bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multiplexed 3D atlas of state transitions and immune interactions in colorectal cancer

Jia-Ren Lin[1,2*], Shu Wang[1,2,3*], Shannon Coy[1,2,4*], Madison Tyler[1,2], Clarence Yapp[1,2], Yu-An Chen[1,2], Cody N. Heiser[1,5], Ken S. Lau,[1,6] Sandro Santagata[1,2,4] and Peter K. Sorger[1,2,7,#]

[1]Human Tumor Atlas Network

[2]Ludwig Center at Harvard and Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, 02115, USA.

[3]Harvard Graduate Program in Biophysics, Harvard University, Cambridge, MA, 02138, USA.

[4]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

[5]Program in Chemical & Physical Biology, Vanderbilt University School of Medicine, Nashville, TN, 37232, USA.

[6]Epithelial Biology Center and Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN, 37232, USA.

[7]Department of Systems Biology, Harvard Medical School, Boston, MA, 02115, USA.

[*]equal contribution

[#] lead contact at peter_sorger@hms.harvard.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SUMMARY

Advanced solid cancers are complex assemblies of tumor, immune, and stromal cells that invade adjacent tissue and spread to distant sites. Here we use highly multiplexed tissue imaging, spatial statistics, and machine learning to identify cell types and states underlying morphological features of known diagnostic and prognostic significance in colorectal cancer. We find that a thorough spatial analysis requires imaging the entire tumor region, not small fields of view (e.g. those found in tissue microarrays). When this condition is met, the data reveal frequent transitions between histological archetypes (tumor grades and morphologies) correlated with molecular gradients. At the tumor invasive margin, where tumor, normal, and immune cells compete, localized features in 2D such as tumor buds and mucin pools are seen in 3D to be large connected structures having continuously varying molecular properties. Immunosuppressive cell-cell interactions also exhibit graded variation in type and frequency. Thus, whereas scRNA-Seq emphasizes discrete changes in tumor state, whole-specimen imaging reveals the presence of large- and small-scale spatial gradients analogous to those in developing tissues.

INTRODUCTION

Solid tumors are complex ecosystems in which the proliferative and invasive properties of cancer cells are controlled by molecular signals from the stroma, surveillance from immune cells, and physical barriers imposed by tissue architecture. Tumor atlases being constructed by the NCI Human Tumor Atlas Network (HTAN) aim to map the interactions of tumor cells with each other and with other cells in the tumor microenvironment based on the identification and quantification of cell types and intracellular states within a preserved 3D environment. A priority is understanding the mechanisms that regulate critical molecular and phenotypic transitions such as evolution of high-grade invasive cancer states and suppression of immune surveillance (Rozenblatt-Rosen et al., 2020a).

Histopathology plays an essential role in the diagnosis, staging, and clinical management of cancer using hematoxylin and eosin (H&E) stained, formalin-fixed paraffin-embedded (FFPE) tissue sections, complemented by immunohistochemistry (IHC). Pathology studies spanning more than a century have determined which morphological features of tumors are of diagnostic or prognostic significance (Amin et al., 2017), but classical histology provides insufficient information to precisely identify cell subtypes, characterize oncogenic mechanisms, and identify potentially druggable vulnerabilities. Multiplexed tissue imaging builds on 75 years of experience with IHC (Coons et al., 1942) and yields subcellular resolution data on the abundance of 20-60 antigens, which is sufficient to identify immune cell types, assay proliferation, measure oncogene expression and deeply phenotype

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

clinical specimens (Angelo et al., 2014; Gerdes et al., 2013; Giesen et al., 2014; Goltsev et al., 2018; Lin et al., 2018; Saka et al., 2019; Schürch et al., 2020; Wagner et al., 2019). When multiplexed images are segmented and quantified, the resulting single cell data are a natural complement to scRNA-seq data. Unlike dissociated scRNA-seq, however, multiplexed tissue imaging preserves detailed spatial information and reveals how cells interact with one another and with tissue structures such as basement membranes and connective tissue.

The multiplexed imaging in this study involves whole slide imaging in which data are collected from the entirety of a sample using tissue-based cyclic immunofluorescence (t-CyCIF) (Lin et al., 2018), a method that is broadly similar to MxIF (Gerdes et al., 2013), CODEX (Goltsev et al., 2018), and Immuno-SABER (Saka et al., 2019). CyCIF assembles high-plex data from multiple rounds of lower-plex fluorescence imaging, resulting in images of large fields of view at high resolution. This contrasts with mass spectrometry-based methods such as Imaging Mass Cytometry (IMC) (Giesen et al., 2014) and Multiplexed Ion Beam Imaging (MIBI) (Angelo et al., 2014) which collect data on all channels at once but typically have relatively small fields of view.

Analysis of multiplexed tissue images can be approached '*bottom-up*' by enumerating cell types and identifying cell-cell interactions using spatial statistics and neighborhood analysis. With a sufficient number of patients in a bottom-up analysis, it is potentially possible to relate such data to cancer outcomes, as measured by Kaplan-Meier estimators. A "*top-down*" approach starts with the identification in H&E data of archetypical histopathologic features (histotypes) of known importance in disease management followed by use of machine learning and other methods to discover underlying molecular patterns in multiplexed data. Bottom-up analysis predominates in multiplexed tissue imaging, perhaps due to the applicability of tools developed for mass cytometry (Bendall et al., 2011) and scRNA-seq. In contrast, digital pathology primarily involves top-down approaches that have the very substantial advantage of leveraging a century of research on tumor morphologies and states known to associate with clinical outcome (Amin et al., 2017). However, pathologists have had to discount many false leads involving recurrent spatial patterns that cannot definitively be shown to have prognostic or diagnostic value. It remains difficult even today to determine which morphological and molecular features are consequential and which are incidental. In this paper we analyze colorectal cancer (CRC) using both approaches and compare the resulting insights.

Classification of CRC increasingly incorporates molecular and genetic information on microsatellite stability (MSI), mismatch repair status (MMR), PD-L1 expression, and other features (Kopetz et al., 2015). Nonetheless, histopathology dominates diagnosis and clinical management. Key

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

histological features include the degree of tumor differentiation (well, moderate, poor), grade (low, high), subtype (mucinous, signet ring cell, etc.), presence of perineural, lymphatic, or vascular invasion, presence of peri-colorectal tumor deposits, the density of tumor infiltrating lymphocytes (TILs), the presence of tumor budding, etc. (Weiser, 2018). Tumor differentiation is assessed with respect to normal epithelial structures and considers changes in single cell morphology (e.g., cell shape, cytoplasmic staining, nuclear size, etc.) and in cell neighborhoods (e.g., loss of glandular organization, hypercellularity, etc.) (Fleming et al., 2012). In many cases the molecular basis of these morphological changes is not fully understood although de-differentiation, including "stemness" (Aponte and Caicedo, 2017), epithelial-mesenchymal transition (EMT) (Kalluri and Weinberg, 2009), changes in nuclear mechanics (Uhler and Shivashankar, 2018), and similar processes, are clearly involved.

Invasive margins are among the most important regions of solid cancers for understanding tumor progression, immune surveillance, and prognosis (Cianchi et al., 1997; Schürch et al., 2020). The deep invasive fronts of CRC and other solid tumors often exhibit single or small clusters of tumor cells surrounded by stromal cells. These cells appear to sprout or "bud" from the tumor mass (Lugli et al., 2017a). Tumor budding is defined by the International Tumor Budding Consensus Conference (ITBCC) (Lugli et al., 2017b) as clusters of ≤4 tumor cells along the invasive front. In CRC, tumor budding cells express cytokeratins specific to normal epithelial cells, but they have lower levels of cell-to-cell adhesion proteins (e.g. E-cadherin, CD44, and Ep-CAM) (Gosens et al., 2007a), elevated expression of EMT markers, and a reduced proliferative index (Rubio, 2007). Because tumor buds appear to be genetically indistinguishable from the parent tumor (Centeno et al., 2017), budding is thought to involve a cell state (epigenetic) change involving interaction with adjacent stromal and immune cells. Tumor budding cells may give rise directly to circulating tumor cells and metastases (Zlobec and Lugli, 2010) and are correlated with poor outcomes in CRC, including increased risk of local recurrence, metastasis, and cancer-related death (Rogers et al., 2016).

In this paper, we address fundamental questions about the construction and analysis of tissue atlases via 3D image-based reconstruction of human colorectal cancer (CRC). We confirm cell type assignments with scRNA-seq and bridge the divide between classical histopathology, with its demonstrated ability to identify disease-relevant spatial patterns, and computational analysis of multiplexed tissue images with spatial statistics. Using supervised machine learning, we show that CRC histotypes can be highly intermixed and that morphological transitions are often coincident with molecular gradients. Tumor budding appears to be a graded phenotype with budding cells, as classically defined, forming one extreme of a gradual molecular and morphological transition in which fibrils

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

branch off from the tumor mass and become increasingly mesenchymal. These phenotypic and molecular transitions are only detected when large contiguous regions of tumor are analyzed in 2D and 3D, thereby providing the context and statistical power needed to chart cell intrinsic and population-based tumor phenomena.

RESULTS

Overview of the specimens and data

Multiplexed CyCIF and H&E imaging were performed on 94 FFPE CRC human specimens, including a wide variety of histologic and molecular subtypes (**Table S1**). The most extensive analysis, involving multi-section 3D imaging and scRNA-seq, was performed on a specimen also available in frozen form (specimen CRC1, **Figure 1A**; **Table S1** and **S2**). Sixteen additional FFPE specimens (CRC2-17) were subjected to single-section whole slide imaging and were also used to construct a tissue microarray (TMA) with the remaining 77 specimens (0.6 mm diameter cores; four cores per patient; TMA) (**Figure 1A**). The resected tumor and adjacent normal tissue in CRC1 is from the cecum of a 69-year old male; pathology revealed a poorly differentiated stage IIIB adenocarcinoma (pT3N1bM0) (Weiser, 2018) with microsatellite instability (MSI-H) and a $BRAF^{V600E}$ (c.1799T>A) mutation. The tumor had an extended front invading into underlying muscle (*muscularis propria*) and connective tissue giving rise to a 'budding margin' (IM-A) adjacent to normal colon epithelium, a 'mucinous margin' in the middle of the specimen (IM-B), and a deep 'pushing margin' (IM-C) (denoted "A", "B" and "C" in **Figure 1B**).

Although it is possible to perform H&E staining after fluorescence imaging, the quality of the images is substandard; H&E and CyCIF were therefore performed on adjacent serial sections. A total of 106 serial sections were cut from an FFPE block of CRC1 (~3cm^2) (**Table S2**), and 25 sections were stained with 24-plex CyCIF using lineage specific antibodies against epithelial, immune, and stromal cell populations, as well as markers of cell cycle state, signaling pathway activity, and immune checkpoint expression (**Figure 1C, Figure S1A; Table S3**). Three additional sections (CRC1, sections 045-047) were subjected to deeper analysis with a total of 102 cell lineage and state markers (**Table S4**). In total, fluorescence and transmitted light microscopy involved acquisition of 361 successive megapixel image tiles followed by stitching and flat-fielding to generate whole-slide images comprising approximately 7 x 10^8 pixels (2.5 terabytes of data).

Using the MCMICRO software pipeline (Schapiro et al., 2021), images were segmented, fluorescence intensities quantified on a per-cell basis, and cell types assigned using lineage-specific

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

marker expression (**Figure 1C; Table S5**); a total of ~5 x 10^7 cells were successfully identified in a tissue volume of 1.2 x $10^{11}\mu m^3$. t-SNE demonstrated clear separation of keratin$^+$ epithelial cells, both normal and transformed, from CD31$^+$ endothelial cells (primarily blood vessels), desmin$^+$ stromal cells and CD45$^+$ immune cells (**Figure 1D; Table S6**). Immune cells could be further divided into biologically important classes such as CD8$^+$PD1$^+$ cytotoxic T cells, CD4$^+$ helper T cells, CD20$^+$ B cells, CD68$^+$ and/or CD163$^+$ macrophages, as well as discrete sub-categories such as CD4$^+$FOXP3$^+$ T regulatory cells (**Table S5**). When scRNA-seq was performed on ~10^4 cells from an adjacent region of CRC1, a high degree of concordance was observed between the cell type abundances estimated from RNA expression and from image data (r^2 = 0.938; **Figure 1E; Figures S1B–D**), indicating that image segmentation and quantification of staining intensities was performed accurately. These images are available for interactive on-line viewing in a web browser (without data or software download) using MINERVA software (Hoffer et al., 2020; Rashid et al., 2020).

Morphological and molecular gradients involving tumor phenotypes

To study the molecular features underlying archetypical histologic patterns and anatomic levels of invasion in H&E data, two board-certified pathologists annotated regions of interest (ROI) which were then subjected to computational analysis using CyCIF data from adjacent sections. Annotated ROIs included normal colonic mucosa (ROI1); three regions of moderately differentiated invasive adenocarcinoma with glandular morphology involving either the luminal tumor surface (ROI2), submucosa (ROI3) or muscle (*muscularis propria)* at the deep invasive margin (ROI4); regions of poorly differentiated (high-grade) adenocarcinoma with solid or signet ring cell architecture (ROI5); and regions of invasive adenocarcinoma with prominent extracellular mucin pools (ROI6) (**Figure 1B)**. A region with prominent tumor budding (TB) near IM-A was also annotated. Note that ROIs 2-4 have similar local morphologies but represent different levels of anatomic invasion that correlate with standard TNM (tumor, node, metastasis) staging (Weiser, 2018); TNM staging has well established value in prognosis and disease management. The location of an extended ROI corresponding to the invasive margin was also determined and is described in greater detail below.

With respect to composition, cell type calling on CyCIF data showed that solid adenocarcinoma (RO15) had the highest proportion of epithelial (tumor) cells, 1.5% of which expressed PD-L1 (**Table S6**). The normal epithelium (ROI1) had the fewest epithelial cells (~40%) and the greatest proportion of stroma and immune cells. The cells in ROIs 2-4 were similar in having a glandular histomorphology characteristic of moderately differentiated adenocarcinoma. Stromal reaction (extracellular matrix and

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

growth factor production in response to tumor) (Ohtani, 1998) was greater in ROI2. However, we could identify no single molecular marker unique to one ROI or specific morphology. We therefore used machine learning to look for multidimensional features predictive of different histotypes.

A k-nearest neighbors (kNN) classifier was trained on CyCIF marker expression in keratin$^+$ normal epithelial and tumor cells (a different classifier was trained and validated for each section to assess the reproducibility of the approach). For each ROI, half of the cells were used for training and half for validation; cells from ROIs 2-4 were combined because their local morphologies were similar. The kNN classifier returned high confidence predictions (Shannon entropy near zero) within and outside of the annotated ROIs for all four morphologies and across all tissue sections: pathology review confirmed that the kNN classifier accurately predicted tumor domains with the expected histotypes and morphologic patterns (**Figures 2A**,**B** and **Figure S2**). Of note, CyCIF data comprised the centroid of each cell and its integrated marker intensity, but no other morphological or environmental information. This demonstrates that disease-relevant morphology in H&E images had been encoded as hyperdimensional patterns in marker space.

Unexpectedly, the kNN classifier scored most regions of the tumor as comprising a mixture of morphological classes (as quantified by the posterior probability; **Figure 2C**) and transitions from one class to another on varying spatial scales. In many regions, Shannon entropy values approached two, demonstrating an equal mixture of all classes (red in **Figure 2B**). This was not a limitation of the markers used for classification, because similar results were obtained with combinations of ~100 antibodies used to stain sections 044 to 047 of CRC1 (**Figures S2A–S2C; Table S4**). When tumor regions with high entropy values were examined in H&E, we found that they corresponded to transitions between classical morphologies (**Figure 2D, E**). These transitions occurred on spatial scales ranging from a few cell diameters (~50 μm) to the whole image (~1 cm) and included transitions from mucinous to glandular, mucinous to solid, and glandular to solid. These transitions were not limited to a single part of the tumor, but were observed multiple times in spatially separated areas, with some contiguous tumor regions appearing to alternate between states (**Figure S2D, E**).

While histotypes could not be distinguished from each other using single CyCIF markers (kNN classification was based on hyperdimensional features) morphological gradients coincided in some cases with local molecular gradients. The existence of chemical and physical gradients in normal tissues is well established (Oudin and Weaver, 2016), but gradients are less well explored in human tumors, perhaps because tumor genetics tends to focus on discrete differences. Three examples of morphologic and molecular gradients from CRC1 are shown in in **Figures 2D** and **2E:** a normal-glandular transition

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coinciding with an E-cadherin to PCNA gradient, a mucinous-solid transition coinciding with a Cytokeratin 20 to Cytokeratin 18 gradient, and a glandular-solid transition coinciding with a transition in epigenomic regulators/modifications (KAP1 and H3K27me3). We conclude that histological archetypes commonly used for pathology grading and clinical planning represent only a minority of the CRC1 tumor and that transitional regions are common. Transitions appear to correspond to gradients in the expression of molecular markers, suggesting the existence of continuous rather than discrete differences in cancer cell biology across a tumor.

Tumor budding and molecular transitions at the deep invasive front

The deep invasive front of a tumor, where cancer cells invade and displace normal tissue, is most clearly defined by morphology and position relative to normal tissue. This area is thought to contain the most aggressive subpopulation of tumor cells able to migrate into normal tissue and evade immune surveillance. In solid tumors, an EMT-like transition at the invasive front gives rise to small clusters of "de-differentiated" budding cells probably involved in metastatic dissemination (Zlobec and Lugli, 2010). To study the molecular properties of tumor buds we annotated regions having large numbers of buds based on established histological criteria and then developed a computational approach to quantify molecular markers.

For diagnostic purposes, tumor buds are defined as clusters of ≤4 tumor cells along the invasive front (Lugli et al., 2017b), and less commonly the non-marginal 'internal' stroma of the tumor mass (Lugli et al., 2011). To identify such cells computationally, we used Delaunay triangulation (Goltsev et al., 2018). This method generates links between immediately adjacent cells; intervening cell free-regions and structures were allowed, but cells farther than 20 μm apart were not linked (**Figure 3A**). In parallel, a pathologist annotated buds near the deep invasive front based on H&E and CyCIF images from CRC1 using ITBCC morphologic criteria, which yielded ~4 x 10^3 budding cells (~0.01% of all tumor cells) across all 25 CyCIF sections (~20 to 400 budding cells per section; **Figure S3**). Buds were largely confined to one ~2.0 x 0.7 x 0.4 mm region of the invasive front proximal to normal colonic epithelium (denoted 'A' in **Figure 1B**).

When Delaunay triangulation was performed for all cells that stained with keratin in CyCIF data (i.e., both tumor and normal epithelium) the smallest clusters contained 1-2 contiguous cells and the largest ~3 x 10^4 cells (**Figures 3A** and **3B**). Large Delaunay clusters mapped to regions of poorly differentiated adenocarcinoma with solid architecture, consistent with the fact that these regions are almost entirely made up of tumor cells, whereas moderately differentiated regions with glandular

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

architecture contained a wide range of cluster sizes. Delaunay clusters meeting the ITBCC definition for buds (≤4 tumor cells) totaled ~10^5 cells across all sections, including those near the IM-A margin, in agreement with pathologist-annotated buds.

CyCIF data showed that E-cadherin and Na-K ATPase levels were sharply lower in both annotated and Delaunay budding cells as compared to adjacent tumor masses and that the proliferative markers PCNA and Ki-67 were also low (**Figure 3C, D**); loss of cell-to-cell adhesion and a non-proliferative phenotype are established features of budding cells in CRC and other tumor types and has been described as a partial EMT phenotype (Gosens et al., 2007b; Rubio, 2008). Across the whole of CRC1, we found that E-cadherin, Na-K ATPase, PCNA and Ki-67 expression varied continuously with Delaunay cluster-size (**Figure 3D,E**), with the lowest value associated with the smallest clusters (2-4) cells. The presence of continuous variation in gene expression and proliferative index with Delaunay cluster-size was reproduced in 16 whole slide images (CRC2-17) from other CRC patients (**Figure S3E**).

