A spatial analytics computational and systems biology platform predicts risk of colorectal cancer recurrence and identifies emergent spatial domain networks associated with recurrence

Shikhar Uttam[1, *], Andrew M. Stern[1,2], Samantha Furman[1], Filippo Pullara[1], Daniel Spagnolo[1], Luong Nguyen[1], Albert Gough[1,2], Christopher J. Sevinsky[3], Fiona Ginty[4], D. Lansing Taylor[1,2], S. Chakra Chennubhotla[1, *]

[1] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260

[2] University of Pittsburgh Drug Discovery Institute, University of Pittsburgh, PA 15261

[3] Biochemistry and Bioanalytics Laboratory, GE Global Research Center, Niskayuna, NY 12309

[4] Life Sciences and Molecular Diagnostics Laboratory, GE Global Research Center, Niskayuna, NY 12309

[*]Email: shf28@pitt.edu (Shikhar Uttam); chakracs@pitt.edu (S. Chakra Chennubhotla)

Abstract

An unmet clinical need in solid tumor cancers is the ability to harness the intrinsic spatial information in primary tumors that can be exploited to optimize prognostics, diagnostics and therapeutic strategies for precision medicine. We have developed a transformational spatial analytics (SpAn) computational and systems biology platform that predicts clinical outcomes and captures emergent spatial biology that can potentially inform therapeutic strategies. Here we apply SpAn to primary tumor tissue samples from a cohort of 432 chemo-naïve colorectal cancer (CRC) patients iteratively labeled with a highly multiplexed (hyperplexed) panel of fifty-five fluorescently tagged antibodies. SpAn predicted the 5-year risk of CRC recurrence with a mean area under the ROC curve of 88.5% (SE of 0.1%), significantly better than current state-of-the-art methods. SpAn also inferred the emergent network biology of the tumor spatial domains revealing a synergistic role of known features from CRC consensus molecular subtypes that will enhance precision medicine.

Main

Colorectal Cancer (CRC) is the second most common type of cancer and the third leading cause of cancer-related deaths worldwide.[1] This multi-factorial disease like other carcinomas, develops and progresses through the selection of epithelial clones with the potential to confer malignant phenotypes in the context of a reciprocally coevolving tumor microenvironment (TME) comprising immune and stromal cells.[2–4] CRC patients are staged using the well-established tumor-node-metastases (TNM) classification.[5,6] However, there is significant variability in patient outcomes within each stage. For example, CRC will recur in up to 30% of Stage II patients despite complete

tumor resection, no residual tumor burden and no signs of metastasis.[7] In contrast, more advanced CRC has been known to show stability or indeed even to spontaneously regress.[7,8]

The intrinsic plasticity of the TME underlying this variability in outcome is controlled by complex network biology emerging from the spatial organization of diverse cell types within the TME and their heterogeneous states of activation.[3,9–11] The important role of the TME in CRC progression and recurrence has recently been highlighted by the identification of four consensus molecular subtypes (CMS) [12,13], functional studies defining the critical role of stromal cells in determining overall survival,[14] and the development of Immunoscore®[14] which quantifies tumor-infiltrating T-lymphocytes in different regions of the tumor and associates their infiltration with CRC recurrence.[15,16] However the TME can be further harnessed to significantly improve CRC prognosis through the identification of biomarkers mechanistically linked to disease progression and the development of novel therapeutic strategies.

Deeper understanding of the TME may arise from imaging methods capable of labeling > 7 cellular and tissue components in the same sample (hyperplexed[17] (HxIF) fluorescence and other imaging modalities).[17–21] To fully extract the intrinsic information within each primary tumor we have developed a spatial analytics (SpAn) computational and systems pathology platform applicable to all solid tumors to analyze the spatial relationships throughout TME signaling networks. SpAn constructs a computationally unbiased and clinical outcome-guided statistical model enriched for a subset of TME signaling networks that are naturally selected as dependencies of the corresponding malignant phenotype.[22–25] Herein we describe the implementation of SpAn to predict 5-year risk of CRC recurrence in patients with resected primary tumor that also enables inference of recurrence-specific network biology.

Results

Hyperplexed immunofluorescence imaging of tissue microarrays generates multidimensional spatial proteomics data with sub-cellular resolution

The acquired data were generated using GE Cell DIVE™ previously named MultiOmyx[20] (GE Healthcare, Issaquah, WA) hyperplexed[17] immunofluorescence (HxIF) imaging and image processing workflow instrument. As previously described,[20] Cell DIVE™ can perform hyperplexed imaging of greater than 50 biomarkers via sequentially multiplexed imaging of 2 to 3 biomarkers plus DAPI nuclear counterstain through iterative cycles of label–image–dye-inactivation visualized in Supplementary Fig. S1.[20] (See Methods for more details.) Extensive validation of this approach has demonstrated that a majority of epitopes tested are extremely robust to the dye inactivation process. The biological integrity of the samples were preserved for at least 50 iterative cycles.[20]

In this study we use 55 biomarkers, which include markers for epithelial, immune and stromal cell lineage, subcellular compartments, oncogenes, tumour suppressors, and posttranslational protein modifications indicative of cellular activation states. Their role in CRC is described in detail in Supplementary Table S1. Figure 1a shows the HxIF image stack of a 5 µm thick and 0.6mm wide tissue microarray[26] (TMA) spot from resected primary tumor of a Stage II CRC patient labelled with the 55 biomarkers plus DAPI. Figure 1b highlights a sub-region of this patient TMA spot enabling optimal visualization of the 55 HxIF biomarker images resulting from the iterative label–image–chemical-inactivation cycles.

