or OS (f, n = 140 (Liu et al., 2018)) analysis of human bulk RNA-seq cohort stratified by malignant-like region profile scores.



Supp. Figure 1. Marker genes, cell states and cell types in the healthy and dysplastic mouse colon atlas.

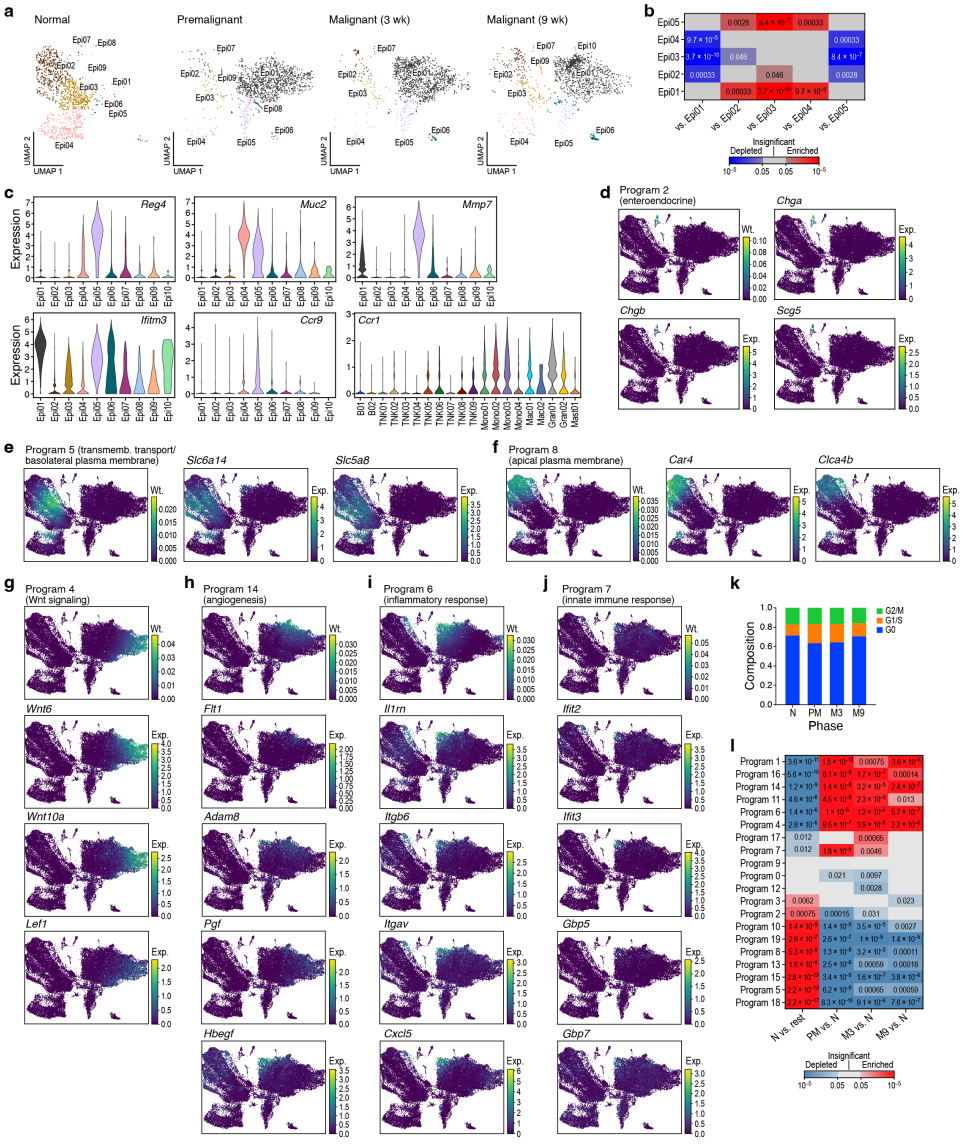
a. Cell-type expression signatures. Scaled log-normalized expression (color bar) of the top 20 differentially expressed genes (rows) in cells (columns) from each cell type. b,c. Distinct condition states and compartments. 2D embedding of all single cell profiles (dots) colored by either condition state (b) or compartment (c). **d.** Changes in cell composition between healthy and dysplastic tissue. Proportion of cells (y axis) from each cell subset (x axis) out of all cells in each sample (x axis). **e.** Changes in cell composition in dysplastic lesions. Significance (P-value, Welch's t-test on CLR-transformed compositions, color bar) of enrichment (red) or depletion (blue) of each cell subset (rows) between samples from different conditions (columns).



Supp. Figure 2. Compositional and cell intrinsic changes in stromal and immune cells.

a-c. Immune cell subsets and composition. a. 2D embedding of immune cell profiles colored by clusters (legend). b. Proportion of cells out of all immune cells (y axis) of each immune cell subset in each sample (x axis). c. Scaled log-normalized expression (color bar) of the top 20 differentially expressed genes (rows) in cells (columns) from each immune cell subset (color bar on top). **d-f**. Monocyte and macrophage cell subsets and composition. d. 2D embedding of monocyte and macrophage cell profiles colored by clusters (legend). e. Proportion of cells out of all monocytes and macrophages (y axis) of each monocyte and macrophage cell subset in each sample (x axis). f. Scaled log-normalized expression (color bar) of the top 20 differentially expressed genes (rows) in cells (columns) from each monocyte and macrophage cell subset (color bar on top). **g-i.** T/NK cell subsets and composition. g. 2D embedding of T/NK cell profiles colored by clusters (legend). h. Proportion of cells out of all T/NK cells (y axis) of each T/NK cell subset in each sample (x axis). i. Scaled log-normalized expression (color bar) of the top 20 differentially expressed genes (rows) in cells (columns) from each T/NK cell subset (color bar on top). **j.** Enrichment of IL17+ γδT cells and depletion of CD8+ γδT cells in dysplastic lesions. Significance (P value, Welch's t-test on CLR-transformed compositions, color bar) of enrichment (red) or depletion (blue) of each TNK cell subset (rows) between samples from different conditions (columns). **k**. TNK cell compositions is similar in human and mouse. Left: 2D embedding of TNK cell composition profiles of human and mouse samples colored by sample type (legend) (Methods). Right: Similarity of TNK cell composition (enrichment z-scores) in the 2D embedding between each set of samples (rows, columns). **l-n.** Stromal cell subsets and composition. l. 2D embedding of stromal cell profiles colored by clusters (legend). m. Proportion of cells out of all stromal cells (y axis) of each stromal cell subset in each sample (x axis). n. Scaled log-normalized expression (color bar)

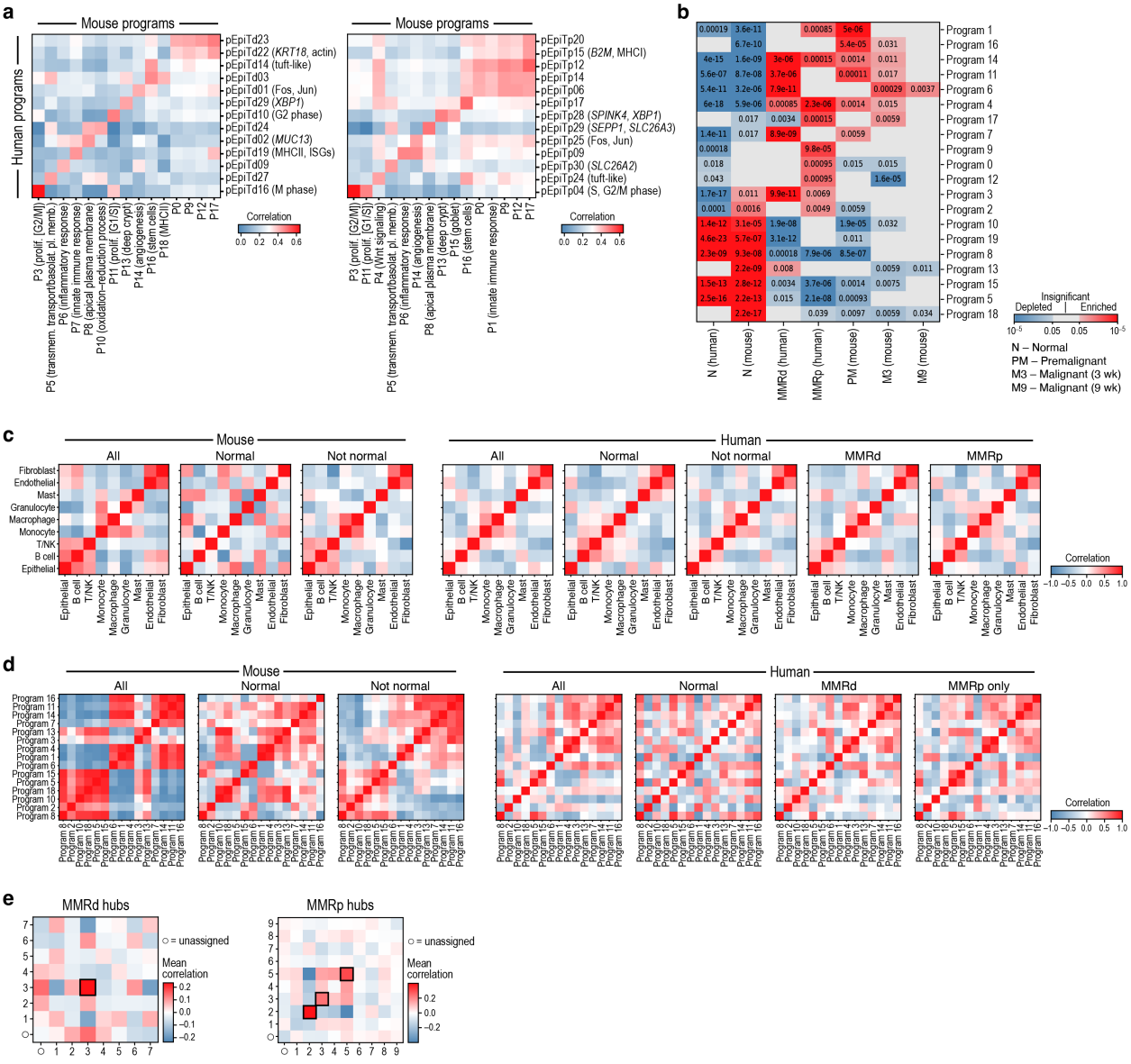
of the top 20 differentially expressed genes (rows) in cells (columns) from each stromal cell subset (color bar on top). **o**. Enrichment of vascular endothelial cells and depletion of myofibroblasts in dysplastic lesions. Significance (P value, Welch's t test on CLR-transformed compositions, color bar) of enrichment (red) or depletion (blue) of each stromal cell subset (rows) between samples from different conditions (columns). **p.** Increased VEGFA expression in monocyte-macrophage populations with dysplasia. Distribution of expression (y axis, log1p(counts)) of *VegfA* in monocytes and macrophages from different conditions (x axis).