Examples of these graded transitions are shown in **Figures 3F** and **3G,** in which cohesive glandular tumor, which involves larger cluster-sizes and a $PCNA^{high}$ state, fragments into fibrillar structures involving smaller clusters and a $PCNA^{low}$ state and ultimately what appear to be bud-like structures having the lowest marker expression and surrounded by stroma (**Figure 3F**) or mucinous spaces (**Figure 3G**). Analogous transitions of tumor masses into small Delaunay clusters occurred throughout the tumor both at the invasive front and along the luminal surface of the tumor in regions corresponding histologically to discohesive growth with focal signet ring cell morphology (so-called because of the resemblance to a signet ring). Signet-ring cells are found as single cells or in loose clusters having reduced cell–cell adhesion and are aggressive with respect to invasion and metastasis (Sung et al., 2008). The small Delaunay clusters found in mucin pools along the invasive margin at IM-B are not distinguishable in size or expression (of cohesive and proliferation markers) from buds as classically defined, but the ITBCC definition encompasses only clusters in fibrous stroma. These data suggest that buds as classically defined represent an extreme of a gradual, morphological transition occurring across much of the tumor.

A 3D reconstruction of CRC1 serial sections (**Figure 4A**) showed that buds meeting ITBCC criteria were frequently components of fibrillar structures emanating from the main tumor mass (**Figure 4A-C**). When visualized using a virtual section of the annotated budding region viewed along the X- and Z-directions (**Figure 4C**) we found that annotated buds (white) were connected to larger tumor structures (red) along Z. This is analogous to fibrillar structures shown **Figures 3F,G**. We conclude that

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

budding cells are distal cross-sections of fibrillar structures, consistent with a previous 3D H&E study of budding (Bronsert et al., 2014); fibrils of similar sizes involving E-cadherin^{low}, Na-K ATPase^{low}, PCNA^{low} and Ki-67^{low} cells are also found surrounded by mucin along the IM-B region of the tumor margin. Thus, EMT at the invasive front is not characterized by isolated cell spheres, as first described by Weinberg and colleagues for cultured cells (Mani et al., 2008), but rather large interconnected structure in which gradual breakdown in cell adhesion leads to ever narrower fibrils.

Many small clusters of cells having an EMT-like phenotype identified by Delaunay triangulation were found within or immediately adjacent to mucin pools. Mucins are large glycoproteins that protect the gastrointestinal epithelium from mechanical, chemical, and microbial damage; they are made by goblet cells but also by some colorectal tumor cells (Bresalier, 2002). Mucins are found in the lumen of the normal colon and also in fissures and isolated pools of many different sizes surrounded by tumor; in CRC1 this is most prominent in the central region involving invasive margin IM-B (**Figure 1B**). When we reconstructed mucin-containing domains in 3D (**Figure 4D,E**), we found that 2D pools were interconnected and continuous with other pools and with the colonic lumen. Fibrillar structures projected into this complex space and were continuous with glandular and mucinous tumor. Thus, this region of the invasive front likely consists of an inter-connected mucin-filled structure invaded by fibular structures comprising tumor cells undergoing partial EMT that are themselves connected to an extended and irregular invasive margin (**Figures 4A-C**). The occurrence of bud-like transitions in distinct microenvironments suggests that invasion is a property of an ensemble of cells rather than small, isolated, clusters of tumor cells. Moreover, entities that appear in 2D to be isolated, such as mucin pools, are in fact cross-sectional views of larger inter-connected structures.

Characterization of cell neighborhoods using spatial statistics

As a complementary approach to working from H&E histotypes to molecular features we used "*bottom-up*" analysis based on segmenting individual cells, identifying cell types and patterns of protein expression and then applying spatial statistics (Jackson et al., 2020; Schürch et al., 2020). Spatial correlation (Genton and Kleiber, 2015; Sethna, 2006; Sokal and Thomson, 1987) quantifies similarity in cell properties, expression of a marker protein for example, as a function of physical distance (Moffitt et al., 2016; Svensson et al., 2018). When we analyzed three exemplary proteins expressed in three different cell types, keratin (epithelial cells), α-SMA (stromal cells) and FOXP3 (immune cells), spatial correlation exhibited a good fit to a single exponential function (staining levels were binarized and co-occurrence was scored as a function of absolute distance); the parameters of these exponentials are

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

length scale and *strength* (the relative spatial contiguity of a cell type on a scale of 0 to 1). Keratin, α-SMA and FOXP3 positivity length scales varied from 100-400 μm and strengths from 0.1 to 0.8 (**Figure 5A)**, properties that were readily apparent in underlying images (**Figure 5B)**. When 16 additional CRC specimens were analyzed using 24-plex whole-slide CyCIF (samples CRC2-17), the length scales for different markers and biologically relevant combinations ranged from r =80 μm for the platelet–endothelial cell adhesion molecule CD31, which corresponded predominantly to small vessels (**Figures 5C-5E** and **Figure S4A**) found throughout the stroma and tumor with enrichment in muscle. The longest length scale feature for a single marker was r = 450 μm for the B-lymphocyte antigen CD20, corresponding to tertiary lymphoid structures found in the peri-tumoral stroma, muscle, and sub-serosal soft tissue. $CD163^+$ macrophages had an intermediate spatial scale (r = 250 μm) and clustered in narrow strips of tumor associated stroma with exclusion from the nearby $keratin^+$ tumor compartment (**Figure 5F** and **5G**). Thus simple spatial statistics uncover spatial motifs that repeat throughout the specimen.

Impact of spatial correlation on statistical power

The impact of spatial inhomogeneity and correlation on statistical power has not been studied extensively in histology as compared to ecology and related fields (Stroup, 2002). However, length scale has a direct impact on the minimum dimension of a sample needed to measure a feature accurately. This is a concern because length scales for many markers in our CRC patients were similar to those commonly encountered in multiplexed imaging: 300 to 500 μm radii for TMAs and ~1 mm^2 for fields of view in MIBI and IMC imaging. To explore the issue of statistical power in image data, we created a "virtual TMA" (vTMA) in which 1 mm diameter field of views (FOVs) were subsampled from an image of CRC1 (section 097). We then computed cell abundances (cell count normalized to the total number of cells) in each virtual core, which contained ~10^3 cells on average as compared to ~5 x 10^5 for an entire CRC1 section. Sampling was performed so that the vTMA would primarily contain keratin-positive cells and therefore derive from the tumor or normal epithelia (**Figure 5H** shows four virtual cores). We observed very poor precision when the abundance of $keratin^+$ cells was estimated from vTMA cores: it varied 20-fold from 5% to 95% (the true value based on counting all cells in section 097 was 45%; **Figure 5I**). Abundance estimates for α-SMA and FOXP3 positivity in vTMA cores were also imprecise but to a lesser extent than for keratin positivity (**Figure 5I**). In contrast, when we repeatedly drew samples of ~10^3 cells from the dataset without regard to position (the actual number in each random sample was matched to the number in each vTMA core), the estimated abundance of $keratin^+$ cells was

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

45 ± ~1%, representing good precision. Thus, imprecision was not a consequence of having too few cells.

These findings can be explained in full based on the concept of an *effective* sample size (N_{eff}), the number of independent samples that can be obtained from correlated measurements (Lavrakas, 2008). We found that the reduction in sample size between a truly random set of cells and a vTMA core (N/N_{eff}) was 10- to 1000-fold depending on the marker (median value ~100; **Figure S4B**) meaning that a core having ~10^3 cells might constitute as few as 1 to 3 independent samples, which inevitably causes high variance. Across all markers, the effective sample size (N_{eff}) closely matched the predictions of the Central Limit Theorem (Ibragimov, 1962) assuming exponential spatial correlations, as demonstrated by a coefficient of determination of $R^2 = 0.97$ for data and theory (**Figure S4B**, regression in green, theory in red). N_{eff} can also be approximated using measurable properties including the spatial correlation strength c_0, length scale l, and average cell size l_{cell} with a simple scaling law (see Methods):

$$\frac{N}{N_{eff}} \sim c_0\left(\frac{l}{l_{cell}}\right)^2 . \qquad (1)$$

This formula yielded estimates of ~10 for FOXP3, ~100 for α-SMA and ~500 for keratin (**Figure S4C**), a reasonable match to observed values that also explains why variance in estimating keratin positivity (with a length scale of 400 μm) is substantially greater than that for α-SMA or FOXP3 positivity (100-200 μm length scales). Taken together, these findings demonstrate that spatial correlation is responsible for reducing effective sample sizes in images of vTMA cores and other similarly small FOVs by ~10-1000 fold relative to independently chosen cells; small effective sample size causes high variance.

Higher-order spatial features, such as the properties of cell neighborhoods (e.g. proximity analysis; **Figure 5J)**, were even more strongly confounded by spatial correlation and sampling error than cell abundance. For example, when we quantified how often keratin$^+$ (tumor) cells were found in the neighborhood of α-SMA$^+$ (stromal) cells as a function of distance, vTMA data were much less precise than random sampling (compare light blue and green in **Figure 6A**; note that distance is plotted as the number of neighboring cells, which is proportional to the square of the distance). The same was true when we searched for neighborhoods containing CD45$^+$ immune cells and CD31$^+$ endothelial cells: spatial cross-correlation was positive in whole-slide data (or random subsamples but equally likely to be positive or negative depending on which vTMA core was analyzed. Inspection of t-CyCIF images showed that these differences arose from very different tissue morphologies and spatial arrangements rather than a statistical quirk (illustrated by four selected cores; **Figures 6B** and **6C)**. We also performed neighborhood analysis based on co-occurrence (correlation > 0) or co-avoidance (correlation < 0) for

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

neighborhoods of different sizes. For example, for a neighborhood of 10 cells (~50 μm in diameter, a common parameter choice) analysis of vTMAs data could not confidently determine the sign of the correlation even though analysis of the whole specimen (or a random subset similar in size to a vTMA) is highly accurate. These findings can all be understood as arising from the uncertainty inherent in attempting to quantify spatial phenomena whose length scales are similar in size or larger than the sampling dimension. Errors may be sufficient in magnitude to lead to false associations with patient survival, based on Kaplan-Meir estimators, for example **(Figure S5).**

Distinguishing biological variation and sampling error

To determine how sampling error compares with patient-to-patient variability we computed cell abundances for single markers and marker combinations in samples CRC2 to 17 (e.g., $CD68^+$ $PDL1^+$ macrophages) and found that values varied by 3 to 10-fold from patient to patient based on whole-slide data (**Figure 6D**, blue). Since these data come from patients with different clinical outcomes, we assumed measured variance to have a biological basis. We then examined data derived from an actual TMA constructed with four 0.6 mm diameter cores from the same tissue blocks and patients (and imaged in duplicate). Inter-core variance was found to be much greater than true patient-to-patient differences and examining four cores as oppose to one did not enable detection of patient to patient variability (**Figure 6D**, green). A similar level of variance was also observed in data from 77 tumors available only as TMA (four cores each; imaged in duplicate). Only one univariate measurement made from TMAs, Ki-67 positivity in $keratin^+$ cells (which in this case were primarily tumor cells), exhibited inter-patient variability (18-61%) greater than sampling error (~30%) (**Figures 6D** and **S5A**). The existence of mutual information between markers should make it possible to better distinguish biological variation from measurement errors. As a proof of concept, we identified a linear combination of markers measurable from a single core with a signal-to-noise ratio comparable to Ki-67 positivity (**Figure S5B;** Ki-67 positivity was excluded, see methods). The best estimator primarily included markers of immune cells but was not immediately interpretable. From these data we conclude that the measurement of basic features of tissues such as cell abundance, spatial co-occurrence and neighborhood exclusion are subject to high sampling error introduced when small fields of view are imaged; this sampling error is greater in many cases than true patient to patient variability.

To what extent are 2D whole-slide images adequate representations of 3D specimens? To study this, we computed cell abundances and spatial correlations in 24 sections from CRC1 and compared this to patient-to-patient variability across specimens CRC2 to 17 (compare red and blue in **Figure S5C** and

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S5D). For all but a few markers we found that variance between Z-sections was smaller than patient-to-patient variability. Some variance with section is expected since the tumor differs in extent and shape from the top of the specimen to the bottom. To determine if the spatial features discussed in this paper project sufficiently in X, Y to be sampled accurately, we computed cell type abundances for the pathologist annotated ROIs used in **Figure 1B** for kNN modeling. In this case we also observed low level of section to section variance (**Figure 6E).** We conclude that 2D whole slide imaging of a 3D specimen does not, in general, suffer from the same problem as subsampling a section with TMAs or small fields of view (see Discussion).

Immune profiling shows substantial spatial variation and an immunosuppressive niche associated with the invasive margin

The morphology of the invasive margin of solid tumors has high prognostic value and differences between irregular invasive and well-delineated pushing margins are used for patient management (Koelzer and Lugli, 2014). In cancers, such as CRC1, that are mismatch repair deficient (dMMR) a high rate of mutation generates abundant neoantigens and promotes a microenvironment rich in tumor-directed effector T-cells and macrophages but also cells expressing immune checkpoint proteins. Classical histopathology and the 3D reconstructions described above show that the invasive margin of CRC1 involves three distinguishable regions. To characterize tumor-host interactions across the margin with sufficient spatial power for statistical significance, we performed whole-tumor spatial analysis of cell abundance and neighborhood relationships (**Figure 7**). We concentrate below on specimen CRC1, since this had the highest level of annotation by a pathologist, but computational analysis of CRC2-17 showed that CRC1 was not an outlier with respect to cellular composition, grade or rate of progression (**Figure S7**). However, immuno-regulatory cell types clearly different in abundance,

In CRC1 we found that CD45-positive immune cells (**Table S5,** cell types 7-21) were broadly distributed throughout the tumor, normal mucosa and stroma although $CD20^+$ B cells were concentrated in tertiary lymphoid structure (among all $CD45^+$ cells: 9% were Tregs and 24% T effector cells; **Figure 7A**). Cells (of all types) expressing PD-L1 and PD-1 were also broadly distributed but PD-$L1^+$ tumor cells were rare (**Figures 7B** and **7C**). Cell states potentially involved in immunosuppressive were more highly localized; for example, $CD8^+$ $PD1^+$ T effector cells were enriched 2.3-fold in the stroma whereas $CD68^+$ PD-$L1^+$ macrophages were 3-fold enriched in the tumor (where they comprised 1.4% of all cells). Immunosuppressive cell types were also non-uniformly distributed along the tumor margin. The tumor budding region (IM-A) was highly enriched for $CD4^+$/$FOXP3^+$ T regulatory cells (**Figure 7D**)

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and the greatest concentration of $CD8^+$ $PD\text{-}1^+$ cytotoxic T cells was adjacent to (but not within) the budding region (as traditionally defined) (**Figure 7E**). IM-A also contained the majority of PD-L1 expressing tumor cells detected (**Figure 7E**) as well as $CD68^+$/$PD\text{-}L1^+$ macrophages (**Figure 7F**). In contrast, the invasive mucinous margin along IM-B exhibited the lowest level of immune cell infiltration suggesting low levels of immune surveillance in this region and consistent with a function for some mucins in immunosuppression (Bhatia et al., 2019). The pushing border along IM-C was largely devoid of T regulatory cells and PD-L1 expressing tumor cells but was enriched in $CD68^+$/$PD\text{-}L1^+$ macrophages (**Figure 7F**).

To determine whether immune and tumor cells are in a position to interact with each other, we performed adjacency analysis, setting a strict 20 μm cutoff for a positive interaction (**Figure 7D-F**). The strongest interaction between $CD4^+$/$FOXP3^+$ T regulatory cells and $CD8^+$ effector T cells was in the immediate vicinity of budding cells at the IM-A margin. The frequency of this interaction was 2.5-fold greater than in normal tissue (ROI1) and also significantly greater than in all other annotated ROIs. Moreover, Tregs were present in excess to effector T cells, making efficient suppression via cell-cell contact feasible. Tumor cells in the budding region expressed high levels of PD-L1 (relative to the tumor as a whole) and the frequency of interaction with $CD8^+$ T $PD\text{-}1^+$ cells was significantly greater here than elsewhere in the tumor (**Figure 7E**). $CD8^+$ $PD\text{-}1^+$ T cells were also more frequently associated $CD68^+$ $PD\text{-}L1^+$ macrophages than elsewhere in CRC1, and the macrophages were in 2-3 fold excess (**Figure 7F**). IM-C (near ROIs 2-4) was a second region with significant enrichment for interaction of $CD8^+$ $PD\text{-}1^+$ T effector and an excess of $CD68^+$ $PD\text{-}L1^+$ macrophages. In contrast, no significant concentration of immunosuppressive interactions was observed in the mucin-rich regions of IM-B (**Figure 7A-C**). Thus, three of the most prevalent mechanisms of immunosuppression involving T effector cells were found in CRC1 including interaction with T regulatory cells, $PD\text{-}L1^+$ tumor cells, and $PD\text{-}L1^+$ macrophages. Immunosuppressive mechanisms varied in a graded manner across the invasive front with the greatest inhibition of effector cells at the area of active budding, where metastasis may originate (**Figure 7H**).

DISCUSSION

Most efforts to generate tumor atlases have focused on systematic enumeration of cell types and states using scRNA-seq, and initial multiplexed imaging studies have also started to adopt this "*bottom-up*" paradigm (HuBMAP Consortium, 2019; Rozenblatt-Rosen et al., 2020b) by relying principally on spatial statistics and neighborhood analysis as routes to biological discovery. In contrast, the current paper combines "*bottom up*" analysis with "*top-down"* histology-driven identification of large-scale

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phenotypic patterns in the tumor and the tumor microenvironment. This approach exploits over a century of intensive investigation of tumors by anatomic pathologists and is made possible by a 3D reconstruction of a colorectal cancer comprising ~5 x 10^7 cells in volume of ~1.2 x 10^{11} μm^3 complemented by whole-slide 2D imaging of 16 additional CRC specimens and TMAs having 77 additional tumors.

Whereas dissociative scRNA-seq clusters cell types into discrete states, we find that molecular and morphologic phenotypes are often graded, with transitions between states spanning spatial scales from a few cell diameters to several mm. These transitions are apparent throughout the tumor but are most notable along the invasive margin, where tumor cells actively invade and compete with normal cells while being constrained by immune surveillance. 3D reconstruction of the CRC1 tumor margin reveals multiple neighboring zones distinguished by both morphology and immune environment. In region IM-A, tumor cells having poorly differentiated morphology invade the underlying stroma and classical budding is observed. This is a consequence of reduced cell-cell cohesion and is associated with metastasis and poor prognosis (Guil-Luna et al., 2020). Tumor buds are classically defined ≤4 cells in a cluster surrounded by stroma (Lugli et al., 2017a), but 3D reconstruction shows that these represent the termini of extended fibrillary structures whose distal tips generate the bud morphology in cross-section (Bronsert et al., 2014). Structurally similar fibrils are found along the IM-B region of the invasive front, which is rich in mucinous tumor morphology, with fibrils projecting into mucin pools rather than into stroma. In 2D sections, mucin pools appear to be fully surrounded by stromal cells, but 3D analysis shows that they actually form a large and elaborate interconnected network. Multiplexed imaging shows that the long fibrillary networks projecting into connective tissue or mucin exhibit graded transitions in molecular state with progressively lower expression of cell adhesion and proliferation markers along the proximal-to-distal axis (from larger to smaller clusters in cross section) with the most pronounced EMT-like states at the very tip. Thus, the leading edge of the CRC1 invasive front involves structures that are much larger than they appear in 2D sections and within these structures, cells undergo graded transitions in phenotype. In some instances, these changes in state recur in a semi-periodic manner, analogous to the recurrence of villi in normal epithelia. Acellular features such as mucin pools also appear to be interconnected with each other and with the lumen of the colon and its microbiome.

Multiple immunosuppressive interactions have been described in colorectal cancer involving ingress of Tregs and expression of checkpoint proteins on both tumor and immune cells. Neighborhood analysis shows high enrichment for Tregs and T effector cells along IM-A, where budding is most pronounced (**Figure 1B**). These tumor cells express elevated levels of PD-L1 and $CD8^+$ $PD1^+$ T cells are

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

found in close proximity (20 μm), providing an additional mechanism of effector cell suppression. Suppression is further enhanced by proximity to PD-L1^{+} macrophages. If we include a ~500 μm distant tertiary lymphoid structure rich in B cells (which are implicated in anti-tumor immunity in CRC) (Edin et al., 2019) the margin along IM-A is both the most immune rich and most immunosuppressive environment in CRC1. Tumor budding is less commonly found in microsatellite unstable (MSI-H) tumors such as CRC1, potentially due to their destruction by abundant cytotoxic T cells and macrophages (Zlobec et al., 2011). The presence of multiple axes of immunosuppression may explain the frequency of buds in CRC1 and the evolution of a poorly differentiated high-grade phenotype that correlates with aggressive clinical outcomes. .