Cell DIVE™ was employed to generate HxIF image stacks of FFPE tissue microarrays from resected tissue samples from 432 chemo-naïve CRC patients at single cell resolution. This 55-

dimensional spatial profiling of the patient-level tumor microenvironment served as input in our study. The 432 retrospectively acquired chemo-naïve CRC patient cohort included patients in Stage I through Stage III of CRC primary tumor growth between the years of 1993 and 2002 acquired from Clearview Cancer Institute of Huntsville Alabama. As shown in Supplementary Table S2, the median patient age and gender proportions were similar across all stages, with CRC recurring in 65 patients. The outcome distribution of the patients and their clinical attributes across the CRC stages are detailed in Table S2. The use of chemo-naïve (no administration of neoadjuvant or adjuvant therapies for the 5+ years of follow-up) CRC patient cohort provides SpAn the opportunity to interrogate unperturbed primary tumor biology.

SpAn uses HxIF spatial proteomics data to learn recurrence-guided and spatially-informed prognosis of CRC recurrence

SpAn performs a virtual three-level spatial-dissection of the tumor microenvironment, by first explicitly decomposing the TMA into epithelial and stromal regions as detailed in Methods and illustrated in Fig. S2. The cells in the epithelial region are identified using E-cadherin cell-cell adhesion labeling and pan-cytokeratin, with individual epithelial cells segmented using a Na^+K^+ATPase cell membrane marker, ribosomal protein S6 cytoplasmic marker, and DAPI-based nuclear staining. The resulting epithelial spatial domain of the TMA in Fig. 1a is shown in Fig. 1c. The remaining cells are assigned to the stromal domain and are visualized in Fig. 1d. These stromal cells have diverse morphologies.[20] Based on the epithelial and stromal domains, SpAn also identifies a third epithelial-stromal domain, shown in Fig. 1e, to explicitly capture a 100 µm boundary wherein the stroma and malignant epithelial cells interact in close proximity. Together these three intra-tumor spatial domains comprise the virtual three-level spatial dissection of the

tumor microenvironment that forms the basis for the SpAn spatial model overviewed in Fig. 1f and detailed in Fig. 2.

Utilizing expression of the 55 hyperplexed biomarkers, SpAn first computes the corresponding 55 mean intensities and 1485 Kendall rank-correlations as features characterizing each of the three spatial domains (see Fig. 2a). The mean intensity captures the average domain-specific expression profile of each biomarker, while the Kendall rank-correlations[27] measure strength of association between any two biomarkers without presuming linearity. (See Method for details.) Importantly, computation of domain-specific rank-correlations as explicit features for SpAn is used in place of the more typical approach of implicitly incorporating correlations as interactions between covariates (average biomarker expressions) within the prediction model[28]. These explicit features not only detect the association between two biomarkers presumably mediated by intracellular and intercellular networks all within the same spatial domain but also by mediators (e.g., exosomes) derived from another spatial domain. As an example, SpAn finds enrichment of KEGG 'microRNAs in cancer' pathway in the epithelial and epithelial-stromal domains (Figure 5), while concurrently selecting correlation between CD163 and PTEN as a feature in the stromal domain for recurrence prognosis (see Fig. 2c). As has been reported in gastrointestinal cancers, tumor cell derived exosomal miRNAs mediate crosstalk between tumor cells and the stromal microenvironment, and induce polarization of the macrophages to the anti-inflammatory and tumor-supportive M2 state via activation of the PTEN-PI3K signaling cascade under hypoxic conditions resulting in enhanced metastatic capacity.[29,30]

SpAn then uses CRC recurrence-guided learning to determine those specific spatial domain features that constitute the optimal subset for prognosis via model selection based on L1-

penalized Cox proportional hazard regression method (Figure 2b).[31,32] See Methods for details on penalized regression, Fig. S3 for validity of the proportional hazard assumptions, and Figure S4 for determination of threshold for concordance with recurrence outcome. A follow-up analysis of the selected features is performed to test the stability of their contribution to recurrence prognosis through testing the stability of the sign of the corresponding coefficients at the 90% threshold. The final domain-specific features are shown in Fig. 2c with additional details described in Methods.

The coefficients that control the contribution of the selected features to each of the domain-specific models for assessing recurrence outcome were learned under L1 penalization and their values are, therefore, dependent on all 1540 features. To remove this dependence, SpAn relearns each of the three domain-specific model coefficients using L2 penalty in our penalized Cox regression model with only the optimally selected features as input. This L2-regularized learning allows SpAn to estimate optimal contribution of the selected features that are 90% concordant with the recurrence outcome. The resulting domain-specific coefficients are shown in Supplementary Fig. S5. As detailed in Methods, SpAn combines these domain-specific features weighted by their corresponding coefficients into a single recurrence guided spatial domain prognostic model, whose performance is shown in Fig. 3a. The results were obtained by bootstrapping (sampling with replacement) patient data set to generate 500 pairs of independent training and testing sets using stratified sampling that ensured the proportion of patients in whom cancer recurred in each of the five years remained the same in each bootstrap. For each bootstrap, SpAn used the training data for learning and the independent testing data to compute the receiver operating characteristic (ROC) curve. These ROC curves are shown in Fig. 3a along with the mean

ROC curve. The mean area under the curve (AUC) for bootstrapped ROC curves is 88.5% with a standard error of 0.1%, demonstrating the stable performance of SpAn. We also maximized Youden's index[33] to identify the clinically relevant operating point on the ROC curves that minimized the overall misdiagnosis rate. Figure 3b shows the resulting sensitivity and specificity values for all bootstrap runs, with mean values respectively of 80.3% (standard error of 0.4%) and 85.1% (standard error of 0.3%). High specificity limits SpAn from misidentifying no-evidence-of-disease patients as being at high risk of CRC recurrence, while at the same time good sensitivity allows SpAn to not miss high-risk patients. This is emphasized by a high positive likelihood ratio value of 7.2 (standard error of 0.23), which quantifies the large factor by which odds of CRC recurring in a patient go up, when SpAn identifies the patient as at being risk of CRC recurrence. At the same time a small negative likelihood value of 0.22 (standard error of 0.003) quantifies the decrease in odds of CRC recurrence in a patient when SpAn identifies the patient as being at low-risk. Finally, these results are brought together in Fig. 3c, which show the large separation in recurrence-free survival curves of patients identified by SpAn at low- and high-risk of five-year CRC recurrence.