Supp. Figure 3. Changes in cell composition and expression programs usage in dysplastic epithelium.

a,b. Changes in epithelial cell composition in dysplastic tissues. a. 2D embedding of all single cell epithelial profiles showing only the profiles (dots) of cells from each condition state, subsampled to equal numbers of cells per condition state, colored by cluster. b. Significance (P value, Welch's t test on ALR-transformed compositions with all non-tdTomato counts used as reference compartment, color bar) of enrichment (red) or depletion (blue) of tdTomato expression in cells

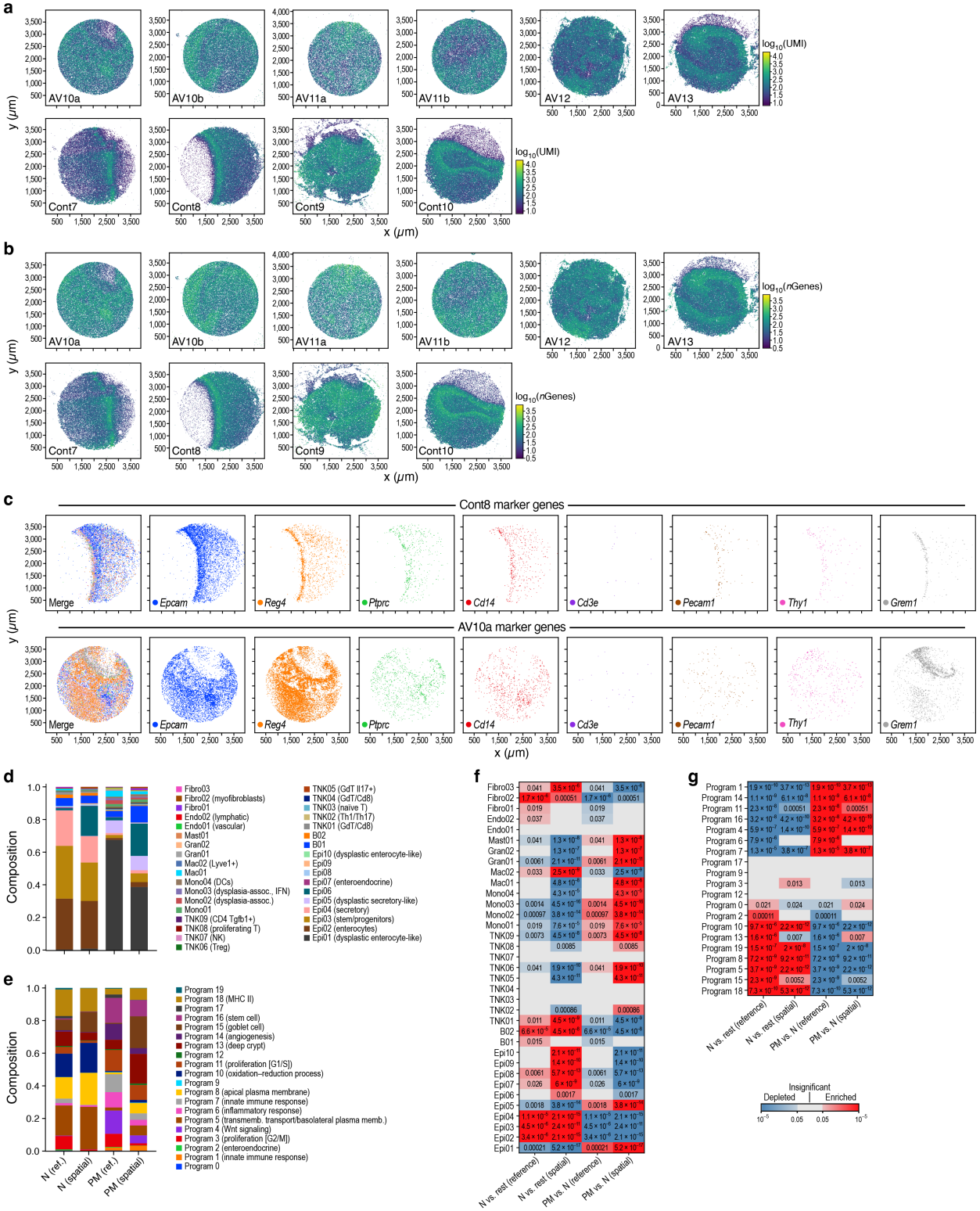
from malignant samples between every pair of epithelial clusters. **c.** Dysplastic secretory like (Epi05) cells express tumor-related genes. Distribution of expression (y axis) of different marker genes in cells from each epithelial/immune cell cluster (x axis). **d-j**. Epithelial gene programs. 2D embedding of all epithelial cells colored by the weight of each program (color bar) or the expression of selected program genes (color bar). **k.** Increase in G1/S cells in dysplastic lesions. Proportion of epithelial cells (y axis) assigned to each phase of the cell cycle in each condition (x axis). **l.** Epithelial programs characteristics of normal and dysplastic colon. Heatmap significance (P value, Welch's t test on CLR-transformed compositions, color bar) of enrichment (red) or depletion (blue) of each epithelial program (rows) between normal vs. dysplastic tissues (columns).



Supp. Figure 4. Conservation of cellular composition and expression programs between mouse and human scRNA-seq data.

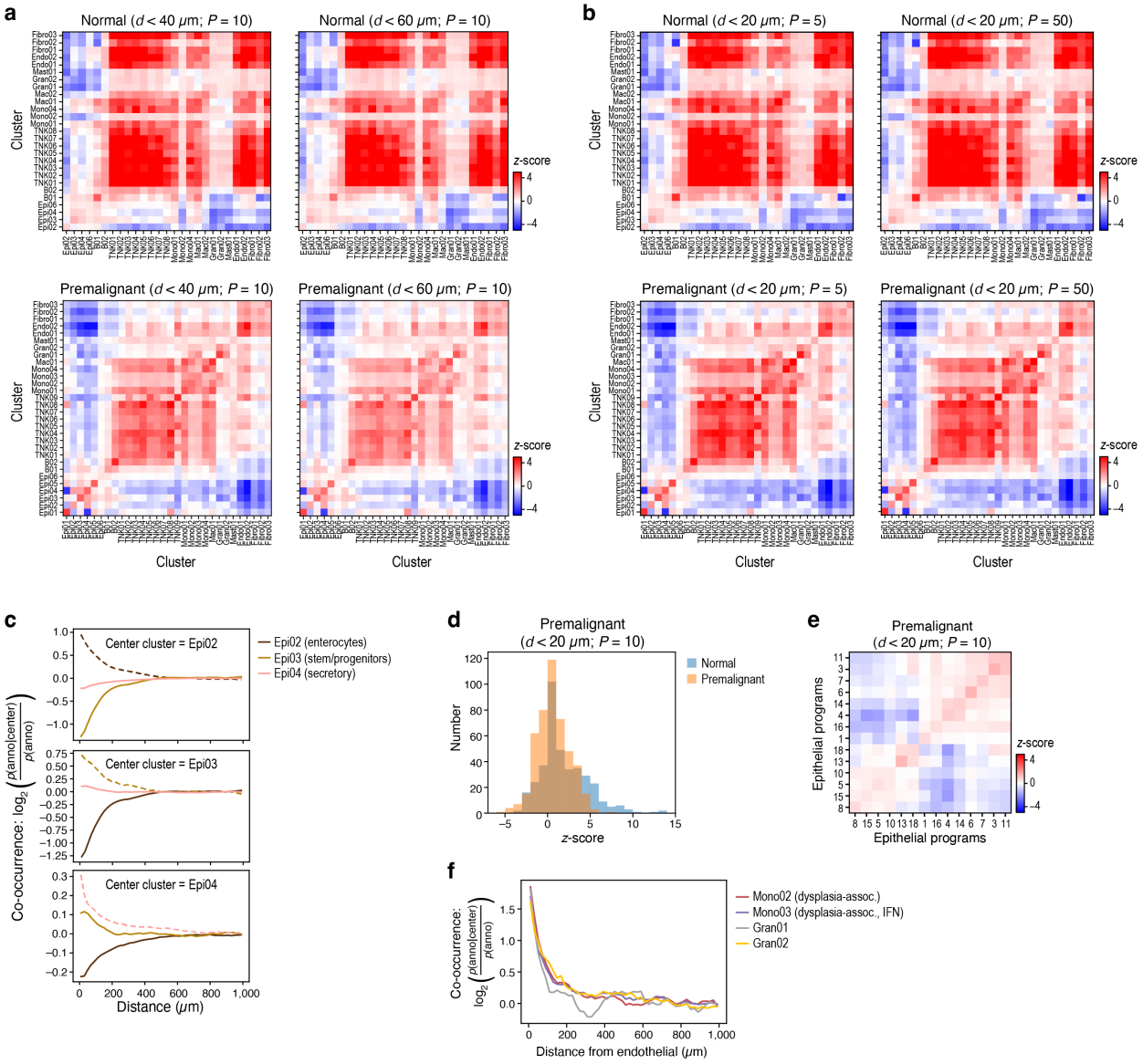
a. Conservation of subtype-specific human and mouse epithelial expression programs. Pearson correlation coefficients (color) between program-specific expression profiles (Methods) of human MMRd- (left) or MMRp- (right) specific programs (rows) and mouse programs (columns). **b.** Mouse epithelial program enrichments in mouse and human tumor and normal samples. Significance (Welch t-test on CLR-transformed compositions) of enrichment (red) or depletion

(blue) of mouse epithelial program expression (rows) in sample classes from human or mouse (columns). **c,d.** Human-mouse conservation of cell type and program associations. Pearson correlation coefficients (color) of the CLR-transformed cell type (c) or epithelial program (d) compositions across samples in mouse (left) or human (right) single cell data. In (d), data are hierarchically clustered for the "not normal" mouse case (premalignant and malignant) and this ordering is applied to all other panels. **e.** Human multicellular hubs conserved in mouse tissue. Mean of the Pearson correlation coefficients (color) between the per sample composition profiles of each human expression program in mouse samples (as in **Fig. 3b**) calculated from all the programs in each pair of hubs defined in human data (excluding the diagonal (program against itself)). Black border: hubs highlighted in **Fig. 3b**.