Few immune cells were found in association with the IM-B region, perhaps due to an abundance of mucin. In contrast, T cells and macrophages along with an immunosuppressive environment were enriched at IM-C; this comprises a well-demarcated 'pushing border' between *muscularis propria* and moderately differentiated tumor. In this case, however, PD-L1^{+} macrophages were the immune cells most frequently in contact with CD8^{+} PD1^{+} T cells; in contrast, PD-L1 expression on tumor cells was low and few Tregs were present. Thus, the established connection between tumor budding, metastasis and disease recurrence is likely to reflect the presence of tumor cells undergoing EMT (and developing the ability to disseminate) as well as a permissive and immunosuppressive environment (Guil-Luna et al., 2020).

Some of the morphological gradients we have identified in images were correlated with molecular gradients detectable by CyCIF, but few of these are likely to be causal associations. Instead, it is much more likely that diffusible molecules (morphogens) and cell-cell contacts similar to those controlling tissue development are involved (Rogers and Schier, 2011). Gradients in growth factors, metabolites and the ECM are readily detected in cultured cells, by intravital imaging in the mouse (Kondo et al., 2021), and by mass spectrometry in frozen human tissue (Randall et al., 2020) and we find that they can be encoded in hyperdimensional features in marker space. However, for none of ~100 proteins examined were we able to identify a unique association between a single marker and an established histomorphology. High-dimensional encodings identified by machine learning were nonetheless fully consistent with current practice in grading CRC, and transitions between learned classes corresponded to transitions between morphological archetypes such as glandular, mucinous and solid tumor. Continuity of states implies that intratumoral heterogeneity (ITH) may have a non-genetic origin as suggested by Sharma et al., 2019 and Black and McGranahan, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The characteristic organization of tissues and tumors arises from the presence of recurrent spatial patterns ranging from individual tumor-immune interactions to organ-scale features; these are visible by eye and quantifiable using spatial statistics. In colorectal cancer we find that the length scale for co-expression of proteins such as cell-type markers is as large as ~500 μm (e.g., tertiary lymphoid structures) and complex morphological gradients are even larger. A corollary is that the effective sample size in images is reduced 10^3-fold compared to dissociative method such as scRNA-seq; the penalty is more severe for complex properties such as neighborhood inclusion and exclusion. Thus, analysis of TMAs and small (sub-millimeter) fields of view is subject to measurement error that is greater than the patient-to-patient variability that we seek to explain; such error can reduce the power of outcome analysis and generate spurious correlations with Kaplan-Meier survival estimators. Imaging entire slides in 2D largely overcomes this problem. Whole-slide analysis is standard pathology practice (Ghaznavi et al., 2013) and the FDA regards it as a diagnostic necessity (Aeffner et al., 2019; Health, 2019) because review of sufficiently large regions of tumor (generally >1 cm^2) is essential to account for heterogeneity, assess degree of invasion and make accurate classifications.

Our findings have implications for the use of TMAs and small specimens recovered via core biopsies (~1-2 mm in diameter) and fine needle aspiration (~300 μm in diameter) for diagnosis. Additional study is required on this topic because accuracy and precision are not only properties of the specimen and measurement method but are also conditioned by the question being asked. Detecting relatively infrequent events, such as tumor budding, is different from quantifying a recurrent clonal event, such as IDH1/2 mutation in glioma (Verhaak et al., 2010) or ALK/ROS expression in lung cancer (Rogers et al., 2015). Our findings argue for the importance of whole-slide 2D data, and in some cases 3D image stacks, for accurate spatial analysis in a research setting. This does not come without cost because processing and analyzing multi-TB datasets comprising 10^6 to 10^7 cells remains a substantial technical challenge.

Conclusions

Although we aspire to identify biologically significant states and neighborhood relationships directly from multiplexed spatial data, 120 years of surgical pathology suggests that this will not be straightforward. A minority of recurrent tumor morphologies associate significantly with disease outcome. Moreover, random subsets of other types of high dimensional data (e.g., gene expression) are well known to associate significantly with disease outcome (Venet et al., 2011). There is therefore merit in working “top down” from established histopathological features to molecular and cellular programs;

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

this is true even when the goal is scientific discovery rather than diagnosis. The current paper represents a first step in this direction and deeper analysis of additional specimens using multiplexed data and machine learning is now warranted. For example, a whole-slide analysis of T cell infiltration in CRC (Galon et al., 2006), which is the basis of an established clinical test (Immunoscore) (Angell et al., 2020) paired with outcome analysis will likely lead to new biological insight and possibly also improved diagnostics.

Limitations of this study

One limitation of the current study is that only one CRC has been reconstructed in 3D. While many spatial features can be accurately discerned from a single 2D section (which we have collected from 17 patients), determining the connectivity of tumor buds and mucin pools in CRC2-17 awaits additional data. Spatial power analysis demonstrates that it is not possible to reliably associate features with outcomes using TMA data, making the 93 patients and 372 distinct cores in our TMA dataset useful for other purposes only; we judge the 16 whole-slide samples to be insufficient and are currently expanding the cohort. However, the primary limitation is computational and not data acquisition per se; processing multi-TB whole-slide images remains time consuming. By releasing all images described in this study at full resolution we hope to attract experts in computation to advance the state of the art. We have recently described the cloud-based MCMICRO pipeline (Schapiro et al., 2021), which is based on the Nextflow language and Docker/Singularity (O'Connor et al., 2017) containers, to promote such research. Our ability to effectively visualize and process 3D tissue data has emerged as another limitation in the current work. We use a simple graph-theoretic approach to find cell clusters similar to buds but other approaches from graph or percolation theory (Plotkin et al., 2002; Reynolds et al., 2009) would probably better capture the interdigitated morphology of fibrils, and both classical (Rohban et al., 2017) and deep learning approaches (Djuric et al., 2017) may better capture key histological features than kNN models.

DATA AVAILABILITY AND ATLAS IMAGE VIEWING

As part of this paper all images at full resolution, all derived image data (e.g. segmentation masks) and all cell count tables will be publicly released via the NCI-recognized repository for Human Tumor Atlas Network (HTAN; https://humantumoratlas.org/) at Sage Synapse. A version of this data is available at https://www.synapse.org/#!Synapse:syn18434611/wiki/597418. ***Note to reviewers***: this data resource is undergoing final review for Private Health Information and requires a (free) Synapse

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

account; public access will be provided as soon as possible. An anonymous "reviewer only" link can be provided prior that by requesting it from the monitoring editor.

Several of the figure panels in this paper are available with text and audio narration for anonymous on-line browsing using MINERVA software (Rashid et al., 2020), which supports zoom, pan and selection actions without requiring the installation of software. A Minerva story with an overview of CRC1 (section 096 & 097) can be found at: cycif.org/crc1-intro and the 25 CRC1 Z-sections can be found at: cycif.org/crc1-3d.

ACKNOWLEDGEMENTS

We thank Alyce Chen, Raquel Arias-Camison, Zoltan Maliga and Jeremy Muhlich and for help in all stages of this project. This work was supported by NIH grants U54-CA225088 (PKS, SS), U2C-CA233262 (PKS, SS), U54-CA225088 (PKS, SS), U2C- CA233291 (CNH, KSL), R01-DK103831 (CNH, KSL), NIH training grant T32-GM007748 (SC)) and the Ludwig Center at Harvard (PKS, SS). All HTAN consortium members are named at (humantumoratlas.org). We thank Dana-Farber/Harvard Cancer Center for the use of the Specialized Histopathology Core, which provided histopathology services supported by P30-CA06516.

AUTHOR CONTRIBUTIONS

JRL, PKS and SS developed the concept for the study. JRL, CY and YC collected image data and also performed analysis in collaboration with SC and SW. CH and KL collected and analyzed scRNA-Seq data. MT developed the MINERVA stories. All authors wrote and edited the manuscript. SS, PKS and KL provided supervision and funding.

DECLARATION OF INTERESTS

PKS is a member of the SAB or BOD member of Applied Biomath, RareCyte Inc., and Glencoe Software, which distributes a commercial version of the OMERO database; PKS is also a member of the NanoString SAB. In the last five years the Sorger lab has received research funding from Novartis and Merck. Sorger declares that none of these relationships have influenced the content of this manuscript. SS is a consultant for RareCyte Inc.

FIGURE TITLES AND LEGENDS

Figure 1. Overview of the data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(A) A formalin-fixed and paraffin-embedded (FFPE) CRC sample (CRC1) was serially-sectioned for H&E staining and CyCIF imaging. Twenty-five sections were stained with the main CyCIF panel described in table S3. Three sections (045, 046, & 047) were stained with three additional panels, detailed in table S4. In addition to CRC1, two duplicate sections of a tissue microarray (TMA) and 16 of the TMA parental whole slide sections (CRC2-17) were stained with the main CyCIF panel. **(B)** Histopathologic annotation of H&E images into three invasive margins (A, B, C) and 6 different ROIs. ROI histology: *1. normal mucosa, 2. superficial (luminal) adenocarcinoma, 3. submucosal adenocarcinoma, 4. muscularis propria adenocarcinoma (deep invasive front), 5. solid adenocarcinoma, 6. mucinous adenocarcinoma.* ROIs 2-4 exhibit a moderately differentiated appearance with glandular architecture, while regions 5-6 exhibit a poorly differentiated appearance with predominantly solid or cribriform architecture. Regions of tumor budding were also noted for subsequent analysis. **(C)** An example of a CyCIF whole-slide image (section 097) and cell type assignment. Thirty-one different cell types from three main categories (tumor, stroma, and immune) were defined and their locations mapped within the tumor section. (**D**) Dimensionality reduction of single-cell data by t-SNE, color coded by staining intensity for the indicated marker. Cell type plot (far right) uses color code show in Figure S1D. **(E)** Comparison of cell-type percentages assessed via single cell RNA-sequencing (scRNA-seq) and CyCIF.

Figure 2. Correlation and prediction of morphologic and molecular tumor phenotypes.
(A) Four examples of tumor morphology in training ROI and non-adjacent regions predicted with high-confidence to have corresponding morphology. kNN classifiers were trained and validated separately for each section to evaluate the reproducibility of the models. **(B)** Prediction confidence for assignment of kNN classes as measured by Shannon entropy (a value of 0 corresponds no classification uncertainty, and a value of 2 to random assignment and equal mixing of all classes). **(C)** Posterior probability that each keratin^{+} cells would belong to normal epithelium or glandular, solid, or mucinous tumor classes. Annotation reflects classifier gradients and alternation corresponding to morphologic phenotype. **(D)** Sample tumor region transitioning from normal morphology (H&E top) with local e-cadherin expression to glandular adenocarcinoma architecture and expression, with increased PCNA (CyCIF, bottom). Contours describe averaged local epithelial cell expression of PCNA. **(E)** Additional examples of transition regions, analogous to (D), with H&E (top), CyCIF (middle), and quantified expression contours (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3. Tumor budding is a distributed phenomenon associated with graded molecular and morphologic transitions

(A) Clusters of keratin^{+} cells (orange) in a local field of view (FOV) of section 097. Neighboring keratin^{+} cells are denoted by edges, along with keratin^{-} cells (blue) and annotated buds (white) for reference. Histogram shows cluster-sizes (log2) across all CRC1 sections. **(B)** Cluster sizes mapped onto CRC1 section 097. Image exaggerates size of single cells for visibility. (**C**) Keratin^{+} cell expression visualized by tSNE performed using 10 tumor-relevant CyCIF markers (two of which are shown; see Methods) with color representing marker intensity. Cluster-size also shown (top panel), with black outline denoting small clusters (including annotated buds) and color representing log2 cluster size. (**D**) Differential expression of markers in cells annotated as tumor buds. The relative expression of indicated markers is represented in heatmap as the log2 ratio of budding tumor cells to all tumor cells. **(E)** Log-intensity of markers and their dependency on cluster-size in CRC1 tumor cells. Expression of annotated buds shown in green for reference. Boxes show 1st-3rd quartiles; points beyond are not shown. Each box represents ~10^5-10^6 cells. **(F)** Example of transition from main tumor mass into fibrils and bud-like cells in the stroma in CyCIF (*left*) and H&E (*right*). There is gradual loss of Na-K ATPase and PCNA from the main tumor mass to the tips of fibrils with decreasing cluster-size (budding cells appear red on CyCIF, arrowheads). Image is oversaturated to make hues more visible. **(F)** Analogous budding structures in mucinous tumor regions, with fibrils and budding cells (arrowheads) extending into mucin pools rather than stroma.

Figure 4. 3D reconstruction of tumor architecture reveals large-scale organization of tumor buds and mucinous regions

(A) 3D overview of CRC1 at at region A of the invasive margin A (see Figure 1B). (*Top)* Surface renderings of keratin (blue), α-SMA (purple), normal mucosal region (green), CD68^{+}PD-L1^{+} cells (yellow), and budding cells (red). *(Bottom)* All annotated buds colored by density and showing interconnected fibril-like networks of budding cells. **(B)** Three-dimensional visualization of annotated buds (pink, middle) or ≤4 cell clusters identified by Delaunay triangulation (pink, bottom) relative to connected tumor masses and other cells with uncertain connectivity (green). Corresponding regions in 2D CyCIF images are shown above with keratin staining fully saturated to make cluster size more apparent. **(C)** Virtual XZ-section of annotated budding region generated from image stack; the locations of annotated tumor buds are marked in white and found to be connected via continuous staining intensity to the main tumor mass. **(D)** Connectivity of mucin pools across serial sections. Largest contiguous

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mucin pool structure (red) extends into the lumen (outline in yellow). Image is mirrored along Z relative to (A) to better visualize details. To the right three exemplary H&E FOVs from different regions of the reconstructed mucin structure with mucin pools that appear isolated in 2D sections (arrowheads).

Figure 5. Spatial heterogeneity and estimation errors for regional sampling.

(A) Spatial correlations of binarized staining intensities for keratin$^+$ (red), α-SMA$^+$ (blue), and FOXP3$^+$ (green) cells, along with their exponential fits (dashed curves). **(B)** FOV portraying four correlation length scales and strengths for keratin$^+$, FOXP3$^+$, αSMA$^+$, and CD163$^+$ cells. Four circles with radii equal to the length scale parameters. **(C)** Length scales for select markers across CRC2-17. **(D)** CyCIF image showing CD20$^+$ a tertiary lymphoid structure (circled in pink) and a CD31$^+$ blood vessel (circled in blue and also magnified in the yellow inset in the upper left) . **(E)** Spatial distribution of CD20$^+$ cells (blue density and contours) and CD31$^+$ cells (red density); annotated ROIs are labelled 1-5; budding region in dark blue. **(F)** Spatial distribution of keratin+ cells (red density and contours) with annotated ROIs in cyan and the three regions of the invasive front labelled A-C. **(G)** Spatial distribution of CD163$^+$ cells labelled as in panel E. **(H)** Virtual 1mm TMA cores from CRC1 section 097, and 0.6mm cores from a real TMA of other CRC patients. **(I)** Estimates of tumor composition using vTMA cores or random sampling. Means indicated in cyan. **(J)** Correlation of select cell-type pairs amongst 10 nearest neighbors. Value from whole section shown in green. unless otherwise noted.

Figure 6. Patient variation compared to correlated sampling error in TMAs.

(A) Correlation functions between select cell-type pairs as estimated from vTMAs or random sampling. Shaded regions represent 95% confidence intervals. **(B)** Estimated correlation functions from several cores, portraying a variety of inter-cellular relations. **(C)** The four cores highlighted in (B). **(D)** *(Left)* Variation in the fraction of various marker-positive cells across specimens CDC2-17 as determined using whole-slide (blue) or TMA (green) data; or among specimens 18 to 93 based on TMA data alone. Data points and 1st-3rd quartiles (box plots) are shown. Proportions <0.0001 are denoted as a single data point along the dotted line. *(Right)* Percent of variance in TMA estimates that can be attributed to sampling error, after removing outliers. Expected improvement from sampling four cores per tumor is shown in beige. **(E)** Variation of each annotated ROI composition across sections of CRC1 for the same cell-types in Figure S1D.

Figure 7. Spatial distribution of immune cell-types and interactions.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

All images and plots of cell coordinates derive from CRC1 section 097. **(A-C)** Distribution of total $CD45^+$ (A), $PD\text{-}L1^+$ (B) and $PD\text{-}1^+$ (C) cells with annotated ROIs in cyan; TB denotes region with a high density of tumor budding; in panel B, the three regions of the invasive front are labelled A, B and C. **(D-F)** Co-occurrence of two cell types using a distance cutoff of 20 µm. *Below*: box plots show the 1^{st}-3^{rd} quartiles of fraction of first cell type that co-occur with the second cell type (as listed in the panel legend) for each annotated ROI well as the *p* value for all two-way comparisons (unpaired t-test). **(D)** $FOXP3^+$ Treg (green) and $CD8^+$ cytotoxic T cells (red); **(E)** $keratin^+$ $PDL1^+$ tumor cells (green) and $CD8^+$ $PD1^+$ T cells (red); **(F)** $CD68^+$ $PDL1^+$ macrophages and $CD8^+$ $PD1^+$ T cells (red). **(G)** Representative image of immune microenvironment surrounding the region containing budding tumor cells (denoted TB). Arrows indicate buds. **(H)** Schematic representation of immune environments associated with each invasive margin; infiltrative budding cells at invasive margin A associate with $FOXP3^+$ Tregs and $PD\text{-}L1^+$ macrophages to suppress $PD\text{-}1^+/CD8^+$ cytotoxic T cells; few immune cells and minimal interaction are identified at invasive margin B, possibly due to physical constraints imposed by mucinous pools surrounding tumor cells; pushing tumor fronts at invasive margin C associate with $PD\text{-}L1^+$ macrophages, but not $FOXP3^+$ Tregs, to suppress cytotoxic T cells.

DATA AND SOFTWARE AVAILABILITY

All full resolution images, derived image data (e.g. segmentation masks) and all cell count tables will be publicly released via the NCI-recognized repository for Human Tumor Atlas Network (HTAN; https://humantumoratlas.org/) at Sage Synapse. A version of this data is available at https://www.synapse.org/#!Synapse:syn18434611/wiki/597418. Several of the figure panels in this paper are available with text and audio narration for anonymous on-line browsing using MINERVA software (Rashid et al., 2020), which supports zoom, pan and selection actions without requiring the installation of software. A Minerva story with an overview of CRC1 (section 096 & 097) can be found at: cycif.org/crc1-intro and the 25 CRC1 Z-sections can be found at: cycif.org/crc1-3d.

All software used in this manuscript is freely available via GitHub as described in (Schapiro et al., 2021) and references therein.

SUPPLEMENTAL INFORMATION

Supplemental information includes seven figures, six tables, and two movies (Lin-Wang-Coy-CRC1-Movie 1-Lumen View and Lin-Wang-Coy-CRC1-Movie 2-Budding). Interactive data viewing is

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

possible via an anonymous Amazon S3 link via https://www.cycif.org/crc1-intro and https://www.cycif.org/crc1-3d.

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Dataset 1: antibodies, images and serial sections. Related to Figure 1.

(A) Representative images of main antibody panel from CRC1. Blue is the nuclear stain (Hoechst 33342). **(B)** UMAP plot of scRNAseq data from CRC1, and cell-types identified by Leiden clustering. **(C)** Marker-guided sub-clustering was performed as described in methods. Positive cells highlighted in yellow. **(D)** Cell-types mapped across CRC1/097. Cell type definitions and main classification markers are as indicated. A detailed marker/reference dictionary is in Table S5.

Figure S2. kNN-classification of epithelial histology. Related to Figure 2.