Validating the rationale behind SpAn

The rationale behind our 'virtual-dissection followed by combination of the three specific spatial domains' approach is motivated by the acknowledged active role of the microenvironment and its spatial organization, and the differential role played by the epithelial and stromal domains in tumor growth and recurrence.[2,3,14,34] We tested the validity of this rationale within the context of our data by comparing the performance of SpAn with the null model, which is based on recurrence-guided learning on the spatially undissected patient TMA spot. The learning procedure

for the null model was identical to the domain-specific learning within SpAn. Figure 3d shows the improvement in performance quantified by the AUC of the bootstrapped ROC curves achieved by SpAn over the null model. This improvement is statistically significant with a p-value less than 0.005, allowing us to reject the null model at the 99% confidence level. The improved performance highlights the importance of explicitly modeling the epithelial and stromal spatial domains associated with the TME. Interestingly, beyond supporting our rationale, this comparative test shows that our approach of exploiting both biomarker expressions and their correlations results in absolute performances of both the null model and SpAn that are better than those from models using biomarker expressions alone, without correlations (Fig. S6) such as published state-of-the-art approaches that include Immunoscore®.[15,16]

SpAn is highly predictive of 5-year CRC recurrence for TNM Stages I-III in individual patients

The ability to identify patients in whom CRC will recur, especially for those patients in Stages II and III of tumor progression is highly clinically relevant. Figure 3e shows that SpAn can consistently identify patients in whom risk of CRC recurrence is high for Stages I through III, with mean AUC of bootstrapped ROC curves for the three stages respectively being 82.1%, 89.4% and 88.6%. Standard error of these mean AUC values respectively is 0.4%, 0.2%, and 0.2%, demonstrating the stability of SpAn performance. Although the overall performance across all three stages is highly significant with the potential of improving prognosis, the relative reduction in Stage I performance is a consequence of a small cohort of only ten patients in Stage I with CRC recurrence.

The ability of SpAn to predict risk of recurrence in individual patients from all three Stages, is relevant in the context of administering adjuvant therapy, especially for Stage II patients. Current

guidelines for treating Stage II CRC patients from The National Comprehensive Cancer Network (NCCN),[35] the American Society of Clinical Oncology (ASCO),[36] and the European Society of Medical Oncology (ESMO)[37] do not recommend routine adjuvant chemotherapy for Stage II patients, but do state that it should be considered for sub-population of Stage II patients that are at higher risk and might benefit from being put on adjuvant therapy regimen.[38] The personalized prognostic potential of SpAn implies that we could triage Stage II patient cohorts into low and high-risk groups, with the latter being further considered for therapy. Furthermore, SpAn could help with postoperative surveillance of high-risk Stage II patients with more intensive follow-up regimes.[39]

While 20% to 30% of Stage II CRC patients are at high-risk of recurrence, there are Stage III patients that have good prognoses of stable 5-year recurrence-free survival. SpAn, therefore, could also be used to fine-tune their postoperative surveillance and adjuvant chemotherapy regimens.

Prognostic performance of SpAn remains stable over the 5-year time period

A majority of CRC recurrence occurs in the first five years, with 90% occurring in the first four.[40,41] We, therefore, consider the time-dependent performance[42] of SpAn during the first five-year period. Figure 3f plots the AUC for time-dependent ROC performance. The performance of SpAn in predicting risk of recurrence remains consistent and stable (95% confidence interval shown) with only a small, and gradual reduction in time-dependent AUC values as we move away from the resection and imaging timepoint. This result suggests SpAn captures the critical biological underpinnings of recurrence in the primary tumor. Supplementary Fig. S7 shows the time-dependent AUCs for domain-specific temporal performance of SpAn.

SpAn identifies spatial domain networks as emergent properties that explain the robust ability to distinguish patients in which CRC recurs

Given the highly prognostic performance of SpAn, we took a systems perspective to understand and to explain the underlying network biology responsible for this performance within each of the three spatial domains. For each domain, we quantified the unique associations between biomarkers included in the selected features through partial correlations between every biomarker pair, when controlling for other biomarkers as described in Methods. This approach was performed for all patients. The resulting partial correlation for every biomarker pair was separated into two groups according to no-evidence-of-CRC and CRC-recurrence patient cohorts and the information distance based on Jensen-Shannon divergence[43] was computed between them (see Methods for more details). The resulting domain-specific distance matrices, shown in Figs. 4a-c, define associated graphs with the nodes being the biomarkers and edge weights quantifying the differential change, the information distance, in biomarker association between patients in which CRC recurred and those in which there was no evidence of recurrence. The stronger the weights, the larger the distance and the more significant the differential change in association between the two markers for the two patient cohorts. We defined the graphs generated by the distance matrices thresholded at the 99th percentile as the spatial domain networks that were most significant for CRC recurrence prognosis. Figures 4d-f show the resulting networks for the three spatial domains that reveal the heterogeneous nature of the cell populations and signaling pathways leveraged by SpAn in CRC recurrence prognosis.

The epithelial-stromal domain network is comprised of three dominant sub-networks associated with tumor-invading T lymphocytes,[44] disruption in DNA mismatch repair cellular

process, and the role of cancer associated fibroblasts (CAFs) in the desmoplastic microenvironment as indicated by the strong edge weight between smooth muscle actin (SMA) and collagen IV. CAFs are well known to promote EMT[45] and the differential expression of beta-catenin and phosphorylated-MET in Fig. 4f is also consistent with the epithelial-stromal domain supporting the mesenchymal phenotype.[46] These features have also been identified with those distinguishing consensus molecular subtype (CMS) 4 that is associated with a poor prognosis in comparison to the other 3 subtypes in the transcriptome-based classification.[12,13] Interestingly, the epithelial-stromal spatial domain also reveals the presence of DNA mismatch repair network that has been associated with regulation of T lymphocyte infiltration, a prominent feature of CMS1. Thus, the epithelial-stromal spatial domain associated with recurrence combines two features, where in contrast, each alone is associated with two different CMS subtypes. This theme extends to the epithelial spatial domain in Fig. 4d, where metabolic deregulation, a prominent feature of CMS3, and DNA mismatched repair, a hallmark of CMS1 are evident. The association of these two subnetworks in the epithelial domain has the potential to promote tumor cell growth while escaping immune surveillance. Finally, we observe a prominent tumor associated macrophage (TAM) network in the stromal spatial domain (Figure 4e). TAM polarization towards the M2 phenotype regulated by AKT/PTEN has been associated with poor prognosis in CRC that could result from their immunosuppressive and matrix remodeling phenotypes.[30]