Supp. Figure 5. **Spatial distributions of cells and programs across regions.**

a,b. Slide-seq quality controls. Slide-seq pucks of premalignant (top) and normal (bottom) mouse colon colored by number of UMIs (A) or of genes (B) per bead. (x and y axis: spatial coordinates (μm)). **c.** Selected marker gene expression. Slide-seq pucks from a normal (top) and premalignant (bottom) sample, colored by marker gene detected per bead. **d,e.** Spatial mapping of cell types and programs yields comparable composition to scRNA-seq. Distribution of the proportion (y axis) of contributions to each cell type (d) or program (e; based on fractional annotations) in cells (for scRNA-seq; "ref") or beads (for Slide-seq; "spatial"; based on fractional annotations) in samples from normal (N) or premalignant (PM) tissue. **f,g.** Distinct cell types and programs associated with premalignant and normal colon. Significance (P value, Welch's t test on CLR-transformed compositions, color bar) of enrichment (red) or depletion (blue) of cell types (f, rows) or epithelial programs (g, rows) in normal (N) or premalignant (PM) tissues based on Slide-seq ("spatial") data or scRNA-seq ("reference").

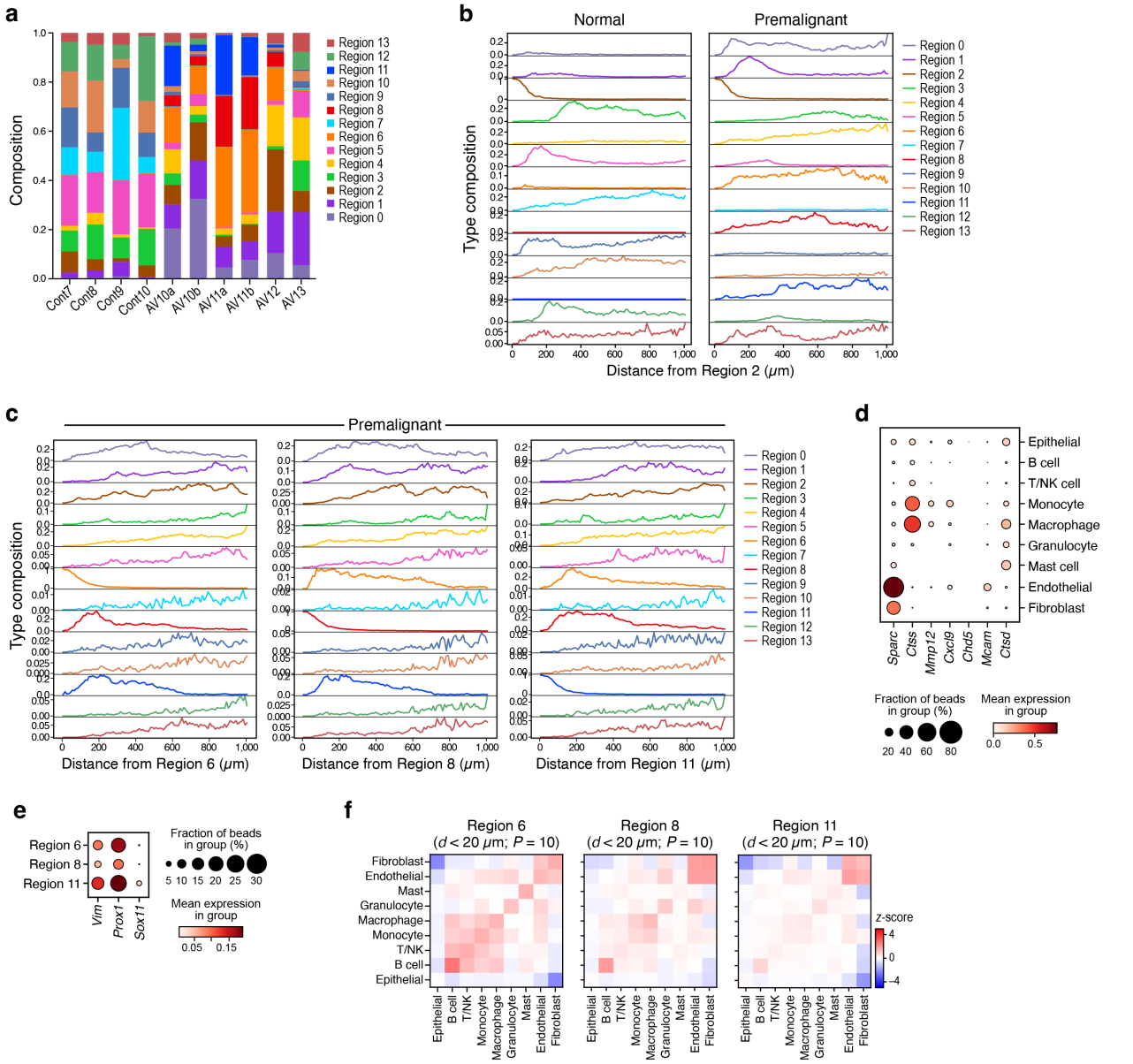


Supp. Figure 6. Distinct cellular layout in normal and premalignant tissues.

a,b. Neighborship analysis is robust to the distance and number of randomizations. Neighborship enrichment (z-scores, color) *vs*. a background of spatially random annotation assignments for each pair of cell annotations (rows, columns) in normal (top) and premalignant (bottom) samples at varying distance (a, left: ≤40μm; right: ≤60 μm) or number of permutations (b, left: 5, right: 50).

c. Spatial arrangement of epithelial cells in normal colon. Co-occurrence (y axis) of normal epithelial cell types (color) at different distances (x axis) from different central normal epithelial

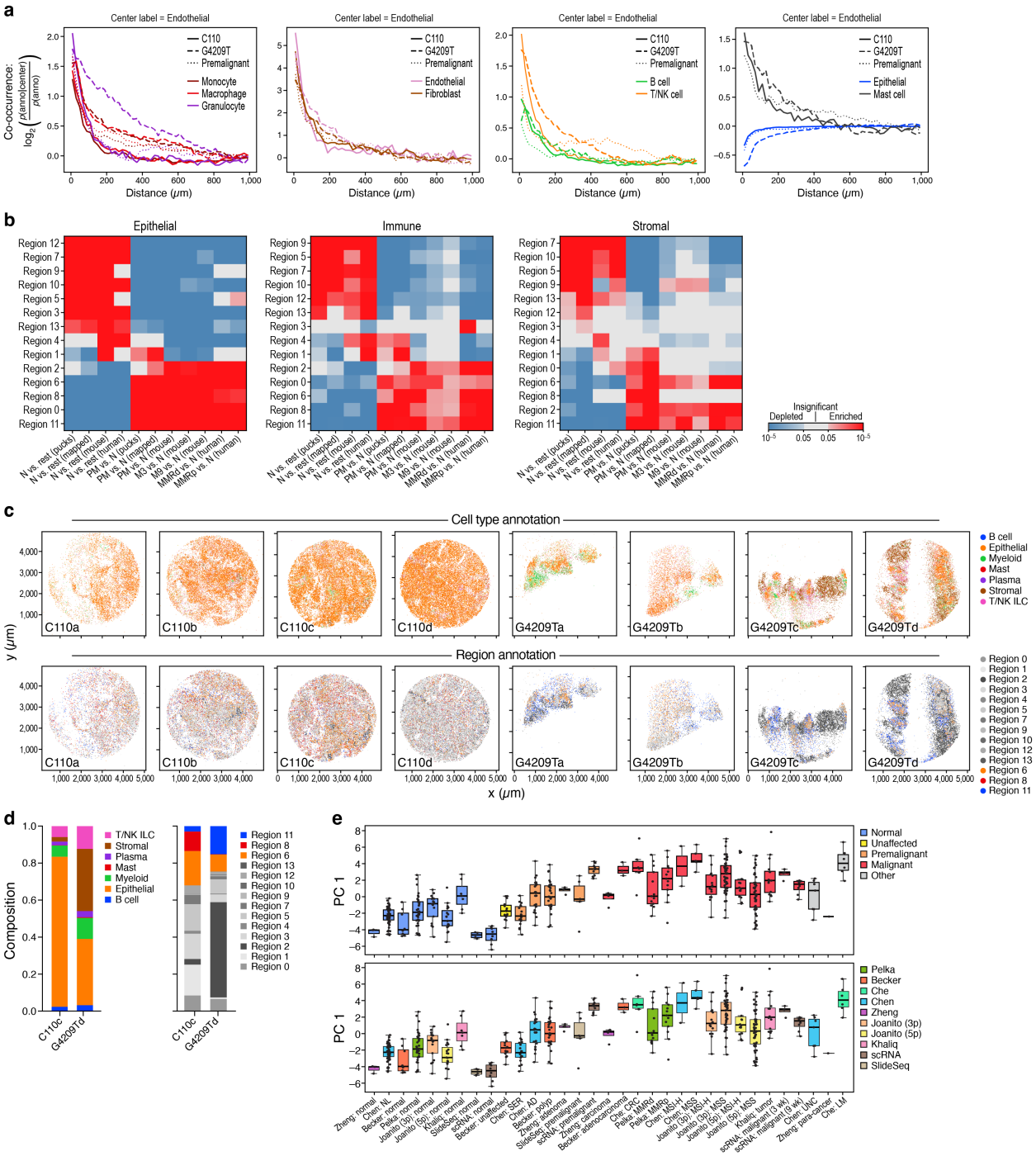
cell type (annotated on top). **d.** Decreased spatial order of cell types in premalignant vs. normal tissue. Distribution of enrichment z-values (x axis) of cluster-cluster interactions (as shown in Fig. 4a,b) in premalignant lesions (orange) and normal colon (blue). **e.** Epithelial program neighborships in premalignant tissue. Short-range (≤20μm) neighborship enrichment z-scores (color) *vs*. a background of spatially random annotation assignments for each pair of epithelial program annotations (rows, columns) in premalignant lesions. **f.** Spatial arrangement of dysplasia-associated immune cells relative to endothelial cells. Co-occurrence (y axis) of dysplasia-associated monocyte or granulocyte cells (color) with endothelial cells at different distances (x axis).