(A) Precision and recall of the classifiers trained on each section. **(B)** *(Left)* Shannon entropy of kNN-classification for each cell. Normal cells have low-entropy, indicating high-confidence classification. Regions used for training were also high-confidence, as expected by definition. Most tumor regions were classified between classes, i.e. high-entropy. *(Right)* The relative weight of each class is visualized by hue. **(C)** Dimensional reduction of subsampled single-cell expression from keratin$^+$ cells by tSNE, with pathologist annotations indicated by color. Each of the four marker panels provide enough information to cluster normal epithelial cells (black) separately from tumor cells, despite limited overlap in markers between panels. Different annotations roughly occupy different regions of expression space, indicating that expression and morphology are correlated, but tumor cells largely form a continuous distribution, supporting the existence of mixed morphologies. **(D)** Region of IM-A predicted to have morphology alternating between glandular and solid morphology on short length scales. **(E)** Contours of epithelial PCNA expression in Figure 2D.

Figure S3. Quantification of pathologist annotated tumor buds. Related to Figure 3.

(A) Proportion of annotated budding cells amongst Keratin$^+$ cells across sections. **(B)** Different magnifications of the annotated budding region in section 097. **(C)** H&E FOV in section 096 IM-A. A few budding cells are indicated by arrowheads. **(D)** Corresponding CyCIF channels. Red outlines separate main tumor mass from buds (yellow outline). **(E)** Log-intensity of markers and their dependency on cluster-size for tumor cells in CRC2-17, as in Figure 3C. Outliers not shown (~5% of

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each box). Trends are the same as in CRC1. **(F)** Location of 3D viewpoints in Figure 4B, represented on section 054. Arrows approximately represent viewing angle.

Figure S4. Spatial correlations and predicted sampling errors. Related to Figure 5

(A) The length scales and correlation strengths of cell-types (top) and marker log-intensities (bottom) for CRC2-17. **(B)** Estimation error of vTMAs summarized by fold-reduction in effective sample size, N/N_{eff}, for marker log-intensities and cell-type compositions. Observed error is compared to that predicted by accounting for exponential fits of spatial correlations in the Central Limit Theorem. Predictions from exponential assumptions fit observations closely ($R^2 = 0.97$, green), and deviations from perfect prediction (red) might be attributed to slight violations of assumptions (e.g. deviations from exponential decay, variation in cell-density). **(C)** Order-of-magnitude estimates of N/N_{eff} for keratin$^+$, FOXP3$^+$, and α-SMA$^+$ based on the scaling law in Eq 1 (shaded boxes) compared to 0.6mm vTMA cores. Order-of-magnitude predictions are good up to a small factor.

Figure S5. Variation of estimated composition and features of patient sections. Related to Figure 5.

(A) FOVs of patient sections with low and high Ki-67$^+$ cell abundance. Green circles visualize the length scale of Ki-67$^+$ cells. **(B)** Variation of weighted cell-type scores between patients, and percent contribution of sampling error to TMA estimates of scores, analogous to Figure 6A. Scores with random weights still perform poorly, with 95% of such scores (magenta) resulting in ~90% sampling error. However, the theoretically optimal score (Best) derived from multivariate correlations gives signal-to-noise comparable to Ki-67$^+$. A Rationalized score (imitating weights one might assign from prior knowledge) diminishes signal-to-noise substantially, despite being constructed to imitate the Best score, demonstrating the importance of quantitatively accounting for multivariate information. Ki-67$^+$ was excluded in all scores. The Best and Rationalized score weights are shown as heatmaps. This score is a conceptual toy example to demonstrate the existence of valuable information amidst multivariate data. **(C)** Variation of cell-type composition between sections of a single tumor (CRC-1) and sections from different patient tumors (CRC2-17). Section sampling error is typically a minority of the variance between patient sections. Indicated outliers were excluded from the percent-variance calculation. **(D)** Variation of cell-type spatial correlation strengths and length scales across Z-sections of the tissue map (blue) and across patients (red). In most cases, variation within the tissue map is smaller than that between patients, and shows no signs of bias.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S6. Kaplan-Meier curves. Related to Figure 6.

Kaplan-Meier (KM) curves for overall survival (OS) in CRC2-17, with patients split by indicated marker abundances estimated from TMA (A&C) or WSI (B&D). **(A&B)** KM curves generated from data stratified with α-SMA$^+$ percentage in each patient sample. The cutoff of α-SMA$^+$ is 40%. **(C&D)** KM curves generated from data stratified with mean CD4 expression level in each patient sample. The cutoff of mean CD4 level is 3500 AFU.

Figure S7. Cell compositions in CRC2-17. Related to Figure 7.

(A) tSNE plots based on CyCIF data for specimens CRC2-17 (excluding data from DNA staining). Cell types are as described in Supplemental Table S5 and Supplemental Figure S1. Tumor/epithelium (T/E), stroma (S) and immune (I) regions are outlined in black. The tSNE plot for CRC1 is reproduced from Figure 1D for reference. **(B)** Cell-type composition for CRC1-17 shown as stacked bar graphs with the same color code as in panel A. **(C)** Abundance of immune cell types involved in immunosuppression based on detailed analysis of specimen CRC1.

SUPPLEMENTARY TABLE LEGENDS

Supplementary Table 1. Clinical information for colorectal cancer cohort. Demographic and diagnostic information for all patient-derived specimens in this study. CRC1 was analyzed in 3D and CRC2-17 in whole-slide 2D. Other specimens were images as TMA cores, as described in the text and Figure 1.

Supplementary Table 2. Sectioning plan for specimen CRC1. Thickness and staining plan for CRC1 sections show in Figure 1. All CyCIF sections other than 44-46 were stained using the Primary CyCIF antibody panel described in supplementary Table 3. Sections 44-46 were stained as described in supplementary Table 4. Numbers refer to the HTAN universal ID scheme used to access underlying Level 2 to Level 4 data.

Supplementary Table 3. Primary CyCIF antibody panel. Antibodies used to stain all CRC1 sections 44-46 including CRC2-17 sections and TMAs. CST refers to Cell Signaling Technologies (Beverley MA USA); RRID refers to the Research Resource Identifier available at https://scicrunch.org/resources.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 4. Supplementary CyCIF antibody panel. Antibodies used to stain CRC1 sections 44-46. Abbreviations as in supplementary Table 2.

Supplementary Table 5. Cell Type Dictionary. Cell type assignments based on marker intensities. The first tab shows the primary discriminating markers and tabs 2 and 3 show assignments based on all markers in the panel.

Supplementary Table 6. Cell type compositions for regions of interest in CRC1. Cell type compositions for pathologist-defined regions of interest (see Figure 1B) across all sections processed for CyCIF. Cell type definitions as in supplementary Table 6

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

This manuscript contains no unique reagents or resources; all antibodies are available commercially and all data can be accessed via the HTAN portal.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Tissue

Unfixed (fresh) tissue from a resection of a colon adenocarcinoma (CRC1) was isolated by the Cooperative Human Tissue Network (CHTN) for single cell RNA-sequencing. A portion of the sample was formalin fixed and paraffin embedded (FFPE) and tissue sections were generated by the CHTN. Additional FFPE colon adenocarcinoma specimens were retrieved from the archives of the Department of Pathology at Brigham and Women's Hospital with Institutional Review Board (IRB) approval as part of a discarded/excess tissue protocol. 93 different tumor samples were used to construct a tissue microarray (HTMA 402; four 0.6 mm diameter cores were extracted from the FFPE donor blocks and assembled into a recipient TMA block). Whole slide sections of 17 of these colon carcinoma specimens were also studied, after the four cores were removed. Clinical metadata was abstracted from the medical record and clinical biospecimen metadata was recorded by CHTN.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHOD DETAILS

t-CyCIF protocol

Tissue-based cyclic immunofluorescence (t-CyCIF) was performed as described (Lin et al., 2018) and in and at protocols.io (dx.doi.org/10.17504/protocols.io.bjiukkew). In brief, the BOND RX Automated IHC/ISH Stainer was used to bake FFPE slides at 60°C for 30 min, dewax using Bond Dewax solution at 72°C, and perform antigen retrieval using Epitope Retrieval 1 (Leica^{TM}) solution at 100°C for 20 min. Slides underwent multiple cycles of antibody incubation, imaging, and fluorophore inactivation. All antibodies were incubated overnight at 4°C in the dark. Slides were stained with Hoechst 33342 for 10 min at room temperature in the dark following antibody incubation in every cycle. Coverslips were wet-mounted using 200µL of 10% Glycerol in PBS prior to imaging. Images were taken using a 20x objective (0.75 NA) on a CyteFinder slide scanning fluorescence microscope (RareCyte Inc. Seattle WA). Fluorophores were inactivated by incubating slides in a 4.5% H_2O_2, 24mM NaOH in PBS solution and placing under an LED light source for 1 hr.

Single-cell RNA-sequencing

Samples for scRNA-seq were processed according to the HTAN protocol (https://www.biorxiv.org/content/10.1101/2021.01.11.426044v1). Surgical tissues were removed and placed into RPMI solution and transported directly to the processing laboratory within 10 min. Tissue samples were immediately minced to approximately 4mm^2 and washed with DPBS. The samples were then incubated in chelation buffer (4mM EDTA, 0.5 mM DTT) at 4°C for 1 hr and 15 min. Then, the resulting suspensions were dissociated with cold protease and DNAse I for 25 min. The suspensions were triturated throughout the process, every 10 min, then washed three times with DPBS before encapsulation. Single cells were encapsulated and barcoded using the inDrop scRNA-seq platform as previously described (Banerjee et al., 2020), targeting about 2,500 cells. Sequencing libraries were prepared using TruDrop library structure (Southard-Smith et al., 2020). Sequencing was performed on the NovaSeq 6000 (150 bp paired end) at a depth of approximately 150 million reads per sample.

QUANTIFICATION AND STATISTICAL ANALYSIS

Image processing and data quantification

Image analysis was performed with the Docker-based NextFlow pipeline MCMICRO)(Schapiro et al., 2021) and with customized scripts in Pytho, ImageJ and MATLAB. All code is available on GitHub

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(https://github.com/sorgerlab/cycif . Briefly, after raw images were acquired, the stitching and registration of different tiles and cycles were done with MCMICRO using the ASHLAR module. The assembled OME.TIFF files from each slide were then passed through quantification modules. For background subtraction, a rolling ball algorithm with 50-pixel radius was applied using ImageJ/Fiji. For segmentation and quantification, UNMICST2 was used (Schapiro et al., 2021) supplemented by customized ImageJ scripts (Lin et al., 2018) to generate single-cell data. More details and source code can be found at www.cycif.org.

Single-cell data QC for CyCIF

Single-cell data was passed through several QC steps during generation of *cell feature table*. Initial QC was done simultaneously with segmentation and quantification, so that cells lost from the specimen in the later cycles would not be included in the output. Next, single-cell data was filtered based on mean Hoechst staining intensity across cycles; cells with coefficient of variation (CV) greater than three standard deviation from the mean were discarded as were any objected identified by segmentation as "cells" but having no DNA intensity. These steps are designed to eliminate cells in which the nuclei are not included as a result of sectioning. Highly auto-fluorescent (AF) cells (measured in cycle 1 or 2) were also removed from the analysis, using a customized MATLAB script that applied a Gaussian Mixture Model (GMM) to identify high-AF population. More details and scripts can be found online (https://github.com/sorgerlab/cycif).

Cell type Identification using CyCIF data

Multiparameter single-cell intensity data was used for generating binary gates. For the main CyCIF panels, 16 measurements (keratin, Ki-67, CD3, CD20, CD45RO, CD4 CD8a, CD68 CD163, FOXP3, PD1, PDL1, CD31, α-SMA, desmin, and CD45) were subjected to binary gating. All samples and markers were gated independently. A customized MATLAB script was used to apply 2-component Gaussian Mixture Modeling and generate the initial gate, followed by human-inspection and adjustment. Double or triple gates were also generated via Boolean operation in single-cell data. For hierarchal-cell-type identification, a modified SYLARAS algorithm (Baker et al., 2020) was applied with these datasets, and a total of 31 different cell types were assigned using the 16 markers described above. Additional markers (e,g E-cadherin) were considered to be continuous variables and used for analysis but not cell type assignment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Delaunay cluster-sizes of Keratin+ cells

A neighborhood graph was first constructed for all segmented cell centroids using a Delaunay triangulation, removing edges whose lengths were greater than 20μm. Then, the keratin$^+$ neighborhood graph was defined as the subgraph restricted to the keratin$^+$ cells (i.e., removing all nodes and edges connected to keratin-negative cells). Each keratin$^+$ cell's cluster-size was defined as the number of nodes in its connected component of the subgraph. For quantification of marker expression dependence on cluster-size, cells annotated as normal mucosa were removed from the keratin$^+$ subgraph. In the 25 CRC1 Z-sections, cells in the upper-left corner (1cm x 1cm) were also removed; this region contained keratin$^+$ cells of reactive, benign, and mesothelial origin, as opposed to tumor cells of interest.

Biased downsampling based on cluster-size for tSNE visualization

By definition, most tumor cells have a large cluster-size. Therefore, to visualize the cluster-size dependence of marker expression with tSNE, we downsampled cells by stochastically rejecting cells at frequency $1 - (1/n_c)^4$, for cluster-size n_c. The power of 4 was chosen empirically to balance the representation of various cluster-sizes. Final t-SNE plots were made by further subsampling 1000 cells from each section uniformly. The t-SNE plots were computed using the following markers: Na-K ATPase, Ki-67, keratin, PD-L1, E-cadherin, vimentin, CDX2, lamin ABC, desmin, PCNA.

kNN-classification of epithelial cell morphologies trained on pathologist annotations

To develop a kNN classifier for pathologist-annotated regions of interest (ROIs) epithelial cells were defined by gating cells with a univariate, 2-component, Gaussian Mixture Model on the relevant marker (keratin, cytokeratin 19, cytokeratin 18, or E-cadherin) in each section. A kNN-classifier was trained on the annotated, epithelial cells using CyCIF marker expression as predictors, and annotated ROI labels as responses. Markers that exhibited unexpected optical artefacts or significant tissue loss were not used. Learning and prediction were performed using MATLAB's $fitcknn()$ and $predict()$ functions, with k=40 neighbors. The prior probability of each label was set as uniform. In each section, there were at least 2,000 annotated cells for each label. Annotated cells were split 50/50 into training and validation sets. Posterior probability colors were visualized based on its vector of classification posterior probabilities (p_1, p_2, p_3, p_4), for 1: normal, 2: glandular classes, 3: solid, and 4: mucinous. The RGB-values of each cell were then defined as:

$$(R, G, B) = (p_2, p_3, p_4)/\max(p_i)$$

to capture the relative weight of each class.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the sections in the primary dataset (e.g. section 044), the following markers were used as predictors: Na-K ATPase, Ki-67, keratin, PD-L1, E-cadherin, vimentin, CDX2, lamin, desmin, PCNA, autofluorescence.

For section 046, which was stained with an extended antibody panel, markers used as predictors were: cyclin B1, cytokeratin 20, cytokeratin 18, NUP98, cytokeratin 8, PD-L1, acetyl-tubulin, p62, pan-keratin, lamin A/C, tubulin.

For sections 045 and 047, which were also stained with different extended antibody panels, we used all artefact-free markers (totaling 29 and 36 respectively).

Contour plots of epithelial cell marker expression gradients

Contours represent level sets for the average marker expression of the 400 nearest tumor cells, and were computed using the MATLAB *contour()* function.

3D registration of CRC1 serial sections

All CyCIF sections were registered using a custom script written in MATLAB 2018 (Mathworks). Briefly, each section was first registered using a rigid transformation followed by elastic deformations starting at section 12 and cascading towards the top and bottom sections. For the rigid transformation, an early cycle Hoechst signal with minimal artefacts from each section was selected. All channels were padded by an equivalent of 1600 pixels along all borders when registering at full resolution. Rigid transformation required consistent landmarks across all sections. Therefore, we identified two such features: the edge of the mucosa section and a point where it transitions into the stromal region. This region was annotated on several downsampled sections, providing training data for a UNet model to estimate fuzzy locations of the transition point and the mucosal edge. Starting from section 12, and taking the centroid of each fuzzy estimate as that section's transition point, all 25 sections were aligned by translation. Each section was then rotated around the transition point until the fuzzy estimates for the edge of the mucosa region overlapped maximally between sections.

For subsequent elastic deformation, we manually selected between 25-35 control points across each section. Most control points were located near the site of budding cells. Then, using local weighted means with these control points via the *fitgeotrans*() MATLAB function, we applied a deformation starting from section 12 towards section 1 and 25. Finally, we applied Demon's algorithm to refine registration further. Images were downsampled by a factor of 0.25 and histogram matched, before

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

applying the *imregdemons*() MATLAB function with an accumulated field smoothing of 1.5 and downsampling with 7 pyramid levels. Demon's algorithm was applied starting from section 12.

3D visualization of registered CRC1 serial sections

Using Imaris, images were Gaussian-blurred, and an intensity threshold was applied to define regions (e.g., keratin^{+}). Connectivity of buds or mucin pools were defined on blurred, thresholded voxels.

Virtual XZ-section

Virtual XZ-sections were constructed from the 3D registered image. Pixels in each XY-section (i.e., original imaging plane) were downscaled such that a downscaled pixel represented the average intensity in a 5μm x 5μm block. A virtual 5μm XZ-section was constructed from the pixels at a fixed Y-position.

Virtual TMA cores and fold-change in effective sample size N/N_{eff}

Virtual TMAs were constructed from whole-slide sections by randomly selecting a central cell, and including all cells within 500μm of the central cell's centroid as one core. For each virtual TMA core, a matching, uniform random sample was generated from the whole-slide section with an equal number of cells. The standard-errors of the mean from virtual TMA (i.e., regional) sampling (σ_{TMA}) or random sampling (σ_{random}) were estimated from the means of 1,000 cores and their matched, random samples. The effective sample size N/N_{eff} was defined as the square of the standard-errors' ratios:

$$\sigma^2{}_{TMA}/\sigma^2{}_{random}.$$

Spatial correlation functions and predicting standard-error of regional sampling

For each sample (whole-slide or TMA core), spatial correlation functions $C_{AB}(r)$ were calculated for a pair of variables A, B and a nearest-neighbor index r. Specifically, $C_{AB}(r)$ was given by the Pearson correlation between cells' A-values and their r^{th} - nearest neighbors' B-values. Each r index was associated to the average, inter-cell-centroid distance $d(r)$ of all r^{th} - nearest neighbors in a sample. Correlations were computed up to $r = 200$. Each $C_{AB}(d(r))$ was fit to an exponential $c_0\exp(c_1 d)$ for parameters c_0, c_1, over the range of $5 < r < 200$ to avoid spurious correlations between adjacent cells that may arise from image segmentation errors. Correlation strength was defined as c_0, and length scale $l = -1/c_1$. Fits were performed with the $fit()$ MATLAB function with default options. We subsequently estimated the standard-error of the mean of a variable A for a regional sample of N correlated cells as follows. First, we computed the $N \times N$ matrix of inter-cellular distances d_{ij}, and then

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

computed the $N \times N$ correlation matrix Σ_N between cells using the fit of the spatial correlation function $C_{AA}(d)$. By the Central Limit Theorem for weakly-dependent variables (Ibragimov, 1962), we expect the standard- error of the mean for N samples to be $\sqrt{|\Sigma_N|}/N$, for $|\Sigma_N|$ the sum of all entries in Σ_N.