Spatial domain networks reveal domain-specific network biology of CRC-relevant pathways

We used STRING[47] and KEGG[48] databases to identify pathways enriched by biomarkers within each of the spatial domain networks and further corroborate their connections to prominent features in the CMS subtype classification. Figure 5 shows the pathways enriched in each of the three

spatial domains, and further identifies those that are common to a majority of at least two of the three spatial domains. Since their identification is based on the spatial domain networks we computationally identified as significant for CRC recurrence prognosis, these pathways play a differentially important role in prognosis of CRC recurrence.

CMS2 tumors are associated with chromosomal instability pathway and enrichment of genes associated with cell cycle and proliferation. Interestingly, both Pi3k-Akt signaling and cell cycle pathways enriched in our analysis are associated with CMS2 tumor subtype, with almost 60%-70% of CRCs associated with dysregulation of Pi3k-Akt signaling pathways.[49]

Tumors associated with the CMS4 mesenchymal phenotype show upregulated expression of genes involved in epithelial-to-mesenchymal transition along with increased stromal invasion, angiogenesis and transforming growth factor-β (TGF-β) activation.[12,13] Interestingly, proteoglycans in cancer, focal adhesion and microRNAs in cancer pathways enriched in our analysis enable the mesenchymal phenotype. For example, non-coding microRNAs both regulate and are targets of upstream regulators for modulating the epithelial to mesenchymal phenotype by targeting EMT-transcription factors such as ZEB1, ZEB2, or SNAIL.[50] Similarly, the focal adhesion pathway through the integrin family of transmembrane receptors mediates attachment to the extracellular matrix, and when dysregulated promotes cell motility and the mesenchymal phenotype.[51,52] Furthermore, extracellular and cell surface proteoglycans with their interaction with cell surface proteins such as CD44 have been known to promote tumor cell growth and migration.[53,54]

Our analysis suggests that by capturing correlation-based crosstalk between heterocellular signaling pathways, SpAn leverages the interconnections between the subtypes for a high performing CRC recurrence prognosis and reveals a synergistic role of the CMS subtypes in CRC progression and recurrence. We note that the ability of SpAn to leverage these interconnections is due to the spatial-context-preserving sampling of a diverse set of CRC-relevant biomarkers enabled by HxIF imaging.

Interestingly, this network biology paradigm also shows enrichment of pathways specific to a single spatial domain whose oncogenic or tumor suppressive roles in CRC is an active area of research but whose differential role in CRC recurrence has not been widely studied. For example, in the epithelial domain our analysis shows the enrichment of Thyroid hormone signaling pathway that has been associated with a tumor suppressive role in CRC development.[55,56] In contrast, the bacterial invasion pathway, enriched in epithelial-stromal boundary region, has been implicated in the oncogenic role of the colonic microbiome in CRC development.[57,58]

Our analysis also reveals enrichment of certain other pathways, such as the hypoxia-inducible factor 1 (HIF-1), human epidermal growth factor receptor 2 (HER2) and T-cell receptor signaling pathways in the epithelial domain. Hypoxia is typical in many solid tumors in CRC with HIF-1 regulating tumor adaptation to hypoxic stress.[59] Alterations in Her-2 signaling, either through genomic amplification or mutations is tumor promoting, and anti-HER2 therapies for preventing CRC recurrence and are a focus of on-going work.[60] We finally note that MAPK and PI3K-AKT signaling cascades are implicated in many of the above discussed pathways.

Discussion

This study highlights the importance of spatial context of the primary tumor microenvironment in conferring distinct malignant phenotypes such as recurrence in CRC. We show how a computationally unbiased approach can be implemented through statistical modeling of spatially defined domains leading to a highly specific and sensitive platform for prognostic and diagnostic tests, as well as potentially inferring therapeutic strategies (Figure 6). SpAn provides testable hypotheses regarding how the spatial association of common networks could potentially lead to emergent signaling networks conferring malignant phenotypes. For example, an epithelial domain network coupling cell metabolism and DNA repair is consistent with tumor cell growth at the expense of T cell exclusion and functional deficiency[12,13,61] (Figs. 4 and 5). Likewise the hijacking of CAFs to support EMT in the context of diminished immune surveillance in the epithelial-stromal spatial domain[45,62,63] (Figs. 4 and 5) and the PI3K/AKT mediated polarization of TAMs within the stromal domain (Figs. 4 and 5) can conspire to facilitate migration of tumor initiating cells to promote both local and distant recurrence.[64]

SpAn, when used in combination with a hyperplexing imaging platform such as Cell DIVE™ allows mechanistic hypotheses to be tested through iterative probing of the same spatial domains with additional biomarkers inferred by the pathway analyses (Figures 5 and 6). We expect even more specific mechanistic biomarkers to emerge based on the iterative hyperplexed imaging approach that incorporates finer stage-based focus, thereby reducing the total number of biomarkers needed for optimal analyses. We will be pursuing this in subsequent studies. This feature of SpAn combined with a hyperplexing imaging platform will not only allow refinement of its prognostic ability, but since the iterative analysis can be potentially conducted in real time with further advances in the technologies, it may allow the prognosis to be specific to individual

patients. Importantly, by enabling the testing of mechanistic hypotheses in patient samples directly connected to a specific clinical outcome, SpAn may inform therapeutic strategies to prevent the outcome. The next phase of this work will take advantage of sampling multiple regions of primary tumors with larger TMAs and/or whole side sections, exploring other spatial analytics ranging from simple to sophisticated spatial heterogeneity metrics[65] and incorporating a combination of protein and nuclei acid biomarkers.[66]

The present retrospective study provides the foundation for the implementation of SpAn in prospective studies predicting disease outcomes in patients with CRC and other carcinomas. The high specificity and sensitivity of SpAn lies in its ability to unbiasedly identify emergent networks that appear to be closely associated and likely to be mechanistically linked to recurrence. We anticipate that hyperplexed datasets based on multiple imaging modalities will be generated faster and become less expensive as the technology evolves to become a mainstay tool to analyze solid tumors.