Supp. Figure 7. Cellular neighborhood archetypes.

a. Distinct characteristic region distributions in healthy and premalignant tissue. Proportion of UMIs assigned to each region (y axis) in each Slide-seq puck (x axis). **b,c.** Spatial relations between the regions. Proportion of beads (y axis) of each region category (color code) at different distances (x axis) from region 2 (b, muscularis), 6 (c, left), 8 (c, middle) or 11 (c, right). **d,e.** Fraction of expressing cells (dot size) and mean expression per celltype (d dot color) or per region

(e, dot color of marker genes (columns) across cell types in region 6 (d, rows) or regions 6, 8 and 11 (e, rows). **f.** Cell type neighborships in different malignant regions. Short-range (≤20μm) neighborship enrichment z-scores (color) *vs*. a background of spatially random annotation assignment for each pair of cell type annotations (rows, columns) in malignant-like regions 6, 8 and 11 within premalignant lesions.



Supp. Figure 8. Transfer of mouse spatial region expression profiles to human patient data.

a. Spatial arrangement of cell types in human and mouse dysplastic lesions. Co-occurrence (y axis) of different cell types with endothelial cells at different distances (x axis) for two human tumors

(solid and dashed lines) and mouse premalignant lesions (dotted lines). **b.** Expression profiles characterizing mouse regions in human and mouse samples. Significance (Welch t test on CLR-transformed compositions) of enrichment (red) or depletion (blue) of mouse region-associated epithelial, immune or stromal profiles (rows) in pucks from normal colon and dysplastic lesions, in the same pucks but after mapping the region annotation to itself (consistency check), in mouse single cell data after mapping region annotation from mouse pucks, and in human single cell data after mapping region annotation from mouse pucks. **c,d.** Mouse defined regions discernable in human tumor spatial data. c. Human cell types (top, color) or mouse regions (bottom, color: malignant-like regions, grayscale: normal-like regions) mapped to Slide-seq pucks of two MMRd tumors (x and y axis are spatial coordinates in μm). d. Composition (y axis) of cell types (left) and regions (right) in each human tumor puck (x axis). e. PC1 scores (y axis; box plots show mean, quartiles, and whiskers for the full data distribution except for outliers outside 1.5 times the interquartile range (IQR)) in a PCA of region scores of mouse and human samples (x axis) sorted by malignant status and colored by status (top, legend) or study (bottom, legend).

Methods

Human subjects

The MGH Institutional Review Board approved protocols for tissue collection used for sequencing (Protocol 02-240). Informed consent was obtained from all subjects prior to collection. Only patients with primary treatment- naive colorectal cancer were included in this study.

Mice

Mice were housed in the animal facility at the Koch Institute for Integrative Cancer Research at MIT. All animal studies described in this study were approved by the MIT Institutional Animal Care and Use Committee (Protocol 1213-106-16**)**. *Apc*$^{fl/fl}$ mice (Kuraguchi et al., 2006) were obtained from NCI mouse repository, *Kras*$^{LSL\text{-}G12D/+}$ Ref. (Johnson et al., 2001), *Rosa26*$^{LSL\text{-}tdTomato}$ Ref. (Madisen et al., 2009) and Trp53$^{fl/fl}$ Ref. (Marino et al., 2000) mice obtained from Jackson, *Villin*^{CreERT2} Ref. (el Marjou et al., 2004) mice were a gift from Dr. Sylvie Robine. All mice were maintained on C57BL/6J genetic background. Approximately equal numbers of male and female mice of 6–10 weeks of age were used for all experiments. Where indicated, mice were injected to the submucosal layer of the colon with 4-hydroxytamoxifen (EMD Millipore # 579002) dissolved in ethanol at a concentration of 100 μM (for the mice that were kept for 3 weeks after injection) or 30 μM (for the mice that were kept for 9 weeks after injection). Tumors were resected at either 3 or 9 weeks after 4-hydroxytamoxifen injection. Colonoscopy and colonoscopy-guided injection methods were previously described in detail (Roper et al., 2017, 2018).

Tissue processing for scRNA-seq

Single-cell suspensions from healthy colon or dysplastic lesions were processed using a modified version of a previously published protocol (Smillie et al., 2019). Tissue samples were rinsed in 30ml of ice-cold PBS (ThermoFisher 10010-049), chopped to small pieces and washed twice in 25 ml PBS, 5mM EDTA (ThermoFisher AM9261), 1%FBS (ThermoFisher 10082-147). To prime tissue for enzymatic digestion, samples were incubated for 10 minutes at 37°C, placed on ice for 10 minutes before shaking vigorously 15 times followed by supernatant removal. Tissues were placed into a large volume of ice-cold PBS to rinse prior to transferring to 5ml of enzymatic digestion mix (Base: RPMI1640, 10 mM HEPES (ThermoFisher 15630-080), 2% FBS), freshly supplemented immediately before use with 100 mg/mL of Liberase TM (Roche 5401127001) and 50 mg/mL of DNase I (Roche 10104159001), and incubated at 37°C with 120 rpm rotation for 30 minutes. After 30 minutes, enzymatic dissociation was quenched by addition of 1ml of 100% FBS and 10mM EDTA. Samples were then filtered through a 40 mM cell strainer into a new 50 mL conical tube and rinsed with PBS to 30 mL total volume. Tubes were spun down at 400 g for 7 minutes, at 4°C. Resulting cell pellets were resuspended in 1ml PBS, placed on ice and counted.

Cell hashing

Cell hashing was performed based on the published protocol (Stoeckius et al., 2018) as summarized below. Dissociated cells were resuspended in 1ml of Cell Hashing Staining Buffer (1× PBS with 2% BSA (New England Biolabs, B9000S) and 0.02% Tween (Tween®-20 Solution, 10%, Teknova, VWR-100216-360) and counted. 500,000 cells were resuspended in 100 µL of Cell Hashing Staining Buffer and incubated for 30 minutes on ice, with 2 µL of the appropriate BioLegend TotalSeq™ Hashing antibody (a 1:50 dilution, using a total of 1 µg of antibody per cell suspension). TotalSeq™-A anti-mouse Hashtag antibodies #1-8 (catalog numbers:155801,

155803, 155805, 155807, 155809, 155811, 155813, 155815) were used. Cells were washed three times with 0.5 mL of Cell Hashing Staining Buffer and filtered through low-volume 40-µm cell strainers. All cell suspensions were recounted to achieve a uniform concentration of 7,000 cells per microliter before pooling for capture by 10x Chromium controller following the manufacturer protocol for the v2 or v3 3' kit (10X Genomics, Pleasanton, CA).

Hashtag oligo (HTO) library preparation

Separation of hashtag oligo (HTO)-derived cDNAs (<180 bp) and mRNA-derived cDNAs (>300 bp) was done after whole-transcriptome amplification by performing 0.6× SPRI bead purification (Agencourt) on cDNA reactions as described in 10x Genomics protocol. Briefly, supernatant from 0.6× SPRI purification contains the HTO fraction, which was subsequently purified using 1.4 and 2× SPRI purifications per the manufacturer's protocol (Agencourt). HTOs were eluted by resuspending SPRI beads in 15 µL TE. Purified HTO sequencing libraries were then amplified by PCR (1µl clean HTO cDNA, 25µl 2X NEBNext Master Mix (NEB #M0541), 10 µM SI-PCR and D701or D704 primers performed dial out PCR (98°C (10 sec), (98°C for 2 sec, 72°C for 15 sec) x 12/18 then 72°C for 1 min) for 12 and 18 cycles, and used the 18 cycles product for sequencing. PCR reactions were purified using another 2× SPRI clean up and eluted in 15 µL of 1× TE. HTO libraries were quantified by Qubit High sensitivity DNA assay (ThermoFisher) and loaded onto a BioAnalyzer high sensitivity DNA chip (Agilent).

SI-PCR: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC*T*C

D701 :

CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGC

D704 :

CAAGCAGAAGACGGCATACGAGATGGAATCTCGTGACTGGAGTTCAGACGTGTGC

Sequencing

Samples were sequenced using HiSeq X (Illumina). Hashing libraries were sequenced with spike-ins of 2.5%.