Scaling analysis of fold-change in effective sample size N/N$_{eff}$

For a variable A with variance $\sigma^2 = 1$, the fold-change N/N$_{eff}$ is defined as:

$$\frac{N}{N_{eff}} = \frac{\sigma^2{}_{TMA}}{\sigma^2{}_{random}} = \frac{|\Sigma_N|/N^2}{1/N} = \frac{|\Sigma_N|}{N}.$$

The final term can be interpreted as the sum of correlations between an average cell and all other cells in the sample region R. Choosing a coordinate system with an average cell at the origin, we approximate the sum as an integral:

$$\frac{|\Sigma_N|}{N} \approx \int^R d^n x\, C_{AA}(|x|)\rho(x)$$

$$= \int^R d^n x\, c_0 \exp(-|x|/l)\rho(x).$$

Where $\rho(x)$ is the density of cells, and n is the spatial dimension of the regional sample. If we assume a uniform density $\rho(x) \sim 1/l^n{}_{cell}$ for a cell length scale l_{cell}, and change variables in the integral to eliminate the length scale l, we have:

$$\frac{|\Sigma_N|}{N} \sim c_0 \left(\frac{l}{l_{cell}}\right)^n \int^{R'} d^n u \exp(-u),$$

which gives us a scaling relation with which we can roughly estimate N/N$_{eff}$ from parameters:

$$\frac{N}{N_{eff}} \sim c_0 \left(\frac{l}{l_{cell}}\right)^n.$$

Variance between patient TMAs due to sampling error, and an optimal score

For any given cell-type's %-composition, we computed the variance of estimates from the whole-slide tumor regions of each patient, $\sigma^2{}_{patient}$, and the variance of estimates from TMA cores, $\sigma^2{}_{TMA}$. We considered $\sigma^2{}_{patient}$ to be the biological variance of $\sigma^2{}_{TMA}$, and remaining variance to be residual error from sampling, $\sigma^2{}_{sampling}$. Percent of variance explained by sampling was given by $\sigma^2{}_{sampling}/\sigma^2{}_{TMA}$. For the hypothetical scenario of averaging 4 cores, $\sigma^2{}_{sampling}$ would be 4-fold lower, and

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

percent variance explained was given by $(\sigma^2{}_{sampling}/4)/\,(\sigma^2{}_{sampling}/4 + \sigma^2{}_{patient}\;)$. Outliers in each distribution, as indicated in each boxplot, were excluded from the variance calculations.

We sought an optimal score that could be constructed by linearly combining the *k* separate cell-types' %-compositions with a *k*-dimensional, weight vector $\vec{w}$. Such a score would be given by maximizing $(\overrightarrow{w^T}\Sigma_{patient}\vec{w})/(\overrightarrow{w^T}\Sigma_{TMA}\vec{w})$, where Σ is the covariance matrix of different cell-types' %-composition across patients or TMA cores respectively. These two covariance matrices are readily computed from the data. Since the quantity we intend to maximize is a generalized Rayleigh quotient, the optimal $\vec{w}$ is the generalized eigenvector of $\Sigma_{patient}$ and Σ_{TMA} with the largest generalized eigenvalue.

Progression-free survival analysis and Kaplan-Meier curve generation

A MATLAB package, MatSurv was used for statistical analysis and creating KM plots with risk tables. The log-rank p-value and hazard ratio (HR) were calculated using Cox proportional hazards regression, as described in (Creed et al., 2020). The stratification of patients was done by thresholding mean intensity data or the mean positive percentage from each sample (TMA or WSI). The CyCIF data from corresponding CRC WSI (16 sections) and CRC TMA (64 cores) were used.

scRNA-seq data analysis

Following sample demultiplexing from the sequencer, reads were filtered, sorted by their barcode of origin, and aligned to the reference transcriptome to generate a counts matrix using the DropEst pipeline(Petukhov et al., 2018). Barcodes containing cells were identified using dropkick (Heiser et al., 2020). Batches were combined and consensus non-negative matrix factorization (cNMF; (Kotliar et al., 2019)) was performed to identify metagenes in the resulting cell matrix, assigning "usage" scores for each factor to all cells. The factors or metagenes contain gene loadings that rank detected genes by their contribution to each factor, which are shown on UMAP embeddings in descending order. CytoTRACE (Gulati, et al. 2020) was also run using the web portal at https://cytotrace.stanford.edu/ to calculate "stemness" or cellular plasticity scores based on genetic diversity. Leiden clustering (Traag et al., 2019) and PAGA (Wolf et al., 2019) graph construction was performed on principal component analysis of the normalized and arcsinh-transformed raw counts matrix. A two-dimensional UMAP (McInnes et al., 2020) embedding was then generated using SCANPY (Wolf et al., 2018) based on principal component analysis and initial cluster positions determined by PAGA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Aeffner, F., Zarella, M.D., Buchbinder, N., Bui, M.M., Goodman, M.R., Hartman, D.J., Lujan, G.M., Molani, M.A., Parwani, A.V., Lillard, K., et al. (2019). Introduction to Digital Image Analysis in Whole-slide Imaging: A White Paper from the Digital Pathology Association. J. Pathol. Inform. *10*, 9.

Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., Meyer, L., Gress, D.M., Byrd, D.R., and Winchester, D.P. (2017). The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA. Cancer J. Clin. *67*, 93–99.

Angell, H.K., Bruni, D., Barrett, J.C., Herbst, R., and Galon, J. (2020). The Immunoscore: Colon Cancer and Beyond. Clin. Cancer Res. *26*, 332–339.

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.

Aponte, P.M., and Caicedo, A. (2017). Stemness in Cancer: Stem Cells, Cancer Stem Cells, and Their Microenvironment. Stem Cells Int. *2017*.

Baker, G.J., Muhlich, J.L., Palaniappan, S.K., Moore, J.K., Davis, S.H., Santagata, S., and Sorger, P.K. (2020). SYLARAS: A Platform for the Statistical Analysis and Visual Display of Systemic Immunoprofiling Data and Its Application to Glioblastoma. Cell Syst. *11*, 272-285.e9.

Banerjee, A., Herring, C.A., Chen, B., Kim, H., Simmons, A.J., Southard-Smith, A.N., Allaman, M.M., White, J.R., Macedonia, M.C., Mckinley, E.T., et al. (2020). Succinate Produced by Intestinal Microbes Promotes Specification of Tuft Cells to Suppress Ileal Inflammation. Gastroenterology *159*, 2101-2115.e5.

Bendall, S.C., Simonds, E.F., Qiu, P., Amir, E.D., Krutzik, P.O., Finck, R., Bruggner, R.V., Melamed, R., Trejo, A., Ornatsky, O.I., et al. (2011). Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science *332*, 687–696.

Bhatia, R., Gautam, S.K., Cannon, A., Thompson, C., Hall, B.R., Aithal, A., Banerjee, K., Jain, M., Solheim, J.C., Kumar, S., et al. (2019). Cancer-associated mucins: role in immune modulation and metastasis. Cancer Metastasis Rev. *38*, 223–236.

Black, J.R.M., and McGranahan, N. (2021). Genetic and non-genetic clonal diversity in cancer evolution. Nat. Rev. Cancer.

Bresalier, R.S. (2002). Intestinal mucin and colorectal cancer: It's not just goo. Gastroenterology *123*, 648–649.

Bronsert, P., Enderle-Ammour, K., Bader, M., Timme, S., Kuehs, M., Csanadi, A., Kayser, G., Kohler, I., Bausch, D., Hoeppner, J., et al. (2014). Cancer cell invasion and EMT marker expression: a three-dimensional study of the human cancer-host interface: 3D cancer-host interface. J. Pathol. *234*, 410–422.

Centeno, I., Paasinen Sohns, A., Flury, M., Galván, J.A., Zahnd, S., Koelzer, V.H., Sokol, L., Dawson, H.E., Lugli, A., Cathomas, G., et al. (2017). DNA profiling of tumor buds in colorectal cancer indicates

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that they have the same mutation profile as the tumor from which they derive. Virchows Arch. Int. J. Pathol. *470*, 341–346.

Cianchi, F., Messerini, L., Palomba, A., Boddi, V., Perigli, G., Pucciani, F., Bechi, P., and Cortesini, C. (1997). Character of the invasive margin in colorectal cancer: does it improve prognostic information of Dukes staging? Dis. Colon Rectum *40*, 1170–1175; discussion 1175-1176.

Coons, A.H., Creech, Hugh J, Jones, Norman, and Berliner, Ernst (1942). THE DEMONSTRATION OF PNEUMOCOCCAL ANTIGEN IN TISSUES BY THE USE OF FLUORESCENT ANTIBODY. J Immunol.

Creed, J.H., Gerke, T.A., and Berglund, A.E. (2020). MatSurv: Survival analysis and visualization in MATLAB. J. Open Source Softw. *5*, 1830.

Djuric, U., Zadeh, G., Aldape, K., and Diamandis, P. (2017). Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. Npj Precis. Oncol. *1*, 1–5.

Edin, S., Kaprio, T., Hagström, J., Larsson, P., Mustonen, H., Böckelman, C., Strigård, K., Gunnarsson, U., Haglund, C., and Palmqvist, R. (2019). The Prognostic Importance of CD20 + B lymphocytes in Colorectal Cancer and the Relation to Other Immune Cell subsets. Sci. Rep. *9*, 19997.

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

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., et al. (2006). Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. Science *313*, 1960–1964.

Genton, M.G., and Kleiber, W. (2015). Cross-Covariance Functions for Multivariate Geostatistics. Stat. Sci. *30*, 147–163.

Gerdes, M.J., Sevinsky, C.J., Sood, A., Adak, S., Bello, M.O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R.J., et al. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. Proc. Natl. Acad. Sci. U. S. A. *110*, 11982–11987.

Ghaznavi, F., Evans, A., Madabhushi, A., and Feldman, M. (2013). Digital imaging in pathology: whole-slide imaging and beyond. Annu. Rev. Pathol. *8*, 331–359.

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.

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.

Gosens, M.J.E.M., van Kempen, L.C.L., van de Velde, C.J.H., van Krieken, J.H.J.M., and Nagtegaal, I.D. (2007a). Loss of membranous Ep-CAM in budding colorectal carcinoma cells. Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc *20*, 221–232.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gosens, M.J.E.M., van Kempen, L.C.L., van de Velde, C.J.H., van Krieken, J.H.J.M., and Nagtegaal, I.D. (2007b). Loss of membranous Ep-CAM in budding colorectal carcinoma cells. Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc *20*, 221–232.

Guil-Luna, S., Mena, R., Navarrete-Sirvent, C., López-Sánchez, L.M., Khouadri, K., Toledano-Fonseca, M., Mantrana, A., Guler, I., Villar, C., Díaz, C., et al. (2020). Association of Tumor Budding With Immune Evasion Pathways in Primary Colorectal Cancer and Patient-Derived Xenografts. Front. Med. *7*.

Health, C. for D. and R. (2019). Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices.

Heiser, C.N., Wang, V.M., Chen, B., Hughey, J.J., and Lau, K.S. (2020). Automated quality control and cell identification of droplet-based single-cell data using dropkick. BioRxiv 2020.10.08.332288.

Hoffer, J., Rashid, R., Muhlich, J.L., Chen, Y.-A., Russell, D.P.W., Ruokonen, J., Krueger, R., Pfister, H., Santagata, S., and Sorger, P.K. (2020). Minerva: a light-weight, narrative image browser for multiplexed tissue images. J. Open Source Softw. *5*, 2579.

HuBMAP Consortium (2019). The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature *574*, 187–192.

Ibragimov, I.A. (1962). Some Limit Theorems for Stationary Processes. Theory Probab. Its Appl. *7*, 349–382.

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.

Kalluri, R., and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. J. Clin. Invest. *119*, 1420–1428.

Koelzer, V.H.M., and Lugli, A.M. (2014). The Tumor Border Configuration of Colorectal Cancer as a Histomorphological Prognostic Indicator. Front. Oncol. *4*.

Kondo, H., Ratcliffe, C.D.H., Hooper, S., Ellis, J., MacRae, J.I., Hennequart, M., Dunsby, C.W., Anderson, K.I., and Sahai, E. (2021). Single-cell resolved imaging reveals intra-tumor heterogeneity in glycolysis, transitions between metabolic states, and their regulatory mechanisms. Cell Rep. *34*, 108750.

Kopetz, S., Tabernero, J., Rosenberg, R., Jiang, Z.-Q., Moreno, V., Bachleitner-Hofmann, T., Lanza, G., Stork-Sloots, L., Maru, D., Simon, I., et al. (2015). Genomic Classifier ColoPrint Predicts Recurrence in Stage II Colorectal Cancer Patients More Accurately Than Clinical Factors. The Oncologist *20*, 127–133.

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

Lavrakas, P.J. (2008). Encyclopedia of Survey Research Methods (SAGE Publications).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lin, J.-R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P.M., Santagata, S., and Sorger, P.K. (2018). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. ELife *7*.

Lugli, A., Vlajnic, T., Giger, O., Karamitopoulou, E., Patsouris, E.S., Peros, G., Terracciano, L.M., and Zlobec, I. (2011). Intratumoral budding as a potential parameter of tumor progression in mismatch repair-proficient and mismatch repair-deficient colorectal cancer patients. Hum. Pathol. *42*, 1833–1840.

Lugli, A., Kirsch, R., Ajioka, Y., Bosman, F., Cathomas, G., Dawson, H., El Zimaity, H., Fléjou, J.-F., Hansen, T.P., Hartmann, A., et al. (2017a). Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. Mod. Pathol. *30*, 1299–1311.

Lugli, A., Kirsch, R., Ajioka, Y., Bosman, F., Cathomas, G., Dawson, H., El Zimaity, H., Fléjou, J.-F., Hansen, T.P., Hartmann, A., et al. (2017b). Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc *30*, 1299–1311.

Mani, S.A., Guo, W., Liao, M.-J., Eaton, E.Ng., Ayyanan, A., Zhou, A.Y., Brooks, M., Reinhard, F., Zhang, C.C., Shipitsin, M., et al. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell *133*, 704–715.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat.

Moffitt, J.R., Hao, J., Wang, G., Chen, K.H., Babcock, H.P., and Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc. Natl. Acad. Sci. *113*, 11046–11051.

O'Connor, B.D., Yuen, D., Chung, V., Duncan, A.G., Liu, X.K., Patricia, J., Paten, B., Stein, L., and Ferretti, V. (2017). The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. F1000Research *6*, 52.

Ohtani, H. (1998). Stromal reaction in cancer tissue: pathophysiologic significance of the expression of matrix-degrading enzymes in relation to matrix turnover and immune/inflammatory reactions. Pathol. Int. *48*, 1–9.

Oudin, M.J., and Weaver, V.M. (2016). Physical and Chemical Gradients in the Tumor Microenvironment Regulate Tumor Cell Invasion, Migration, and Metastasis. Cold Spring Harb. Symp. Quant. Biol. *81*, 189–205.

Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D.T., Samsonova, M.G., and Kharchenko, P.V. (2018). dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. Genome Biol. *19*, 78.

Plotkin, J.B., Chave, J., and Ashton, P.S. (2002). Cluster analysis of spatial patterns in Malaysian tree species. Am. Nat. *160*, 629–644.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Randall, E.C., Lopez, B.G.C., Peng, S., Regan, M.S., Abdelmoula, W.M., Basu, S.S., Santagata, S., Yoon, H., Haigis, M.C., Agar, J.N., et al. (2020). Localized Metabolomic Gradients in Patient-Derived Xenograft Models of Glioblastoma. Cancer Res. *80*, 1258–1267.

Rashid, R., Chen, Y.-A., Hoffer, J., Muhlich, J.L., Lin, J.-R., Krueger, R., Pfister, H., Mitchell, R., Santagata, S., and Sorger, P.K. (2020). Interpretative guides for interacting with tissue atlas and digital pathology data using the Minerva browser. BioRxiv 2020.03.27.001834.

Reynolds, A.M., Sword, G.A., Simpson, S.J., and Reynolds, D.R. (2009). Predator Percolation, Insect Outbreaks, and Phase Polyphenism. Curr. Biol. *19*, 20–24.

Rogers, K.W., and Schier, A.F. (2011). Morphogen gradients: from generation to interpretation. Annu. Rev. Cell Dev. Biol. *27*, 377–407.

Rogers, A.C., Winter, D.C., Heeney, A., Gibbons, D., Lugli, A., Puppa, G., and Sheahan, K. (2016). Systematic review and meta-analysis of the impact of tumour budding in colorectal cancer. Br. J. Cancer *115*, 831–840.

Rogers, T.-M., Russell, P.A., Wright, G., Wainer, Z., Pang, J.-M., Henricksen, L.A., Singh, S., Stanislaw, S., Grille, J., Roberts, E., et al. (2015). Comparison of Methods in the Detection of ALK and ROS1 Rearrangements in Lung Cancer. J. Thorac. Oncol. *10*, 611–618.

Rohban, M.H., Singh, S., Wu, X., Berthet, J.B., Bray, M.-A., Shrestha, Y., Varelas, X., Boehm, J.S., and Carpenter, A.E. (2017). Systematic morphological profiling of human gene and allele function via Cell Painting. ELife *6*, e24060.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al. (2020a). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al. (2020b). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell *181*, 236–249.

Rubio, C.A. (2007). Further studies on the arrest of cell proliferation in tumor cells at the invading front of colonic adenocarcinoma. J. Gastroenterol. Hepatol. *22*, 1877–1881.

Rubio, C.A. (2008). Arrest of cell proliferation in budding tumor cells ahead of the invading edge of colonic carcinomas. A preliminary report. Anticancer Res. *28*, 2417–2420.

Saka, S.K., Wang, Y., Kishi, J.Y., Zhu, A., Zeng, Y., Xie, W., Kirli, K., Yapp, C., Cicconet, M., Beliveau, B.J., et al. (2019). Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. Nat. Biotechnol. *37*, 1080–1090.

Schapiro, D., Sokolov, A., Yapp, C., Muhlich, J.L., Hess, J., Lin, J.-R., Chen, Y.-A., Nariya, M.K., Baker, G.J., Ruokonen, J., et al. (2021). MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging. BioRxiv 2021.03.15.435473.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Sethna, J. (2006). Statistical Mechanics: Entropy, Order Parameters, and Complexity (OUP Oxford).

Sharma, A., Merritt, E., Hu, X., Cruz, A., Jiang, C., Sarkodie, H., Zhou, Z., Malhotra, J., Riedlinger, G.M., and De, S. (2019). Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung Tumors. Cell Rep. *29*, 2164-2174.e5.

Sokal, R.R., and Thomson, J.D. (1987). Applications of Spatial Autocorrelation in Ecology. In Develoments in Numerical Ecology, P. Legendre, and L. Legendre, eds. (Berlin, Heidelberg: Springer), pp. 431–466.

Southard-Smith, A.N., Simmons, A.J., Chen, B., Jones, A.L., Ramirez Solano, M.A., Vega, P.N., Scurrah, C.R., Zhao, Y., Brenan, M.J., Xuan, J., et al. (2020). Dual indexed library design enables compatibility of in-Drop single-cell RNA-sequencing with exAMP chemistry sequencing platforms. BMC Genomics *21*, 456.

Stroup, W.W. (2002). Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. J. Agric. Biol. Environ. Stat. *7*, 491–511.

Sung, C.O., Seo, J.W., Kim, K.-M., Do, I.-G., Kim, S.W., and Park, C.-K. (2008). Clinical significance of signet-ring cells in colorectal mucinous adenocarcinoma. Mod. Pathol. *21*, 1533–1541.

Svensson, V., Teichmann, S.A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. Nat. Methods *15*, 343–346.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

Uhler, C., and Shivashankar, G.V. (2018). Nuclear Mechanopathology and Cancer Diagnosis. Trends Cancer *4*, 320–331.

Venet, D., Dumont, J.E., and Detours, V. (2011). Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLOS Comput. Biol. *7*, e1002240.

Verhaak, R.G.W., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110.

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.

Weiser, M.R. (2018). AJCC 8th Edition: Colorectal Cancer. Ann. Surg. Oncol. *25*, 1454–1455.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Williams, E., Moore, J., Li, S.W., Rustici, G., Tarkowska, A., Chessel, A., Leo, S., Antal, B., Ferguson, R.K., Sarkans, U., et al. (2017). The Image Data Resource: A Bioimage Data Integration and Publication Platform. Nat. Methods *14*, 775–781.