Acknowledgements

The work of D.M.S., A.G., D.L.T. and S.C.C. was supported by grant NIH/NCI U01CA204836. S.C.C. was also supported in part by NIH/NHGRI U54HG008540, and UPMC Center for Commercial Applications of Healthcare Data 711077. D.L.T. was supported in part by NIH P30CA047904, and PA DHS 4100054875. D.M.S. was supported in part by NIH NIBIB 5T32EB009403-07. F.G was supported by NIH NCI RO1CA208179.

Author Contributions

S.U., A.M.S., D.L.T. and S.C.C. conceived and designed the study and wrote the manuscript. S. U. and S.C.C. developed the computational platform. S.U. performed the formal analyses. S.U., A.M.S., F.P., D.L.T. and S.C.C developed the systems biology framework. C.J.S and F.G. performed imaging and acquired the data. S.U., C.J.S., S.F., F.P., D.M.S. and L.N. were responsible for data quality control and providing analysis support. A.G., C.J.S. and F.G. helped with technical issues and edited and reviewed the manuscript.

Competing interests

S.C. C. and D.L.T. have ownership interest in SpIntellx, Inc., a computational and systems pathology company. The other authors disclose no potential conflicts of interest.

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68,** 394–424 (2018).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
3. Weber, C. E. & Kuo, P. C. The tumor microenvironment. *Surg. Oncol.* **21,** 172–177 (2012).
4. Tape, C. J. The Heterocellular Emergence of Colorectal Cancer. *Trends in cancer* **3,** 79–88 (2017).
5. Weiser, M. R. AJCC 8th Edition: Colorectal Cancer. *Ann. Surg. Oncol.* **25,** 1454–1455 (2018).
6. Gospodarowicz, M. K., Brierley, J. D. & Wittekind, C. *TNM Classification of Malignant Tumours*. (Wiley, 2017).
7. Mlecnik, B., Bindea, G., Pagès, F. & Galon, J. Tumor immunosurveillance in human cancers. *Cancer Metastasis Rev.* **30,** 5–12
8. Bir, A. S., Fora, A. A., Levea, C. & Fakih, M. G. Spontaneous Regression of Colorectal Cancer Metastatic to Retroperitoneal Lymph Nodes. *Anticancer Res.* **29,** 465–468 (2009).
9. Stanta, G. & Bonin, S. Overview on Clinical Relevance of Intra-Tumor Heterogeneity . *Frontiers in Medicine* **5,** 85 (2018).

10. Carmona-Fontaine, C. *et al.* Metabolic origins of spatial organization in the tumor microenvironment. *Proc. Natl. Acad. Sci.* **114,** 2934 LP-2939 (2017).

11. Marusyk, A. *et al.* Spatial proximity to fibroblasts impacts molecular features and therapeutic sensitivity of breast cancer cells influencing clinical outcomes. *Cancer Res.* canres.1457.2016 (2016). doi:10.1158/0008-5472.CAN-16-1457

12. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21,** 1350–1356 (2015).

13. Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer* **17,** 79 (2017).

14. Tauriello, D. V. F., Calon, A., Lonardo, E. & Batlle, E. Determinants of metastatic competency in colorectal cancer. *Mol. Oncol.* **11,** 97–119 (2017).

15. Pagès, F. *et al.* International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* **391,** 2128–2139 (2018).

16. Jorissen, R. N., Sakthianandeswaren, A. & Sieber, O. M. Immunoscore—has it scored for colon cancer precision medicine? *Ann. Transl. Med.* **6,** S23 (2018).

17. Gough, A. *et al.* High-content analysis with cellular and tissue systems biology:a bridge between cancer cell biology and tissue-based diagnostics. in *The Molecular Basis of Cancer* 369–392 (Elsevier, 2015).

18. Schubert, W. *et al.* Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat. Biotechnol.* **24,** 1270 (2006).

19. Lin, J.-R., Fallahi-Sichani, M. & Sorger, P. K. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nat. Commun.* **6,** 8390 (2015).

20. Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 11982–11987 (2013).

21. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174,** 968–981.e15 (2018).

22. Buchberger, A. R., DeLaney, K., Johnson, J. & Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal. Chem.* **90,** 240–265 (2018).

23. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348,** aaa6090 (2015).

24. Wang, G., Moffitt, J. R. & Zhuang, X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci. Rep.* **8,** 4847 (2018).

25. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11,** 360–361 (2014).

26. Bubendorf, L., Nocito, A., Moch, H. & Sauter, G. Tissue microarray (TMA) technology: miniaturized pathology archives for high-throughput in situ studies. *J. Pathol.* **195,** 72–79 (2001).

27. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30,** 81–93 (1938).

28. Chambers, J. M. & Hastie, T. J. Statistical Models. in *Statistical Models in S* 13–44 (1992).

29. Wang, X. *et al.* Hypoxic Tumor-Derived Exosomal miR-301a Mediates M2 Macrophage Polarization via PTEN/PI3Kγ to Promote Pancreatic Cancer Metastasis. *Cancer Res.* **78,** 4586 LP-4598 (2018).

30. Maia, J., Caja, S., Strano Moraes, M. C., Couto, N. & Costa-Silva, B. Exosome-Based Cell-Cell Communication in the Tumor Microenvironment . *Frontiers in Cell and Developmental Biology* **6,** 18 (2018).

31. Goeman, J. J. L1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical J.* **52,** 70–84 (2010).

32. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **39,** 1–13 (2011).

33. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3,** 32–35 (1950).

34. Peddareddigari, V. G., Wang, D. & Dubois, R. N. The tumor microenvironment in colorectal carcinogenesis. *Cancer Microenviron.* **3,** 149–166 (2010).

35. Benson, A. B. *et al.* Colon Cancer, Version 1.2017, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw. J Natl Compr Canc Netw* **15,** 370–398 (2017).

36. Benson, A. B. *et al.* American Society of Clinical Oncology Recommendations on Adjuvant Chemotherapy for Stage II Colon Cancer. *J. Clin. Oncol.* **22,** 3408–3419 (2004).

37. Group, on behalf of the E. G. W. *et al.* Early colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Ann. Oncol.* **24,** vi64-vi72 (2013).

38. Varghese, A. Chemotherapy for Stage II Colon Cancer. *Clin. Colon Rectal Surg.* **28,** 256–261 (2015).

39. Gan, S., Wilson, K. & Hollington, P. Surveillance of patients following surgery with curative intent for colorectal cancer. *World J. Gastroenterol.* **13,** 3816–3823 (2007).

40. Ryuk, J. P. *et al.* Predictive factors and the prognosis of recurrence of colorectal cancer within 2 years after curative resection. *Ann. Surg. Treat. Res.* **86,** 143–151 (2014).

41. Hammond, K. & Margolin, D. A. The role of postoperative surveillance in colorectal cancer. *Clin. Colon Rectal Surg.* **20,** 249–254 (2007).

42. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* **56,** 337–344 (2000).

43. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37,** 145–151 (1991).

44. Sevinsky, C. *et al.* Abstract 1467: Multiplexed immunofluorescence quantitation and validation of multiple immune cell types in colon cancer epithelium and stroma. *Cancer Res.* **76,** 1467 LP-1467 (2016).

45. Tommelein, J. *et al.* Cancer-associated fibroblasts connect metastasis-promoting communication in colorectal cancer. *Front. Oncol.* **5,** 63 (2015).

46. Bradley, C. A. *et al.* Transcriptional upregulation of c-MET is associated with invasion and tumor budding in colorectal cancer. *Oncotarget* **7,** 78932–78945 (2016).

47. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47,** D607–D613 (2019).

48. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47,** D590–D595 (2019).

49. Colakoglu, T. *et al.* Clinicopathological significance of PTEN loss and the phosphoinositide 3-kinase/Akt pathway in sporadic colorectal neoplasms: is PTEN loss predictor of local recurrence? *Am. J. Surg.* **195,** 719–725 (2008).

50. Vu, T. & Datta, P. K. Regulation of EMT in Colorectal Cancer: A Culprit in Metastasis. *Cancers (Basel).* **9,** 171 (2017).

51. Bonnans, C., Chou, J. & Werb, Z. Remodelling the extracellular matrix in development and disease. *Nat. Rev. Mol. Cell Biol.* **15,** 786–801 (2014).

52. Morgan, M. R., Humphries, M. J. & Bass, M. D. Synergistic control of cell adhesion by integrins and syndecans. *Nat. Rev. Mol. Cell Biol.* **8,** 957–969 (2007).

53. Chen, C., Zhao, S., Karnad, A. & Freeman, J. W. The biology and role of CD44 in cancer progression: therapeutic implications. *J. Hematol. Oncol.* **11,** 64 (2018).

54. Theocharis, A. D., Skandalis, S. S., Tzanakakis, G. N. & Karamanos, N. K. Proteoglycans in health and disease: novel roles for proteoglycans in malignancy and their pharmacological targeting. *FEBS J.* **277,** 3904–3923 (2010).

55. Krashin, E., Piekiełko-Witkowska, A., Ellis, M. & Ashur-Fabian, O. Thyroid Hormones and Cancer: A Comprehensive Review of Preclinical and Clinical Studies . *Frontiers in Endocrinology* **10,** 59 (2019).

56. Brown, A. R., Simmen, R. C. M. & Simmen, F. A. The role of thyroid hormone signaling in the prevention of digestive system cancers. *Int. J. Mol. Sci.* **14,** 16240–16257 (2013).

57. Iftekhar, A., Sperlich, A., Janssen, K.-P. & Sigal, M. Microbiome and Diseases: Colorectal Cancer BT - The Gut Microbiome in Health and Disease. in (ed. Haller, D.) 231–249 (Springer International Publishing, 2018). doi:10.1007/978-3-319-90545-7_15

58. Dejea, C., Wick, E. & Sears, C. L. Bacterial oncogenesis in the colon. *Future Microbiol.* **8,** 445–460 (2013).

59. Vaupel, P. The Role of Hypoxia-Induced Factors in Tumor Progression. *Oncol.* **9,** 10–17 (2004).

60. Ross, J. S. *et al.* Targeting HER2 in colorectal cancer: The landscape of amplification and short variant mutations in ERBB2 and ERBB3. *Cancer* **124,** 1358–1373 (2018).

61. Park, J. H. *et al.* Mismatch repair status in patients with primary operable colorectal cancer: associations with the local and systemic tumour environment. *Br. J. Cancer* **114,** 562–570 (2016).

62. Chen, X. & Song, E. Turning foes to friends: targeting cancer-associated fibroblasts. *Nat. Rev. Drug Discov.* **18,** 99–115 (2019).

63. Haviv, I., Polyak, K., Qiu, W., Hu, M. & Campbell, I. Origin of carcinoma associated fibroblasts. *Cell Cycle* **8,** 589–595 (2009).

64. Vergadi, E., Ieronymaki, E., Lyroni, K., Vaporidi, K. & Tsatsanis, C. Akt Signaling Pathway in Macrophage Activation and M1/M2 Polarization. *J. Immunol.* **198,** 1006 LP-1014 (2017).

65. Spagnolo, D. *et al.* Pointwise mutual information quantifies intratumor heterogeneity in tissue sections labeled with multiple fluorescent biomarkers. *J. Pathol. Inform.* **7,** 47 (2016).