Tissue processing for Slide-seq

Colons were flashed with cold PBS and a segment that includes the lesion and surrounding tissue (or a respective healthy segment from normal mice) was dissected. Samples were then mounted in cold OCT, flash frozen on dry ice covered with ETOH until and long-term stored in -80℃. For human experiments, fresh human tumor samples were snap frozen in Tissue Tek Optimal Cutting Temperature (Sakura) and stored at -80°C until sectioning.

Slide-seq

For mouse and human experiments, 10 μm sections were cut and the Slide-seq V2 protocol was used as previously described (Stickels et al., 2021). For mouse experiments, four and six arrays were collected from normal colons and premalignant lesions respectively. The muscularis was fit onto the array of both healthy and dysplastic lesions to allow appropriate orientation.

scRNA-seq pre-processing and quality control filtering

Count matrices for scRNA-seq were generated using the Cumulus feature barcoding workflow v0.2.0(Li et al., 2020) with CellRanger v3.1.0 and the mm10_v3.0.0 mouse genome reference. Cell profiles were quality filtered by requiring between 1,000 and 50,000 counts, and between 500

and 7,000 genes, less than 20% mitochondrial counts, and less than 10% hemoglobin counts. Cell profiles that did not meet all these criteria were discarded. The top 5,000 highly variable genes were annotated on the remaining cells after normalization to 10,000 counts and log1p transform using Scanpy's "highly_variable_genes" function (Wolf et al., 2018) and providing the chemistry (v2/v3) by hashing (True/False) combination as batch-annotation. Putative doublets were removed using Scrublet (Wolock et al., 2019) with default parameters.

Selection of variable genes, dimensionality reduction and clustering

A preliminary clustering using the Leiden algorithm with resolution 1.0 was performed after normalization to 10,000 counts, log1p transform, correction for number of counts and percentage of mitochondrial genes, scaling with a max_value of 10, and generating a *k*-nearest neighbors (*k*-NN) graph with 15 neighbors on a PCA of the previously annotated 5000 highly variable genes with 50 components using Scanpy (Wolf et al., 2018). The single cell profiles were provisionally annotated with SingleR (Aran et al., 2019) cell-wise (i.e. without using clustering information) using the SingleR built-in MouseRNAseqData and an intestine specific dataset from Tabula Muris (Tabula Muris Consortium et al., 2018). https://figshare.com/ndownloader/files/13092143. For further processing, the dataset was then split into the three compartments, epithelial, immune and stromal, using the provisional SingleR annotations.

For each compartment, the top 5,000 highly variable genes were annotated using Scanpy's "highly_variable_genes" function on cells normalized to 10,000 counts after log1p-transformation and providing the chemistry (v2/v3) by hashing (True/False) combination as batch-annotation.

Expression programs and batch correction

For the dataset of each compartment separately (generated as described above), an integrative NMF was performed (using a part of the LIGER implementation) with k=20 and lambda=5 to identify 20 programs and their respective weights per cell. The same approach was also used with a higher *k* (epithelial and immune: 200, stromal: 50) to yield a detailed and batch corrected decomposition of expression which was then combined to obtain a count-like corrected expression matrix for the top 5,000 highly variable genes. For each compartment separately, these batch-corrected data were normalized to 10,000 counts, log1p transformed, corrected for number of counts and percentage of mitochondrial genes by linear regression, scaled with a max_value of 10, followed by a PCA of the previously annotated 5,000 highly variable genes. A a *k*-nearest neighbors (*k*-NN) graph was constructed from the top 50 PCs, with *k*=15 neighbors using Scanpy, and clustered using a compartment-specific Leiden resolution parameter (epithelial: 0.2, immune: 0.4, stromal: 0.1). This clustering was used as the cluster level annotation of the mouse scRNA-seq data for the epithelial and stromal compartment. Separately per compartment the data were annotated with SingleR using the cluster information. The same per-compartment batch-corrected and preprocessed data from the Leiden clustering was used to create UMAP embeddings with PAGA initialization using Scanpy.

To improve the clustering and annotation in the immune compartment and to filter out additional doublets not detected by Scrublet, the immune data were separately filtered and clustered using information from the compartment level clustering and annotation. To that end, myeloid and TNK cells were partitioned separately and further processed, and additional likely doublet cells were labeled and removed by the following procedure:

1. Cells were labeled as doublets based on higher number of UMIs of marker genes for other compartments than the 95th percentile observed in this immune partition (*i.e.*, Epcam and Cdh1 to remove immune-epithelial doublets and Cav1 and Kdr to remove immune-stromal doublets) and other immune partitions (i.e., Cd3d, Cd3e, and Cd3g to remove myeloid-lymphoid doublets from the myeloid cells). This type of filter criterion for lowly expressed genes ("larger than some percentile" on integer counts) also allows to keep more than 95% of the cells if, for example, all cells of this partition happened to have 0 UMIs of a particular marker gene.
2. Cell were labeled as doublets if they had inconsistent cell-wise and cluster-wise SingleR annotations.
3. Cells were labeled as doublets if they had significantly (Benjamini-Hochberg FDR=0.05, Fisher's exact test) more neighbors in the k-NN graph from the immune compartment that were already marked as doublets.
4. All cells labeled as doublets were removed.

After filtering, the count matrices were batch corrected as above using the integrative NMF from LIGER with k=20 and lambda=5, and clustered like above with group specific Leiden resolution (myeloid: 0.2, TNK: 0.4). For myeloid and TNK cells, this clustering superseded the original clustering. The integrative NMF result here was only used for updating the clustering and not for generating an extra set of expression programs.

Cell cycle phase annotation

Annotation of single cell profiles with a cell cycle phase was performed with scvelo's (Bergen et al., 2020) score_genes_cell_cycle function. The resulting cell cycle annotation was corrected

according to the original gene list annotation from (Tirosh et al., 2016): “G1” was renamed to “G0” and “S” was renamed to “G1/S”.

Selection of human single cell data for the comparison of cell type and epithelial program composition

ScRNA-seq data from Ref (Pelka et al., 2021) was used as reference for human CRC. To avoid biases in cell type compositions, only the subset of the data where "PROCESSING_TYPE = unsorted" were used.

Comparison of human and mouse samples by cell type composition

To compare human and mouse samples by composition of T/NK cell subsets, T/NK annotations from mouse and human data (Pelka et al., 2021) were matched by TACCO (Mages. et al., 2022), using optimal transport (OT). First, human expression data were mapped to mouse genes using MGI homology information [subsection “Mapping of mouse and human orthologs”]. Then, human cell cluster annotations ('cl295v11SubFull') (Pelka et al., 2021) were mapped from the subset of human cells annotated as T/NK/ILC to the subset of mouse cells annotated as T/NK using TACCOs "annotate" function with OT as core method, basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 4 iterations of bisectioning with a divisor of 3. Annotation with maximum probability per cell was used as the unique cluster level annotation for mouse T/NK cells. Annotations were aggregated per sample to yield a compositional annotation over the identical cluster annotation categories (from the human dataset) for the T/NK subsets of human and mouse samples. Annotations vectors were then processed using the sc.pp.neighbours and sc.tl.umap functions from Scanpy (Wolf et

al., 2018) to yield a 2D sample embedding with respect to T/NK cell composition. Using the coordinates in the UMAP in place of spatial coordinates, neighborship enrichment z-scores were computed with TACCO's co_occurrence_matrix function with max_distance=2 and n_permutation=100.

Slide-seq annotation

For the compositional annotation of Slide-seq beads with the categorical cell clusters from the single cell data, the "annotate" function of TACCO with OT was used as core annotation method per Slide-seq puck with the subset of the single cell data with matching disease state, with basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.001, and 4 iterations of bisectioning with a divisor of 3.

For the compositional annotation of Slide-seq beads with the compositional epithelial programs, the annotated beads were split using the "split_observations" function of TACCO on the cluster-level annotation, aggregated to compartment level, and the epithelial part was then annotated using again the "annotate" function with OT as core annotation method, basic platform normalization, entropy regularization parameter epsilon 0.01, and a marginal relaxation parameter lambda of 0.01.

Region annotation was done for all pucks (with normal and premalignant pucks) in one step to get comparable region annotations across pucks. This is done with the "find_regions" function of TACCO, using a position weight of 0.7, a Leiden resolution of 1.3, and 15 nearest neighbors per bead in position space and epithelial program space. To determine the neighbors in epithelial program space, the square-roots of the program weights were used for neighbor finding which

effectively uses Bhattacharyya coefficients as overlap in epithelial program space instead of the Euclidean scalar products used for position space. These regions are defined by construction only on beads with a large enough epithelial contribution and are then extended to all beads by assigning unannotated beads the region from the nearest bead with region annotation.

To determine region composition at a certain distance of a reference region, TACCO's "annotation_coordinate" function is used with max_distance=1000 and delta_distance=10.

Region- and cell type- characterizing genes in Slide-seq data

Genes to characterize regions on Slide-seq pucks irrespective of compartment composition were found using Scanpy's rank_genes_groups function on the full bead expression profiles. To find them separately for each compartment, the compartment-level split beads [sub-section "Slide-seq annotation"] were used instead of the full beads. To compare gene expression between cell types on Slide-seq pucks, cluster-level split beads [sub-section "Slide-seq annotation"] were aggregated to cell type level.

Cell-type neighborships in Slide-seq data

To evaluate the local cell-type neighbourship relations in the different disease states on the cluster level, the clusters were filtered per disease state to contain only clusters which account for at least 1% of the UMIs in that state. Then neigbourhood-enrichment z-scores were calculated using TACCO's "co_occurrence_matrix" function with max_distance=20 and n_permutation=10. To evaluate the stability of the result, this is also repeated for (max_distance, n_permutation)=(40,10),(60,10),(20,5), and (20,50). To get the significance of the overall change

in z-scores between the states, a Mann-Whitney-U test was performed on the values of the upper triangular half of the matrix between the two disease states for (max_distance, n_permutation)=(20,10).