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

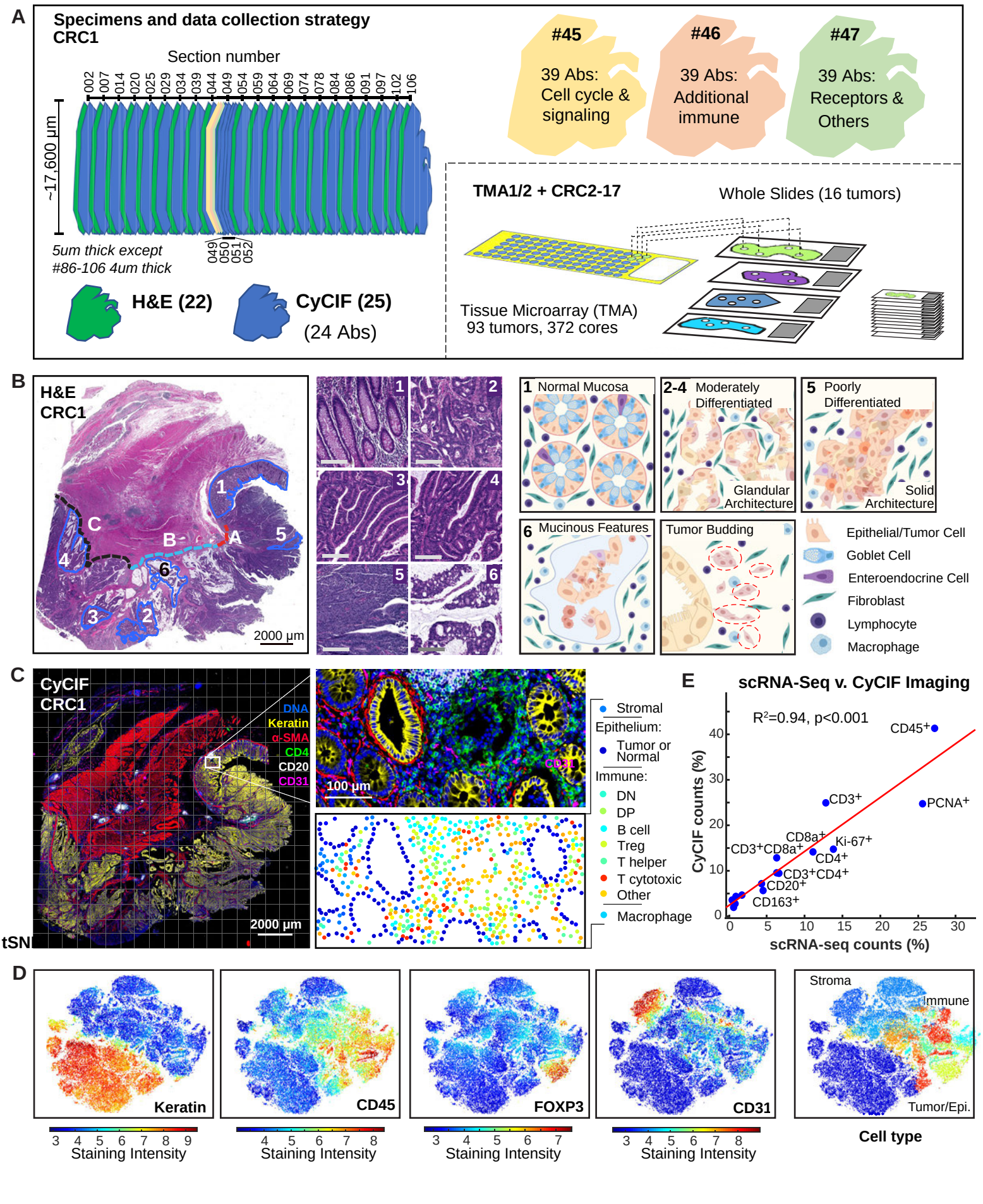
Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59.

Zlobec, I., and Lugli, A. (2010). Epithelial mesenchymal transition and tumor budding in aggressive colorectal cancer: Tumor budding as oncotarget. Oncotarget *1*, 651–661.

Zlobec, I., Minoo, P., Terracciano, L., Baker, K., and Lugli, A. (2011). Characterization of the immunological microenvironment of tumour buds and its impact on prognosis in mismatch repair-proficient and -deficient colorectal cancers. Histopathology *59*, 482–495.

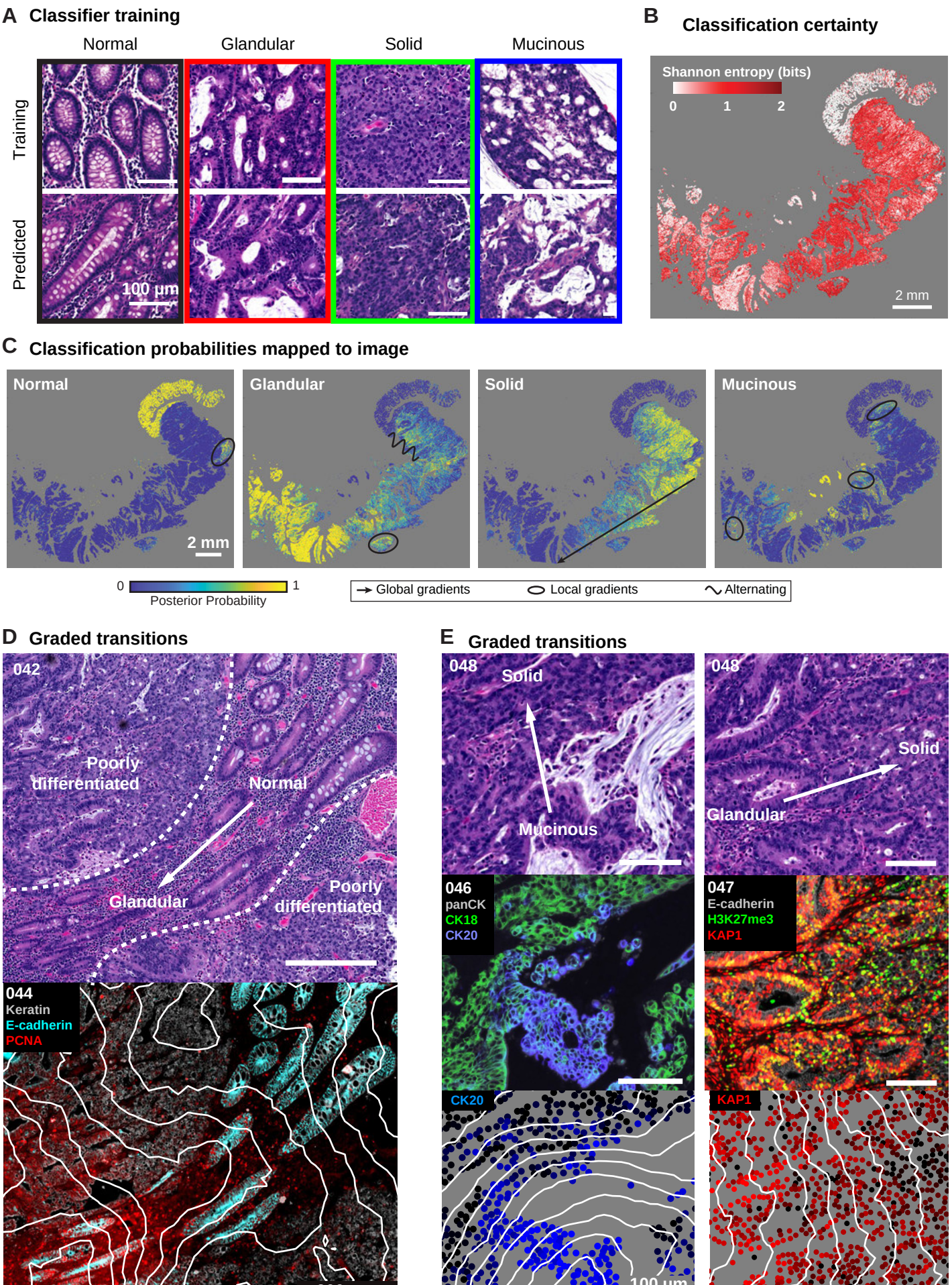
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1



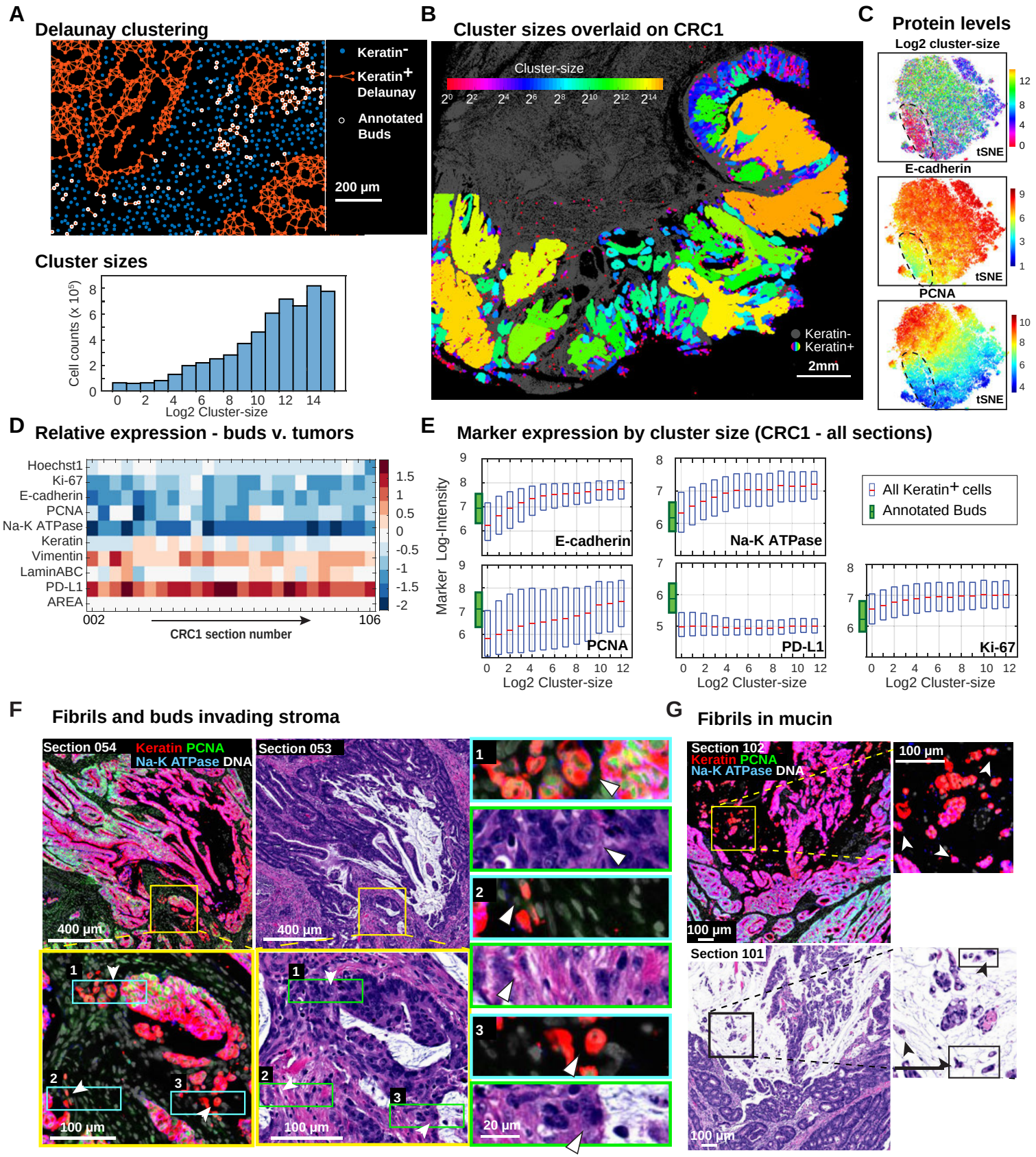
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



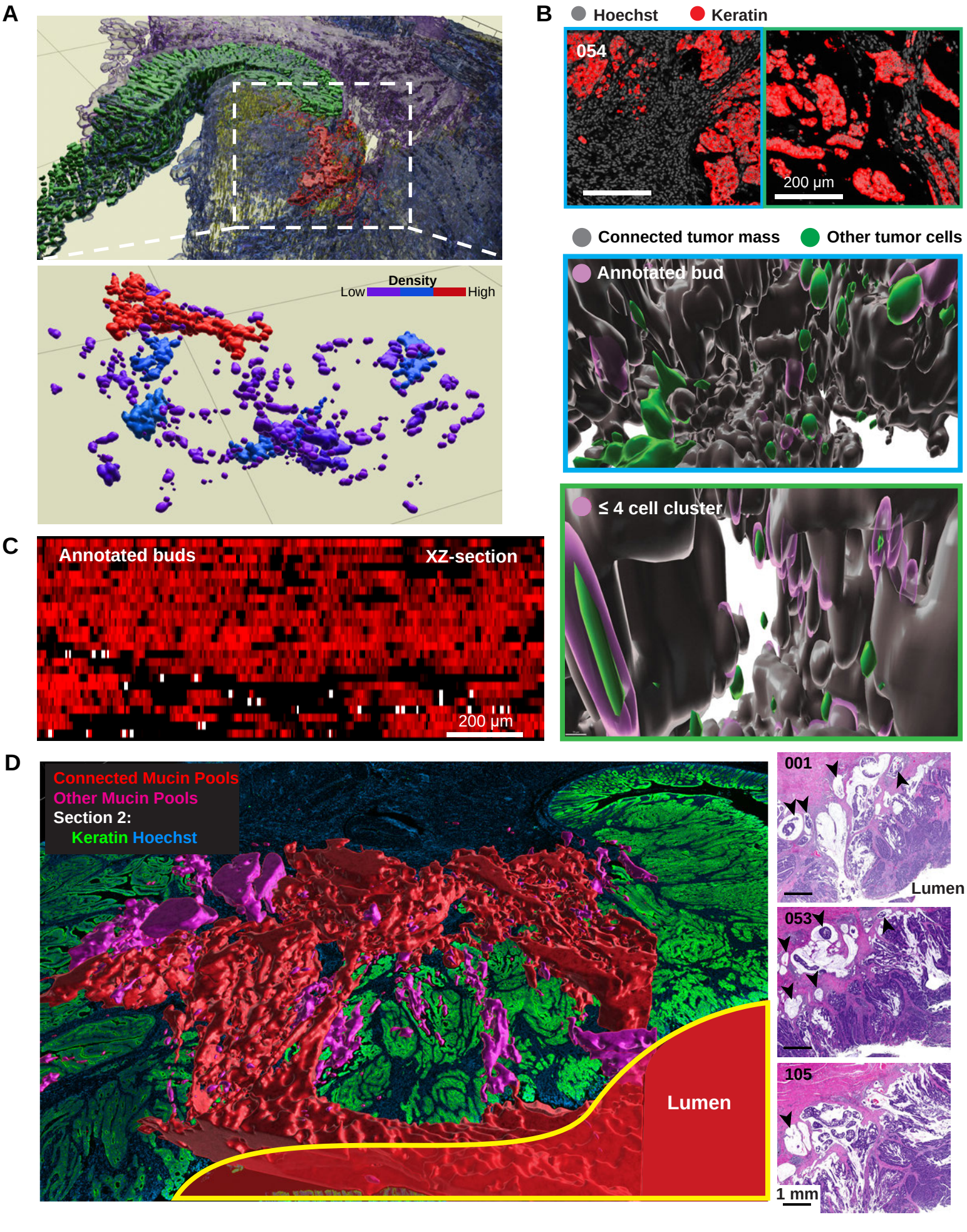
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3



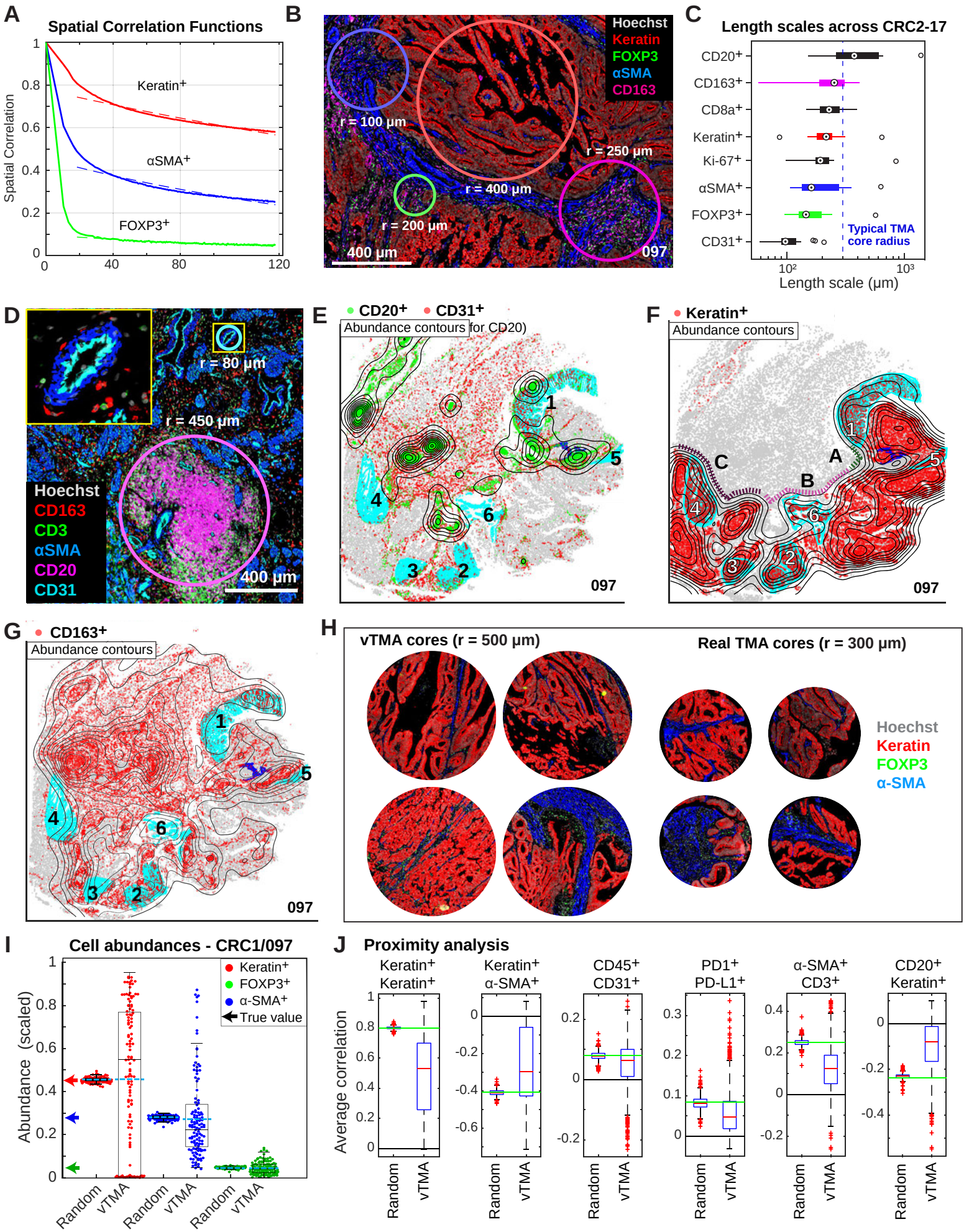
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4