66. Janiszewska, M. *et al.* In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. *Nat. Genet.* **47,** 1212–1219 (2015).

Methods

Cell DIVE-based hyperplexed imaging (HxIF) of tissue microarrays (TMA). The 55 biomarkers plus DAPI nuclear counterstain included in this study are described in Supplementary Table S1. HxIF imaging of a TMA slide was performed using sequentially multiplexed labeling and imaging of 2 to 3 biomarkers along with DAPI counterstain through a label–image–chemical-inactivation iterative cycle visualized in Supplementary Fig. S1, and previously described[1] in detail in the supporting information therein. Broadly, the supporting information details the hyperplexed immunofluorescent workflow with information on iterative cycles of antibody labeling of single 5 µm formalin-fixed and paraffin embedded tissue sections and TMA slides, autofluorescence removal, imaging, and dye inactivation in tissue. All samples were stained and imaged in a single batch for 2 to 3 biomarkers and DAPI at a time.

Image processing and single cell analysis. DAPI based nuclear staining was used to register and align sequentially labeled and imaged TMA spots prior to downstream image analysis steps.[1,2] Autofluorescence was removed from the stained images,[1,3,4] which were then segmented into epithelial and stromal regions (Fig. S2), differentiated by epithelial E-cadherin staining. This was followed by segmentation of individual cells in both the epithelium and stroma. Epithelial cells were segmented using Na^+K^+ATPase based cell-membrane staining to delineate cell borders and membrane regions, the cytoplasmic ribosomal protein S6 for cytoplasm identification, and DAPI stain for nuclear regions.Protein expression level and standard deviation were subsequently quantified in each cell.

Quality checks and data normalization. Following single cell segmentation, several data pre-processing steps were conducted. These included cell filtering, spot exclusion, log2 transformation

and slide to slide normalization. Cells were included for downstream analysis if their size was greater than 10 pixels at 20X magnification. The hyperplexing process can result in the tissue being damaged, folded or lost. Image registration issues can also result in poor-quality cell data. Therefore, a tissue quality index based on the correlation of that image with DAPI was calculated for each cell for each round. Only those cells whose quality index equals to or greater than 0.9 (meaning that at least 90% of the cells overlapped with DAPI) were included. All the slides for all the biomarkers were adjusted to a common exposure time per channel. The data were then log2 transformed. A median normalization that equalizes the medians of all the slides was performed to remove slide to slide non-biological variability.

SpAn Input Features. For each of the epithelial, stromal and epithelial-stromal spatial domains, SpAn used $M = 1540$ domain-specific biomarker feature vector $\boldsymbol{f}$ as input. This input feature vector comprised of (1) mean intensity value of 55 biomarkers averaged across all cells within the spatial domain, and (2) 1485 (= 55*54/2) Kendall rank-correlations between all 55 biomarker pairs. Kendall rank-correlation was chosen as the correlation metric because it is a non-parametric measure of association between two biomarkers. Moreover, its use of concordant and discordant pairs of rank-ordered biomarker expression for computing correlation coefficients allows it to robustly capture biomarker associations in presence of measurement noise and small sample size. Rank-correlation for each pair of biomarkers was computed for each spatial-domain from all cells across the spatial-domain expressing the biomarkers. This approach is distinctly different from prediction models that typically consider correlations via interactions, implicit within the models, between mean biomarker intensity expressions – with the biomarker expressions being the only covariates of the model.[5] We emphasize that we did *not* compute correlations through mean

intensity biomarker expression across the spatial-domain, but instead used biomarker expressions across individual cells of the spatial domain to explicitly compute domain-specific rank-correlation values between every pair of biomarkers to form the SpAn correlation feature set.

Before computing these two sets of features, SpAn analysis workflow included an initial intensity threshold step to ensure feature robustness. Specifically, we computed intensity-based distribution of cell-level biomarker expression separately for every biomarker across each patient TMA spot. Only intensities above the 85th percentile on this distribution were considered as biomarker expression and included in computing the intensity features. This focus on the right-tail of the intensity distribution was deliberately conservative, and although it might have potentially excluded low-intensity biomarker expression, it minimized inclusion of false-positive expressions into the analysis.

Penalized Cox proportional hazard regression. For each spatial-domain, SpAn implemented the Cox proportional hazard model via the partial likelihood function $L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{e^{\left(f_{i_k}^{\mathrm{T}}\boldsymbol{\beta}\right)}}{\sum_{i\in R_k} e^{\left(f_i^{\mathrm{T}}\boldsymbol{\beta}\right)}}$ with the penalty given by $P_{\lambda,\alpha}(\boldsymbol{\beta}) = \sum_{m=1}^{M} \lambda \left(\alpha|\beta_m| + \frac{1}{2}(1-\alpha)\beta_m^2\right)$, and $\alpha = \{0,1\}$. (The validity of using the Cox proportional hazard regression model is demonstrated in Fig. S3.) Given feature vector $\boldsymbol{f}$ as input, the partial likelihood $L(\boldsymbol{\beta})$ quantifies the conditional probability of observing CRC recur in a patient at time t_k (proportional to the numerator $e^{\left(f_{i_k}^{\mathrm{T}}\boldsymbol{\beta}\right)}$ of $L(\boldsymbol{\beta})$), given the risk that a patient will recur from the set R_k of patients at risk at time t_k (proportional to the denominator $\sum_{i\in R_k} e^{(f_i^{\mathrm{T}}\boldsymbol{\beta})}$), over all time $t_k, k = 1, \ldots, K$, as quantified by the product over time index k. The partial likelihood is a function of the coefficient vector $\boldsymbol{\beta}$, whose penalized estimate