A similar neighbourship analysis was performed on the coarser cell-type level separately for the three malignant regions 6, 8, and 11, using TACCO's "co_occurrence_matrix" function with max_distance=20 and n_permutation=10.

Cell-type co-occurrence in Slide-seq data

Cell-type compositions relative to a spatial landmark, Region 2=muscularis, was evaluated using TACCO's "annotation_coordinate" function with max_distance=1000 and delta_distance=10. To reduce tissue structure bias from the muscularis, the distance dependency of cell-type frequency relations was evaluated only for beads deep in the "epithelial domain", defined as follows. The effective distance from stromal annotation was computed using TACCO's "annotation_coordinate" function (with max_distance=100, delta_distance=10, critical_neighbourhood_size=4.0) and only beads with a distance of at least 75µm were used. On these remaining beads, TACCO's "co_occurrence" function was used (with delta_distance=20, max_distance=1000) to compute cell types co-occurrence as a function of their distance.

Epithelial program neighborships in Slide-seq data

As for cell types above, neighborship relations were evaluated for epithelial programs in the premalignant Slide-seq samples using TACCO's "co_occurrence_matrix" function with

max_distance=20 and n_permutation=10, after selecting only the programs which make up at least 1% of the UMIs in the premalignant Slide-seq samples.

Mapping of mouse and human orthologs

We mapped human to mouse genes using the ortholog mapping from Mouse Genome Informatics (http://www.informatics.jax.org/homology.shtml), downloaded April 26th, 2021.

Comparison between mouse and human programs

To compare mouse and human epithelial expression programs (Pelka et al., 2021), genes were mapped to mouse homologs using MGI homology information [subsection "Mapping of mouse and human orthologs"]. Mouse and human programs were then characterized by a single vector of mean expression per program in mouse gene space. Specifically, both mouse and human programs were defined such that their weighted sum approximates the expression profiles of the cells without any transformations. Programs and weights were normalized to sum to 1. To reduce batch-effects (including species-specific ones), a background expression profile was defined for each species dataset as the pseudo-bulk epithelial expression profile in the respective scRNAs-seq data. Program and background profiles were normalized to 10,000 counts and the log ratio of the normalized program and background expression vectors was used to define a vector for each species. Pearson correlation coefficients were calculated for each pair of program vectors (mouse vs. human).

Human expression program associations across mouse scRNA-seq

All sets of programs that were previously used to define human CRC tissue hubs (Pelka et al., 2021) (epithelial, T/NK cells and myeloid cells) were mapped to mouse genes with MGI homology

information [subsection "Mapping of mouse and human orthologs"] and then to mouse single cell data with TACCO, using TACCOs platform normalization to account for batch effects. The "annotate" function in TACCO was used with OT as the core annotation method, on the comparable subsets of cells from mouse and human single-cell datasets (*e.g.*, myeloid cells from mouse and human), with basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 4 iterations of bisectioning with a divisor of 3, and flat annotation prior distribution. The resulting probabilistic per-cell program annotations were aggregated to get probabilistic per-sample program annotations for all dysplastic mouse samples and CLR-transformed. For each pair of programs, the Pearson correlation coefficient was calculated on these transformed values. Using the published mapping from programs to hubs (Pelka et al., 2021), these per program-pair correlation coefficients were aggregated per hub-pair (excluding the values of identical programs) by computing their mean.

Annotating human scRNA-seq data with mouse-derived region information

Human scRNA-seq profiles (Pelka et al., 2021) were mapped to mouse gene space using MGI homology information [subsection "Mapping of mouse and human orthologs"]. Working in the same expression space, the "annotate" function in TACCO with OT as core annotation method was used on the full human scRNA-seq and mouse Slide-seq dataset with basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 7 iterations of bisectioning with a divisor of 3, and 10-fold sub-clustering of the region annotations. The region transfer is done separately per compartment, with the Slide-seq compartment split as described above and the human scRNA-seq data split using the cell type

annotation of the data. For validation, mapping was also performed with mouse scRNA-seq data, as well as mapping the region information from the mouse pucks back to themselves.

To test for enrichments of region annotations across disease state, region composition was aggregated to sample-level (for Slide-seq to 4-way split pucks), CLR-transformed, and enrichment was calculated using Welch's t-test. This was done for region annotation on human and mouse scRNA-seq data, and on the original and mapped region annotation on the mouse Slide-seq data.

Annotating human Slide-seq data

Human Slide-seq data were annotated with cell type clusters using the annotation ("cl295v11SubFull") from the human single cell reference (Pelka et al., 2021). Cluster level annotations were mapped to the human pucks using TACCOs "annotate" function with OT as core method, basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 4 iterations of bisectioning with a divisor of 3, and 10-fold sub-clustering of the annotations. This fine grained cluster-level annotation was aggregated to cell type level ("clTopLevel" in Ref. (Pelka et al., 2021)).

To annotate human Slide-seq data with mouse region information, human genes were mapped to mouse orthologs using MGI homology information [subsection "Mapping of mouse and human orthologs"], and mouse region annotations were mapped to human pucks using TACCOs "annotate" function with OT as core method, basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 4 iterations of

bisectioning with a divisor of 3. Cell type and region information were aggregated to a pseudo bulk cell type and region composition per sample.

Cell-type associations across samples

To compare associations of cell types across samples in human and mouse scRNA-seq the "clMidwayPr" cell type annotation in the human data (Pelka et al., 2021) was aggregated to the same level as mouse cell type annotation, and then aggregated per sample and CLR-transformed. Pearson correlation coefficients were calculated for every cell type pair for different subsets of samples: all samples, normal samples, dysplastic, and for human MMRd/MMRp samples.

Epithelial program associations across human samples

To determine epithelial program associations across human samples, TACCO's "annotate" function was used to annotate human epithelial scRNA-seq (after mapping to mouse orthologs using MGI homology information [subsection "Mapping of mouse and human orthologs"]) with mouse epithelial programs from mouse scRNA-seq data using OT as core method, basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 4 iterations of bisectioning with a divisor of 3. The remaining steps were performed as for cell-type association (subsection "Cell-type associations across samples").

Comparing cell-type spatial organization in human and mouse

To compare the spatial co-occurrence of cell types in human and mouse, human Slide-seq data was mapped to mouse orthologs using MGI homology information [subsection "Mapping of mouse and human orthologs"]. Cell subset (cluster) annotation was mapped from mouse scRNA-

seq to human Slide-seq using TACCO's "annotate" method with OT as core method, basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 4 iterations of bisectioning with a divisor of 3. Cluster annotation was aggregated to cell type and compartment levels. To reduce bias from muscularis beads, the comparison was restricted to beads deep in the "epithelial domain", identified by using TACCO's "annotation_coordinate" function (with max_distance=100, delta_distance=10, critical_neighbourhood_size=4.0) to compute the effective distance from stromal annotation and including only beads with a distance of at least 75µm. TACCO's "co_occurrence" function was applied to the selected beads (with delta_distance=20, max_distance=1000) to determine cell type co-occurence as a function of distance.

Comparing spatial organization of epithelial programs in human and mouse

To compare the spatial organization of epithelial programs in human and mouse, epithelial programs were defined based on the epithelial component of beads. To extract the epithelial component, human Slide-seq data was annotated with human cell cluster (cell subset) annotations from human scRNA-seq (Pelka et al., 2021) using TACCO's "annotate" function with OT as core method, basic platform normalization, entropy regularization parameter epsilon 0.005, marginal relaxation parameter lambda of 0.1, and 4 iterations of bisectioning with a divisor of 3. Using this annotation, beads were split into contributions of every cluster using TACCO's "split_observations" function, and these split contributions were aggregated up to compartment level. Epithelial contributions were mapped to mouse orthologs using MGI homology information [subsection "Mapping of mouse and human orthologs"]. Epithelial program annotations were mapped from mouse scRNA-seq (from dysplastic samples only) to the epithelial fraction of each

human Slide-seq bead (using compartment-level aggregated split beads, including from every bead those UMIs that were assigned to any of the epithelial clusters), using TACCO's "annotate" method with OT as core method, basic platform normalization, entropy regularization parameter epsilon 0.005 and marginal relaxation parameter lambda of 0.001. These data and the mouse data were subsetted to the same "epithelial domain" beads as in the cell-type case [subsection "Comparing cell-type spatial organization in human and mouse"] and TACCO's "co_occurrence" function was used (with delta_distance=20, max_distance=1000) to determine the co-occurrence of epithelial programs as a function of their distance.

Scoring epithelial mouse regions in mouse and human epithelial pseudo-bulk data

The published processed and filtered count matrices were used (where available) or instead raw count matrices for single cell/nucleus RNA seq data from Pelka (Pelka et al., 2021), Chen (Chen et al., 2021), Khaliq (Khaliq et al., 2022b), Becker (Becker et al., 2022), Zheng (Zheng et al., 2022) (excluding 'blood' samples), Che (Che et al., 2021) (only 'CRC' and 'LM' samples) and Joanito (Joanito et al., 2022) (excluding the 'LymphNode' sample).

To subset the human single cell data to epithelial cells, the epithelial annotation was used where readily available (Chen et al., 2021; Pelka et al., 2021). For the remaining datasets (Becker et al., 2022; Che et al., 2021; Joanito et al., 2022; Khaliq et al., 2022b; Zheng et al., 2022), TACCOs tc.tl.annotate function was used with default parameters to transfer the 'cl295v11SubShort' annotation from Pelka (Pelka et al., 2021), from which a compositional compartment annotation was constructed, and then a cell was assigned to the epithelial compartment if it had more than 95% epithelial fraction.