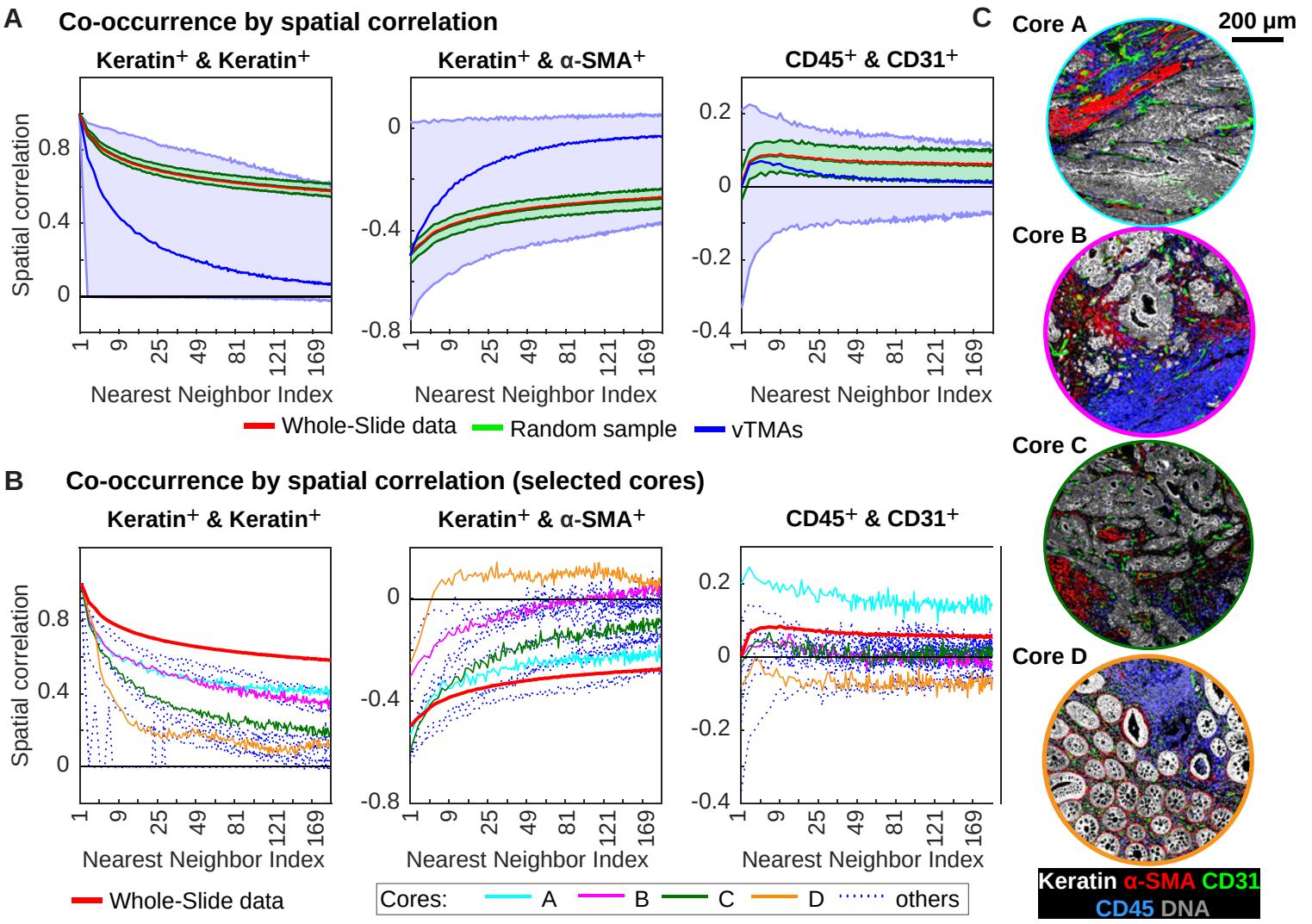
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5



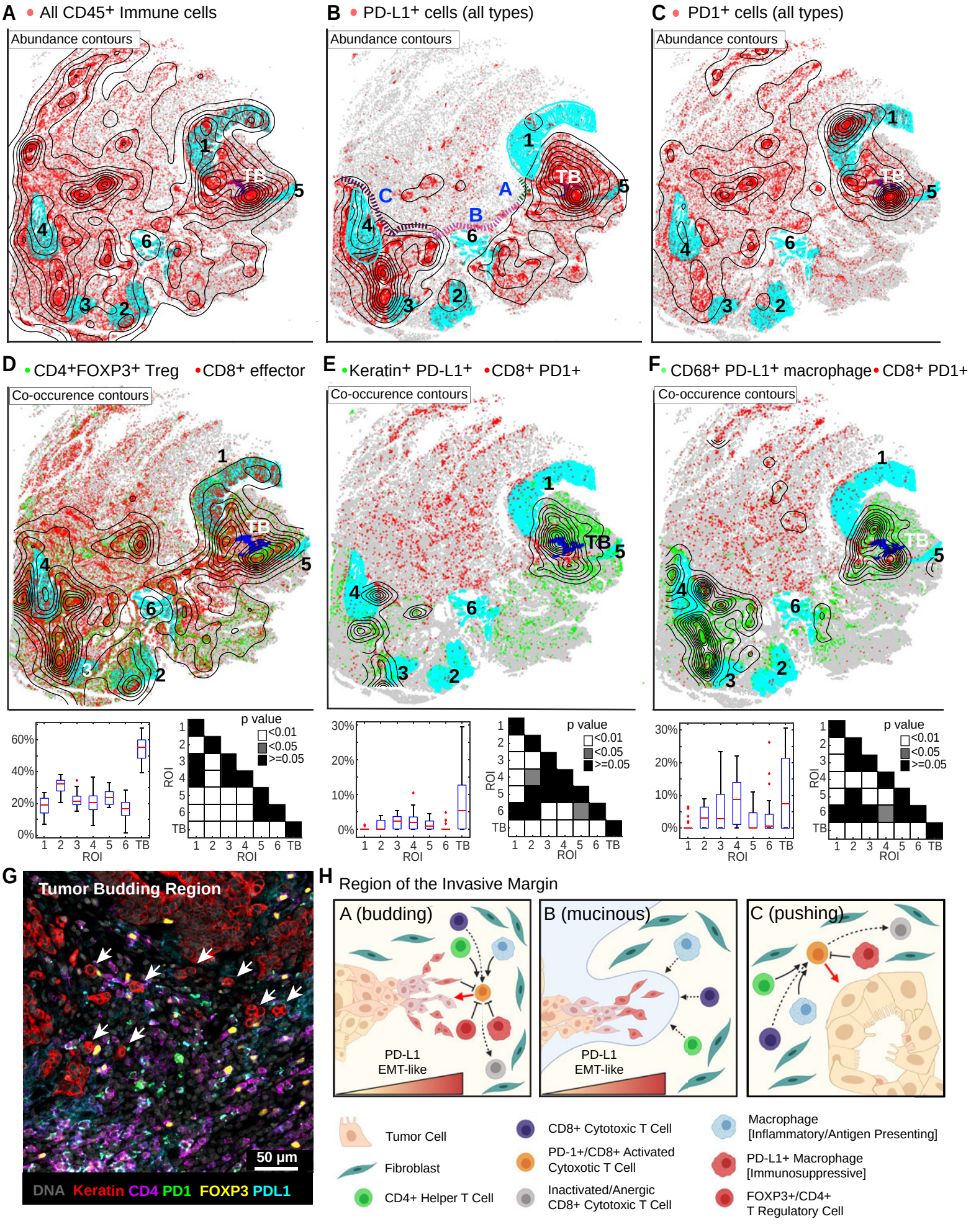
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

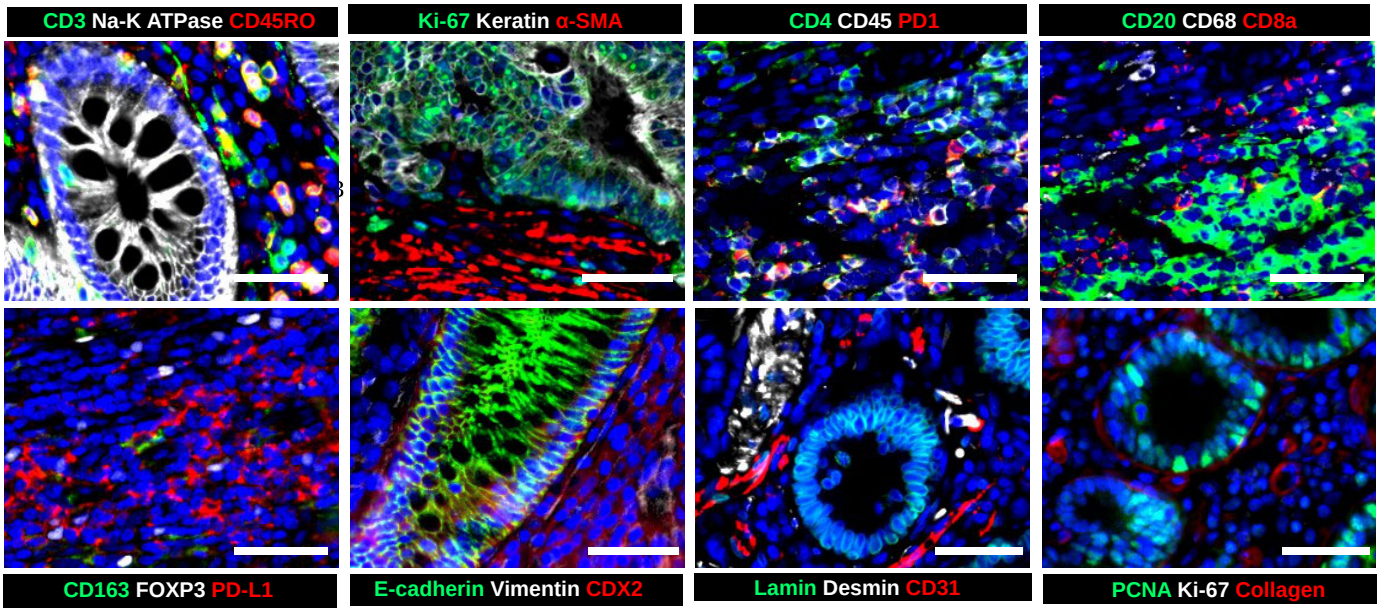
Figure 7



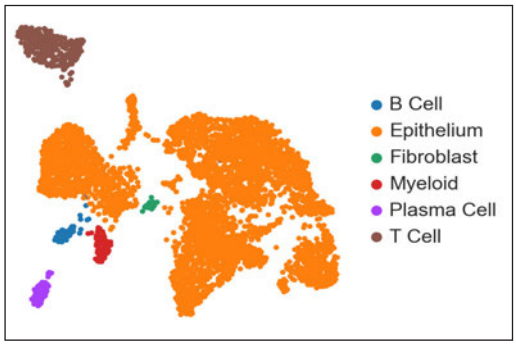
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S1

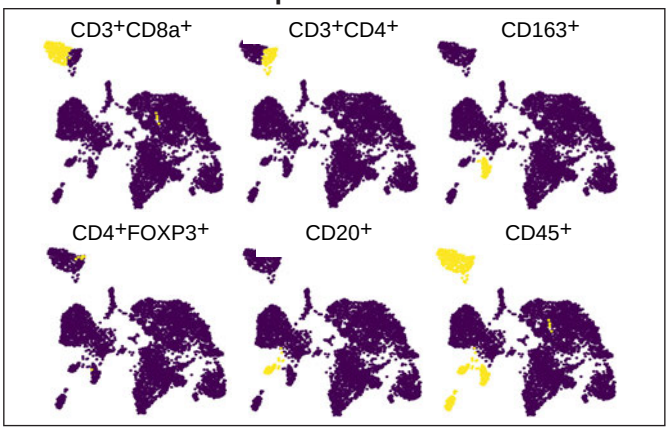
A. Markers and typical morphologies



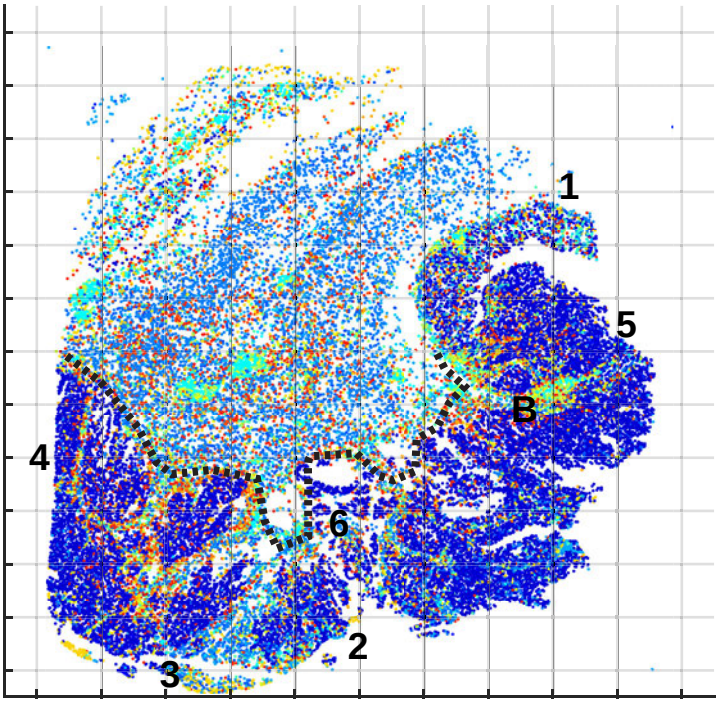
B. UMAP: ScRNASeq



C. UMAP: ScRNASeq



D. Cell type map (CRC1/097)



Tumor

Epithelium

- 1. Tumor/Epi.
- 2. Ki-67+ Tumor/Epi
- 3. PD-L1+ Tumor/Epi

Stroma

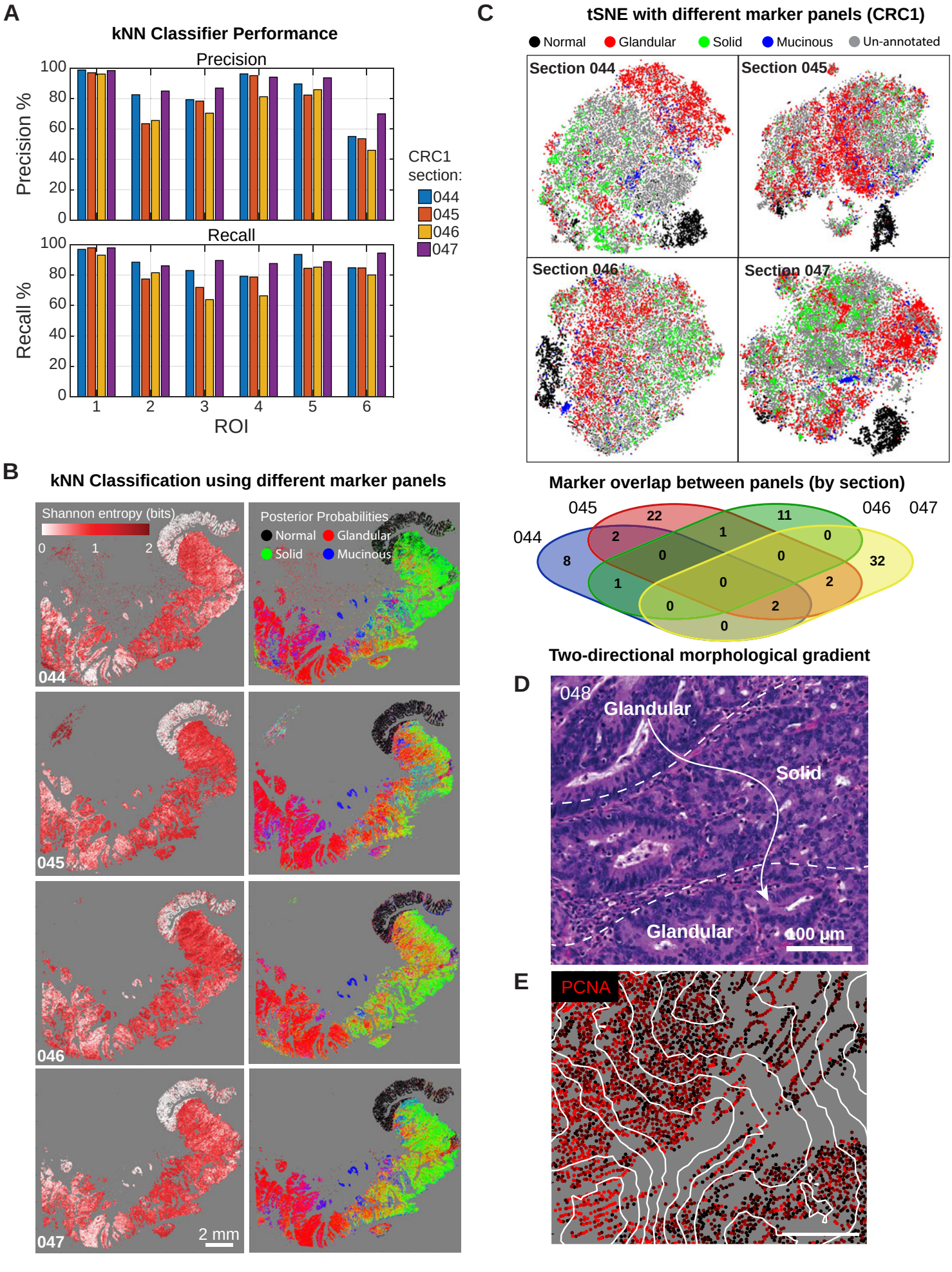
- 4. Endothelial
- 5. Muscle/Fibroblast
- 6. Other

Immune

- 7. Macrophage (I)
- 8. B cell
- 9. DN Lymphocyte
- 10. T helper
- 11. Macrophage (II)
- 12. DP Lymphocyte
- 13. Treg
- 14. PD1+ T helper
- 15. Lymphocyte (III)
- 16. Macrophage (III)
- 17. PD-L1+ Macrophage
- 18. Macrophage (V)
- 19. Tc cell
- 20. PD1+ Tc cell
- 21. PD-L1+ Lymphocyte

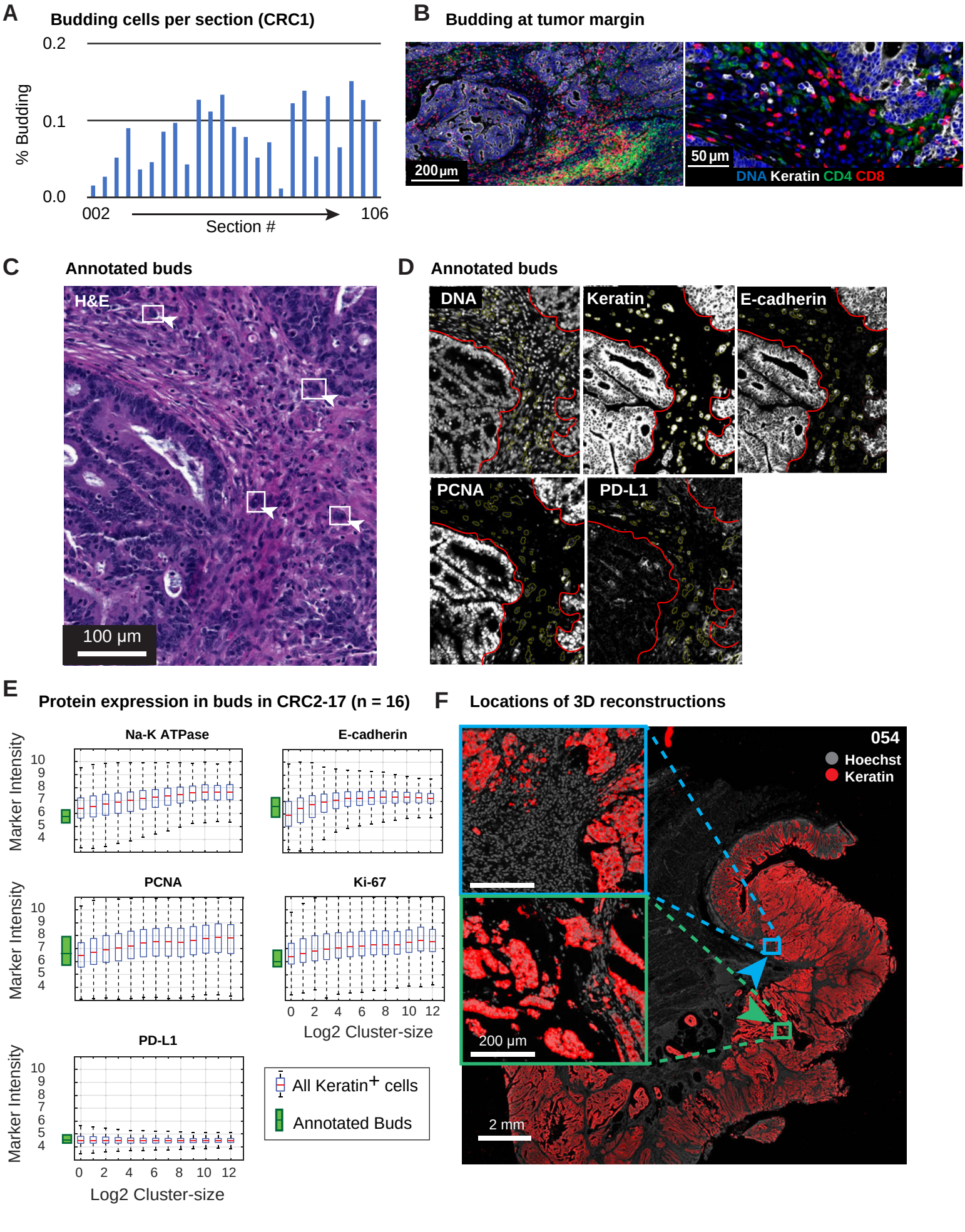
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S2



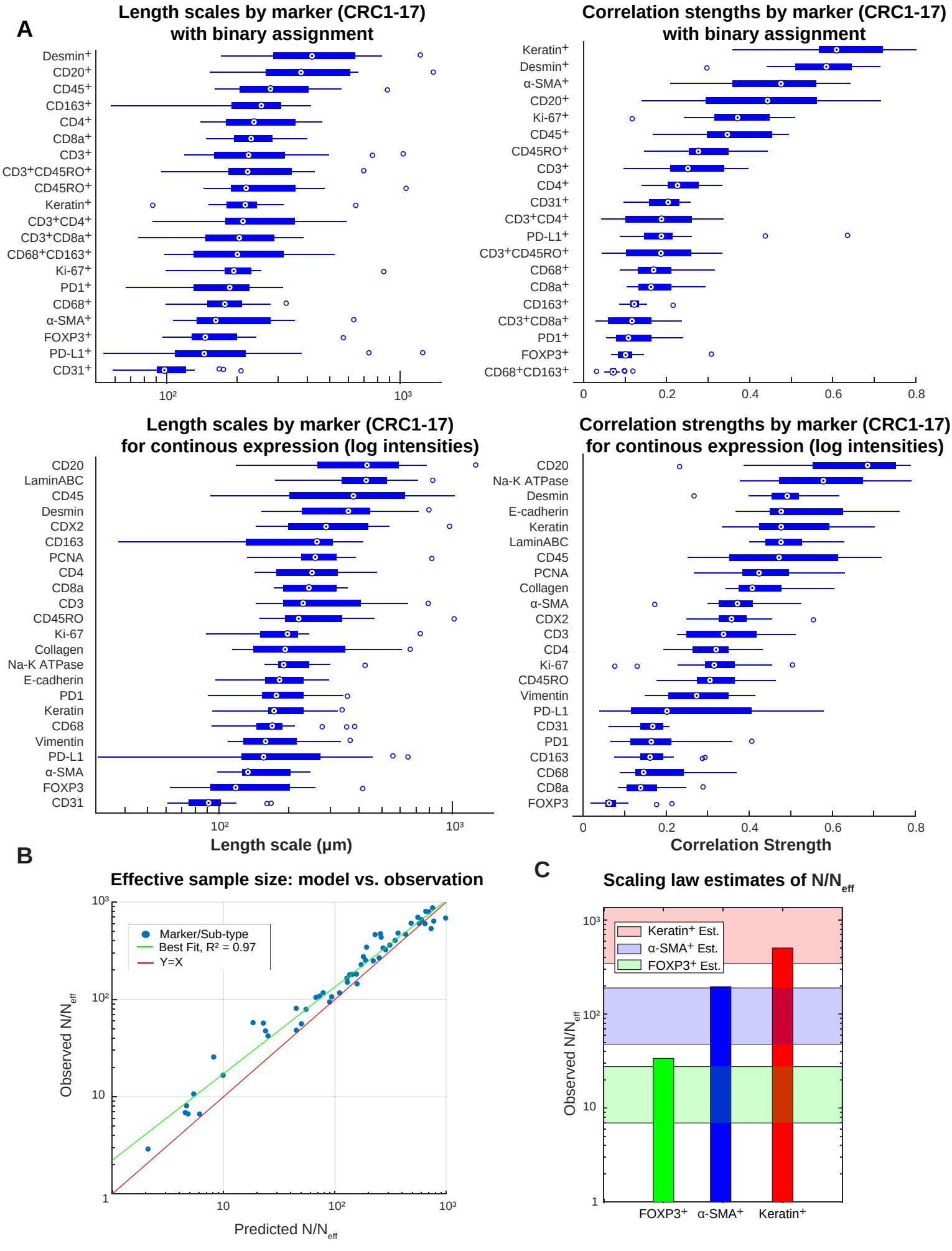
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S3



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

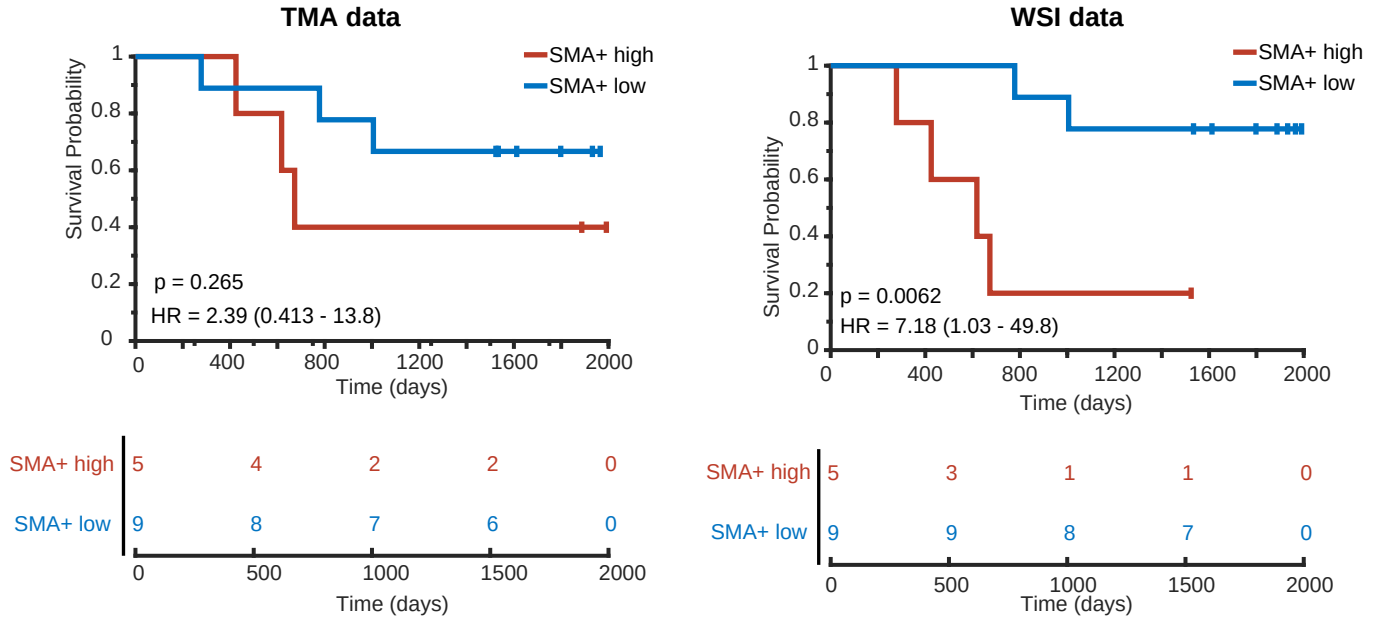
Figure S4



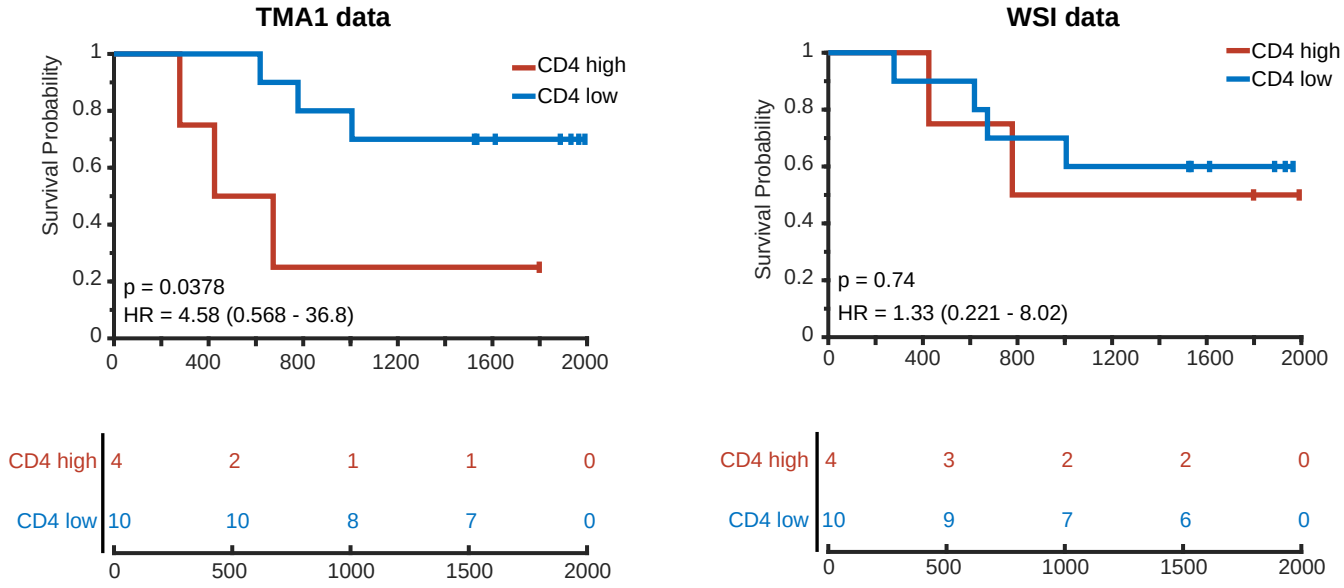
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S5

A. False negative due to reduced power due to use of TMA data (n = 14 patients)

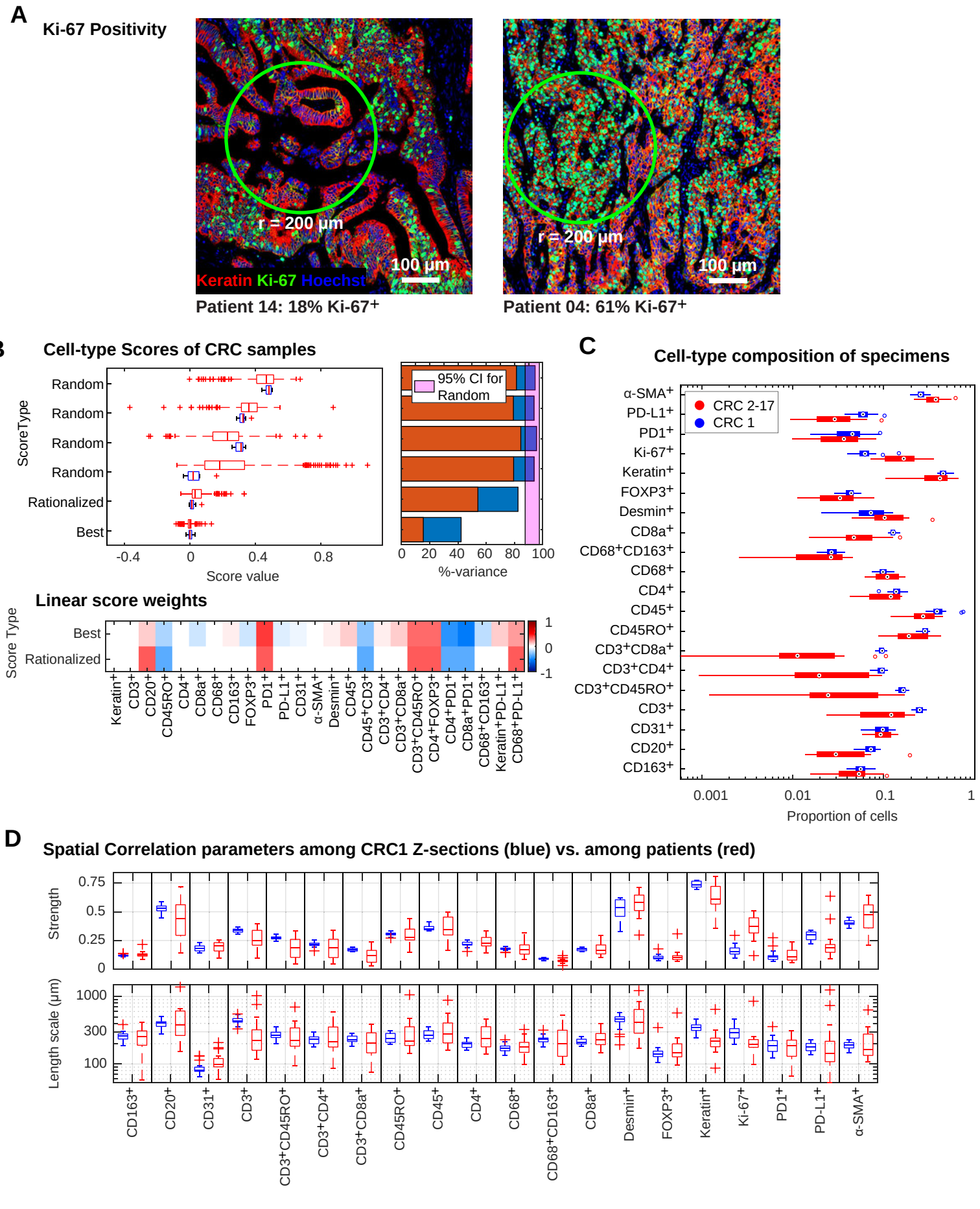


B. False positive due to use of TMA data (n = 14 patients)



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

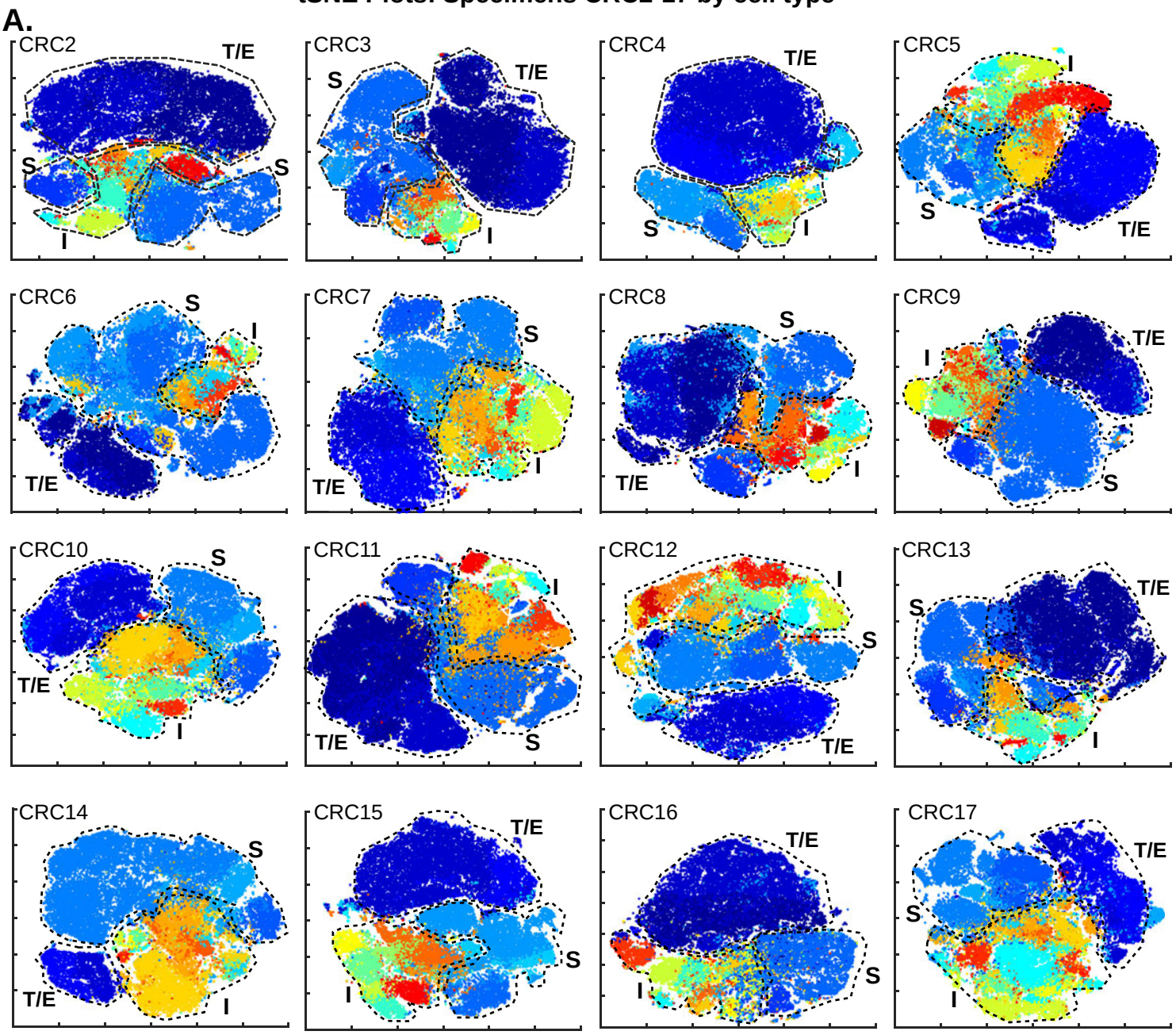
Figure S6

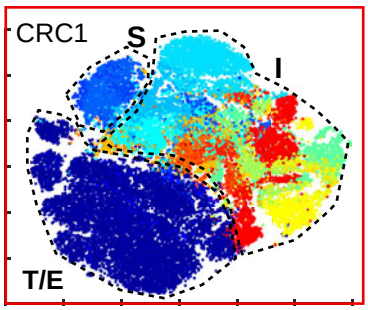


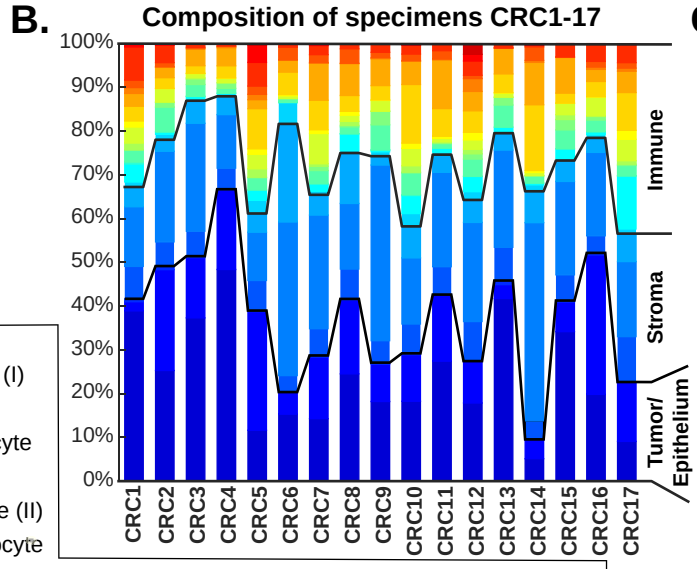
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437984; this version posted April 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

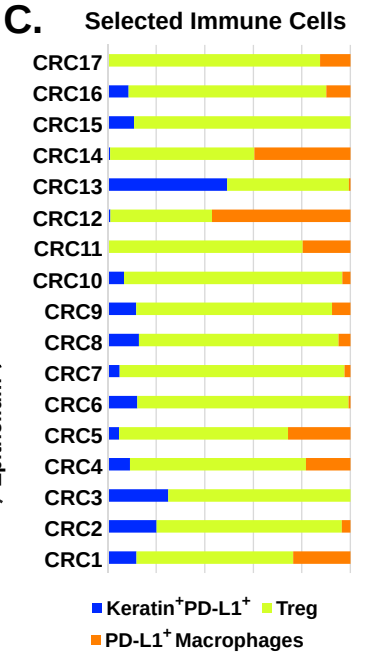
Figure S7

tSNE Plots: Specimens CRC2-17 by cell type









Tumor/ Epithelium (T/E)
- 1. Keratin+
- 2. Keratin+ Ki-67+
- 3. Keratin+ PD-L1+

Stroma (S)
- 4. Endothelial
- 5. Muscle/Fibroblast
- 6. Other

Immune (I)
- 7. Macrophage (I)
- 8. B cell
- 9. DN Lymphocyte
- 10. T helper
- 11. Macrophage (II)
- 12. DP Lymphocyte
- 13. Treg
- 14. PD1+ T helper
- 15. Lymphocyte (III)
- 16. Macrophage (III)
- 17. PD-L1+ Macrophage
- 18. Macrophage (V)
- 19. Tc cell
- 20. PD1+ Tc cell
- 21. PD-L1+ Lymphocyte