To correct for batch effects between the different data sources, first batches were defined by species times protokoll: 'mouse-10x3p', 'mouse-SlideSeq', 'human-10x3p' (Che et al., 2021; Joanito et al., 2022; Pelka et al., 2021; Zheng et al., 2022), 'human-10x5p' (Joanito et al., 2022; Khaliq et al., 2022b), 'human-inDrop' (Chen et al., 2021), and 'human-snRNA' (Becker et al., 2022). Then TACCO's "tc.pp.normalize_platform" function was used to determine per gene batch normalization factors using only the normal samples of one data source per batch (choosing the normal samples from Zheng for 'human-10x3p' and the normal 5' samples from Joanito for 'human-10x5p'). The resulting factors are then used to rescale the sample-by-gene count matrices for the full dataset per batch, i.e. including non-normal samples. The normalization factors are calculated with respect to an (arbitrarily chosen) normal reference dataset (Zheng et al., 2022).

The epithelial mouse region score was defined as the mean of the clr-transformed expression values in the pseudo-bulk expression profile of the epithelial part of a dataset using the top 200 differentially expressed genes between all regions by Fisher's exact test.

To account for species-specific biases (in-set *vs*. out-of-set prediction: the DEGs are calculated in mouse), the scores per region across samples were zero-centered and scaled to unit variance across all samples (including normal and non-normal samples and all batches) per species. A Principal Components Analysis (PCA) of the region scores across all species, batches and samples was conducted and the values for the first PC were compared between different conditions using Mann-Whitney U test with Benjamini-Hochberg FDR.

Assessing the relation between mouse regions and clinical endpoints in human bulk RNA-seq

Published RNA-seq data from the COAD and READ cohorts of TCGA PanCancerAtlas were used. Mouse region scores were defined as the mean of the log1p-transformed, zero-centered and scaled expression values in the bulk expression profile using the top 200 differentially expressed genes between the malignant mouse regions (6, 8, 11) by Fisher's Exact test (comparing each of the three regions to the other two). Scores were stratified into quartiles. PFI was compared between patients with tumors whose scores were in the lowest and highest quartiles using the Logrank test as implemented in the lifelines package, followed by Benjamini-Hochberg FDR.

Compositional enrichment analyses

Enrichments on compositional data (cell type compositions, etc.) were evaluated with a one-sided Welch's t test on sample level using CLR-transformed compositions followed by Benjamini-Hochberg FDR. For the enrichment of tdTomato, counts and ALR-transformation were used instead with all non-tdTomato counts used as reference compartment. Enrichment analyses were performed using TACCO's "enrichments" function.

GO term enrichment analyses

GO term enrichment analyses are performed using TACCO's setup_goa_analysis and run_goa_analysis methods, which use GOATOOLS (Klopfenstein et al., 2018) and a Scipy's two-sided Fishers Exact test stats.fisher_exact internally.

Pearson correlation

For all analyses, Pearson correlation coefficients were calculated using TACCO's utils.cdist which internally used Scipy's spatial.distance.cdist function.

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. (2020). Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. N. Engl. J. Med. *383*, 2207–2218. .

Angelo, M., Bendall, S.C., Finck, R., Hale, M.B., Hitzman, C., Borowsky, A.D., Levenson, R.M., Lowe, J.B., Liu, S.D., Zhao, S., et al. (2014). Multiplexed ion beam imaging of human breast tumors. Nat. Med. *20*, 436–442. .

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. .

Arenberg, D.A., Keane, M.P., DiGiovine, B., Kunkel, S.L., Strom, S.R., Burdick, M.D., Iannettoni, M.D., and Strieter, R.M. (2000). Macrophage infiltration in human non-small-cell lung cancer: the role of CC chemokines. Cancer Immunol. Immunother. *49*, 63–70. .

Basu, S., Cheriyamundath, S., Gavert, N., Brabletz, T., Haase, G., and Ben-Ze'ev, A. (2019). Increased expression of cathepsin D is required for L1-mediated colon cancer progression. Oncotarget *10*, 5217–5228. .

Becker, W.R., Nevins, S.A., Chen, D.C., Chiu, R., Horning, A.M., Guha, T.K., Laquindanum, R., Mills, M., Chaib, H., Ladabaum, U., et al. (2022). Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. Nat. Genet. 1–11. .

Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414. .

Biton, M., Haber, A.L., Rogel, N., Burgin, G., Beyaz, S., Schnell, A., Ashenberg, O., Su, C.-W., Smillie, C., Shekhar, K., et al. (2018). T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation. Cell *175*, 1307-1320.e22. .

Burden, R.E., Gormley, J.A., Jaquin, T.J., Small, D.M., Quinn, D.J., Hegarty, S.M., Ward, C., Walker, B., Johnston, J.A., Olwill, S.A., et al. (2009). Antibody-mediated inhibition of cathepsin S blocks colorectal tumor invasion and angiogenesis. Clin. Cancer Res. *15*, 6042–6051. .

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

Che, L.-H., Liu, J.-W., Huo, J.-P., Luo, R., Xu, R.-M., He, C., Li, Y.-Q., Zhou, A.-J., Huang, P., Chen, Y.-Y., et al. (2021). A single-cell atlas of liver metastases of colorectal cancer reveals reprogramming of the tumor microenvironment in response to preoperative chemotherapy. Cell Discovery *7*, 1–21. .

Chen, B., Scurrah, C.R., McKinley, E.T., Simmons, A.J., Ramirez-Solano, M.A., Zhu, X., Markham, N.O., Heiser, C.N., Vega, P.N., Rolong, A., et al. (2021). Differential pre-malignant

programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. Cell *184*, 6262-6280.e26. .

Clevers, H. (2006). Wnt/beta-catenin signaling in development and disease. Cell *127*, 469–480. .

DiMeo, T.A., Anderson, K., Phadke, P., Fan, C., Perou, C.M., Naber, S., and Kuperwasser, C. (2009). A novel lung metastasis signature links Wnt signaling with cancer cell self-renewal and epithelial-mesenchymal transition in basal-like breast cancer. Cancer Res. *69*, 5364–5373. .

Drev, D., Harpain, F., Beer, A., Stift, A., Gruber, E.S., Klimpfinger, M., Thalhammer, S., Reti, A., Kenner, L., Bergmann, M., et al. (2019). Impact of Fibroblast-Derived SPARC on Invasiveness of Colorectal Cancer Cells. Cancers *11*. https://doi.org/10.3390/cancers11101421.

Fearon, E.R. (2011). Molecular genetics of colorectal cancer. Annu. Rev. Pathol. *6*, 479–507. .

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759–767. .

Flanagan, D.J., Pentinmikko, N., Luopajärvi, K., Willis, N.J., Gilroy, K., Raven, A.P., Mcgarry, L., Englund, J.I., Webb, A.T., Scharaw, S., et al. (2021). NOTUM from Apc-mutant cells biases clonal competition to initiate cancer. Nature *594*, 430–435. .

Fleming, M., Ravula, S., Tatishchev, S.F., and Wang, H.L. (2012). Colorectal carcinoma: Pathologic aspects. J. Gastrointest. Oncol. *3*, 153–173. .

Folkman, J. (2002). Role of angiogenesis in tumor growth and metastasis. Semin. Oncol. *29*, 15–18. .

Giesen, C., Wang, H.A.O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P.J., Grolimund, D., Buhmann, J.M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat. Methods *11*, 417–422. .

Golby, S.J.C., Chinyama, C., and Spencer, J. (2002). Proliferation of T-cell subsets that contact tumour cells in colorectal cancer. Clin. Exp. Immunol. *127*, 85–91. .

Golovko, D., Kedrin, D., Yilmaz, Ö.H., and Roper, J. (2015). Colorectal cancer models for novel drug discovery. Expert Opin. Drug Discov. *10*, 1217–1229. .

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. Cell *174*, 968-981.e15. .

Grünwald, B.T., Devisme, A., Andrieux, G., Vyas, F., Aliar, K., McCloskey, C.W., Macklin, A., Jang, G.H., Denroche, R., Romero, J.M., et al. (2021). Spatially confined sub-tumor microenvironments in pancreatic cancer. Cell *184*, 5577-5592.e18. .

Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., et al. (2017). A single-cell survey of the small intestinal epithelium. Nature *551*, 333–339. .

Hagerling, C., Gonzalez, H., Salari, K., Wang, C.-Y., Lin, C., Robles, I., van Gogh, M., Dejmek, A., Jirström, K., and Werb, Z. (2019). Immune effector monocyte–neutrophil cooperation induced by the primary tumor prevents metastatic progression of breast cancer. Proc. Natl. Acad. Sci. U. S. A. *116*, 21704–21714. .

Hara, T., Chanoch-Myers, R., Mathewson, N.D., Myskiw, C., Atta, L., Bussema, L., Eichhorn, S.W., Greenwald, A.C., Kinker, G.S., Rodman, C., et al. (2021). Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma. Cancer Cell *39*, 779-792.e11. .

Hunter, M.V., Moncada, R., Weiss, J.M., Yanai, I., and White, R.M. (2021). Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. Nat. Commun. *12*, 1–16. .

Jackson, H.W., Fischer, J.R., Zanotelli, V.R.T., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., et al. (2020). The single-cell pathology landscape of breast cancer. Nature *578*, 615–620. .

Jerby-Arnon, L., Shah, P., Cuoco, M.S., Rodman, C., Su, M.-J., Melms, J.C., Leeson, R., Kanodia, A., Mei, S., Lin, J.-R., et al. (2018). A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell *175*, 984-997.e24. .

Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C.L.P., Macalinao, D.C., Kahraman, M., Srinivasan, H., et al. (2022). Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nat. Genet. *54*, 963–975. .

Johnson, L., Mercer, K., Greenbaum, D., Bronson, R.T., Crowley, D., Tuveson, D.A., and Jacks, T. (2001). Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. Nature *410*, 1111–1116. .

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S.-R., Kurian, A., Van Valen, D., West, R., et al. (2018). A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. Cell *174*, 1373-1387.e19. .

Khaliq, A.M., Erdogan, C., Kurt, Z., Turgut, S.S., Grunvald, M.W., Rand, T., Khare, S., Borgia, J.A., Hayden, D.M., Pappas, S.G., et al. (2022a). Correction: Refining colorectal cancer classification and clinical stratification through a single-cell atlas. Genome Biol. *23*, 156. .

Khaliq, A.M., Erdogan, C., Kurt, Z., Turgut, S.S., Grunvald, M.W., Rand, T., Khare, S., Borgia, J.A., Hayden, D.M., Pappas, S.G., et al. (2022b). Refining colorectal cancer classification and clinical stratification through a single-cell atlas. Genome Biol. *23*, 113. .

Kitamura, T., Kometani, K., Hashida, H., Matsunaga, A., Miyoshi, H., Hosogi, H., Aoki, M., Oshima, M., Hattori, M., Takabayashi, A., et al. (2007). SMAD4-deficient intestinal tumors recruit CCR1+ myeloid cells that promote invasion. Nat. Genet. *39*, 467–475. .

Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., et al. (2018). GOATOOLS: A Python library for Gene Ontology analyses. Sci. Rep. *8*, 10872. .

Kuraguchi, M., Wang, X.-P., Bronson, R.T., Rothenberg, R., Ohene-Baah, N.Y., Lund, J.J., Kucherlapati, M., Maas, R.L., and Kucherlapati, R. (2006). Adenomatous polyposis coli (APC) is required for normal development of skin and thymus. PLoS Genet. *2*, e146. .

Kwong, L.N., and Dove, W.F. (2009). APC and its modifiers in colon cancer. Adv. Exp. Med. Biol. *656*, 85–106. .

Lasry, A., Zinger, A., and Ben-Neriah, Y. (2016). Inflammatory networks underlying colorectal cancer. Nat. Immunol. *17*, 230–240. .

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. .

Li, D., Peng, Z., Tang, H., Wei, P., Kong, X., Yan, D., Huang, F., Li, Q., Le, X., Li, Q., et al. (2011). KLF4-mediated negative regulation of IFITM3 expression plays a critical role in colon cancer pathogenesis. Clin. Cancer Res. *17*, 3558–3568. .

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400-416.e11. .

Lu, M.-H., Huang, C.-C., Pan, M.-R., Chen, H.-H., and Hung, W.-C. (2012). Prospero homeobox 1 promotes epithelial-mesenchymal transition in colon cancer cells by inhibiting E-cadherin via miR-9. Clin. Cancer Res. *18*, 6416–6425. .

Madisen, L., Zwingman, T.A., Sunkin, S.M., Oh, S.W., Zariwala, H.A., Gu, H., Ng, L.L., Palmiter, R.D., Hawrylycz, M.J., Jones, A.R., et al. (2009). A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. Nat. Neurosci. *13*, 133–140. .

Mages, S., Moriel, N., Avraham-Davidi, I., Murray, E., Chen, F., Rozenblatt-Rosen, O., Klughammer, J., Regev, A., Nitzan, N. (2022). TACCO: Unified annotation transfer and decomposition of cell identities for single-cell and spatial omics. Unpunished.

Marino, S., Vooijs, M., van Der Gulden, H., Jonkers, J., and Berns, A. (2000). Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. Genes Dev. *14*, 994–1004. .

Marjanovic, N.D., Hofree, M., Chan, J.E., Canner, D., Wu, K., Trakala, M., Hartmann, G.G., Smith, O.C., Kim, J.Y., Evans, K.V., et al. (2020). Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. Cancer Cell *38*, 229-246.e13. .

el Marjou, F., Janssen, K.-P., Chang, B.H.-J., Li, M., Hindie, V., Chan, L., Louvard, D., Chambon, P., Metzger, D., and Robine, S. (2004). Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. Genesis *39*, 186–193. .

Marx, V. (2021). Publisher Correction: Method of the Year: spatially resolved transcriptomics. Nat. Methods *18*, 219. .

McAllister, S.S., and Weinberg, R.A. (2014). The tumour-induced systemic environment as a critical regulator of cancer progression and metastasis. Nat. Cell Biol. *16*, 717–727. .

Mendez, M.G., Kojima, S.-I., and Goldman, R.D. (2010). Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition. FASEB J. *24*, 1838–1851. .

van Neerven, S.M., de Groot, N.E., Nijman, L.E., Scicluna, B.P., van Driel, M.S., Lecca, M.C., Warmerdam, D.O., Kakkar, V., Moreno, L.F., Vieira Braga, F.A., et al. (2021). Apc-mutant cells act as supercompetitors in intestinal tumour initiation. Nature *594*, 436–441. .

Oliemuller, E., Newman, R., Tsang, S.M., Foo, S., Muirhead, G., Noor, F., Haider, S., Aurrekoetxea-Rodríguez, I., Vivanco, M. dM, and Howard, B.A. (2020). SOX11 promotes epithelial/mesenchymal hybrid state and alters tropism of invasive breast cancer cells. Elife *9*, e58374. .

Palla, G., Fischer, D.S., Regev, A., and Theis, F.J. (2022). Spatial components of molecular tissue biology. Nat. Biotechnol. https://doi.org/10.1038/s41587-021-01182-1.

Pelka, K., Hofree, M., Chen, J.H., Sarkizova, S., Pirl, J.D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W.H., Xu, K.H., et al. (2021). Spatially organized multicellular immune hubs in human colorectal cancer. Cell *184*, 4734-4752.e20. .

Penninger, J.M., and Crabtree, G.R. (1999). The Actin Cytoskeleton and Lymphocyte Activation. Cell *96*, 9–12. .

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. .

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467. .

Roper, J., Tammela, T., Cetinbas, N.M., Akkad, A., Roghanian, A., Rickelt, S., Almeqdadi, M., Wu, K., Oberli, M.A., Sánchez-Rivera, F.J., et al. (2017). In vivo genome editing and organoid transplantation models of colorectal cancer and metastasis. Nat. Biotechnol. *35*, 569–576. .

Roper, J., Tammela, T., Akkad, A., Almeqdadi, M., Santos, S.B., Jacks, T., and Yilmaz, Ö.H. (2018). Colonoscopy-based colorectal cancer modeling in mice with CRISPR–Cas9 genome editing and organoid transplantation. Nat. Protoc. *13*, 217–234. .

Sasaki, N., Sachs, N., Wiebrands, K., Ellenbroek, S.I.J., Fumagalli, A., Lyubimova, A., Begthel, H., van den Born, M., van Es, J.H., Karthaus, W.R., et al. (2016). Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon. Proc. Natl. Acad. Sci. U. S. A. *113*, E5399-407. .

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. .

Schwab, R.H.M., Amin, N., Flanagan, D.J., Johanson, T.M., Phesse, T.J., and Vincan, E. (2018). Wnt is necessary for mesenchymal to epithelial transition in colorectal cancer cells. Dev. Dyn. *247*, 521–530. .

Smigiel, J.M., Parameswaran, N., and Jackson, M.W. (2017). Potent EMT and CSC phenotypes are induced by oncostatin-M in pancreatic cancer. Mol. Cancer Res. *15*, 478–488. .

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. .

Smith, P.M., and Garrett, W.S. (2011). The gut microbiota and mucosal T cells. Front. Microbiol. *2*, 111. .

Spencer, J., and Sollid, L.M. (2016). The human intestinal B-cell response. Mucosal Immunol. *9*, 1113–1124. .

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82. .

Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol. *39*, 313–319. .

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., 3rd, Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*, 224. .

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type

annotation, Writing group, Supplemental text writing group, and Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372. .

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196. .

Wagner, J., Rapsomaniki, M.A., Chevrier, S., Anzeneder, T., Langwieder, C., Dykgers, A., Rees, M., Ramaswamy, A., Muenst, S., Soysal, S.D., et al. (2019). A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. Cell *177*, 1330-1345.e18. .

Waylen, L.N., Nim, H.T., Martelotto, L.G., and Ramialison, M. (2020). From whole-mount to single-cell spatial assessment of gene expression in 3D. Commun Biol *3*, 602. .

Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell *177*, 1873-1887.e17. .

West, N.R., Murray, J.I., and Watson, P.H. (2013). Oncostatin-M promotes phenotypic changes associated with mesenchymal and stem cell-like differentiation in breast cancer. Oncogene *33*, 1485–1494. .

West, N.R., Hegazy, A.N., Owens, B.M.J., Bullers, S.J., Linggi, B., Buonocore, S., Coccia, M., Görtz, D., This, S., Stockenhuber, K., et al. (2017). Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor-neutralizing therapy in patients with inflammatory bowel disease. Nat. Med. *23*, 579–589. .

Winkler, J., Abisoye-Ogunniyan, A., Metcalf, K.J., and Werb, Z. (2020). Concepts of extracellular matrix remodelling in tumour progression and metastasis. Nat. Commun. *11*, 1–19. .

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. .

Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst *8*, 281-291.e9. .

Zeng, Z.-S., Shu, W.-P., Cohen, A.M., and Guillem, J.G. (2002). Matrix metalloproteinase-7 expression in colorectal cancer liver metastases: evidence for involvement of MMP-7 activation in human cancer metastases. Clin. Cancer Res. *8*, 144–148. .

Zheng, X., Song, J., Yu, C., Zhou, Z., Liu, X., Yu, J., Xu, G., Yang, J., He, X., Bai, X., et al. (2022). Single-cell transcriptomic profiling unravels the adenoma-initiation role of protein tyrosine kinases during colorectal tumorigenesis. Signal Transduction and Targeted Therapy *7*, 1–14. .

Ziyad, S., and Iruela-Arispe, M.L. (2011). Molecular mechanisms of tumor angiogenesis. Genes Cancer *2*, 1085–1096. .