is then used to compute the proportional hazard ratio $HR = e^{(\boldsymbol{f}^{\mathrm{T}}\boldsymbol{\beta})}$. SpAn computed this estimate via a two-step process: (1) Model selection based on L1-penalized (LASSO) Cox regression where $\boldsymbol{\alpha}$ is set to 1 in penalty $P_{\lambda,\alpha}(\boldsymbol{\beta})$, followed by (2) L2-regularized $\boldsymbol{\beta}$ estimation, with $\boldsymbol{\alpha}$ set to 0 in the penalty term.[6,7] As part of model selection, SpAn selected the model that maximally explained the deviance of the null model (model with only intercept and no predictive features) from the biased but perfect model with an exact fit to the recurrence data. Due to the LASSO penalty, the resulting optimal model forces the coefficients of vector $\boldsymbol{\beta}$ that correspond to features that play a minimal role in predicting risk of recurrence to zero. The subset of features with non-zero coefficients defines the selected model. However, to ensure that this subset identifies a stable set of features, SpAn repeated model selection over 500 bootstraps, and included only those features that were stably concordant at the 90% level with the recurrence outcome. The rationale for 90% concordance is discussed in supplementary Fig. S4. The values of coefficients corresponding to the features stably selected using L1-penalized Cox regression, however, are conditioned on the 1540 input features. To remove this dependence, the second step of SpAn re-estimated the beta coefficients corresponding to the selected features only, by maximizing the partial likelihood function with L2-regularization as the penalty. The resulting beta-coefficients were passed through a final stability check, where the stability of the coefficient sign in 90% of the 500 bootstrap runs was tested, and only features that passed this threshold (Fig. 2c) were included in the spatial domain model. SpAn performed this process independently for each of the three spatial-domains resulting in domain-specific recurrence-guided features (Fig. 2c) and their coefficients (Fig. S5).

Spatial-Model. Each of the three recurrence-guided domain-specific models defined a hazard risk given by $e^{\left(\boldsymbol{f}_{\boldsymbol{epithelial}}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{epithelial}}\right)}$, $e^{(\boldsymbol{f}_{\boldsymbol{stromal}}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{stromal}})}$ and $e^{\left(\boldsymbol{f}_{\boldsymbol{epi-stromal}}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{epi-stromal}}\right)}$ for the epithelial,

stromal, and epithelial-stromal domains respectively. SpAn then defined the final overall risk of recurrence model as $\prod_{s \in S} e^{(f_s^{\mathrm{T}} \boldsymbol{\beta}_s)}$, with $S = \{epithelial, stromal, epi - stromal\}$.

Partial correlations and spatial-domain networks. For each spatial domain, the selected features identified a set of biomarkers specific to predicting risk of CRC recurrence. SpAn used them to define a space of biomarkers within which *partial* correlations between every pair was computed by controlling for confounding effect of biomarkers not defining the pair.[8] The process performed on each patient was as follows: Let the set of biomarkers identified by the selected features be N (<= 55). Using the already computed Kendall rank-correlations between the 55 biomarkers, an $N \times N$ correlation matrix $\boldsymbol{C}$ corresponding to the N biomarkers was constructed, with small shrinkage-based modification to guarantee its positive definiteness, and therefore, its invertibility. Next, the $N \times N$ precision matrix $\mathbf{P}$ was computed by inverting $\boldsymbol{C}$. The partial correlation between any two biomarkers bm_i and bm_j within the set identified by the selected features, was then computed using $\rho_{bm_i,bm_j} = \frac{-p_{bm_i,bm_j}}{\sqrt{p_{bm_i,bm_i} \cdot p_{bm_j,bm_j}}}$, where p_{bm_i,bm_j} is the $(i,j)^{\mathrm{th}}$ element of the precision matrix $\mathbf{P}$. The partial correlations were performed for all patients and were then separated into two groups corresponding to patients with no evidence of disease and those patients in which CRC recurred. Probability distributions of the partial correlations – on the compact set $[-1,1]$ – within each group were computed and the information distance between these two distributions was computed using the Jensen-Shannon divergence. This information distance defines the differential change in the association – partial correlation – between biomarkers bm_i and bm_j in the two patient cohorts. Greater the distance, larger the differential change. Repeating this process for all $N(N-1)/2$ biomarker pairs resulted in the information

distance matrices shown in Figs. 4a-c for the three spatial domains. These information distance matrices were thresholded at the 99th percentile resulting in the computationally inferred spatial-domain networks shown in Figs. 4d-f. The high percentile was chosen to ensure that most discriminative networks are captured.

Enrichment analysis. The STRING database[9,10] was queried with the set of proteins identified by the spatial domain networks generated by thresholding the domain-specific information distance matrices at the 99th percentile, to perform functional enrichment analysis using Fisher's exact test with multiple testing correction.

References

1. Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 11982–11987 (2013).

2. Bello, M., Can, A. & Tao, X. Accurate registration and failure detection in tissue micro array images. in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 368–371 (2008). doi:10.1109/ISBI.2008.4541009

3. Woolfe, F., Gerdes, M., Bello, M., Tao, X. & Can, A. Autofluorescence Removal by Non-Negative Matrix Factorization. *IEEE Trans. Image Process.* **20,** 1085–1093 (2011).

4. Pang, Z., Laplante, N. E. & Filkins, R. J. Dark pixel intensity determination and its applications in normalizing different exposure time and autofluorescence removal. *J. Microsc.* **246,** 1–10 (2012).

5. Chambers, J. M. & Hastie, T. J. Statistical Models. in *Statistical Models in S* 13–44 (1992).

6. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** 1–22 (2010).

7. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **39,** 1–13 (2011).

8. Whittaker, J. *Graphical models in applied multivariate statistics*. (Wiley, 1990).

9. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).

10. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47,** D607–D613 (2019).

Figure 1: Spatial analytics (SpAn) platform based on hyperplexed immunofluorescence imaging for predicting risk of 5-year CRC recurrence. (a) Hyperplexed image-stack of a TMA spot generated by iteratively multiplexed (Fig. S1) HxIF imaging using the Cell DIVE platform.[20] **(b)** Close-up view of a TMA region in **(a)** outlined in white (approx. 110µm by 110µm), labelled with 55 biomarkers (plus DAPI nuclear counterstain) that include epithelial, immune and stromal cell lineage, subcellular compartments, oncogenes, tumour suppressors, and posttranslational protein modifications described in detail in supplementary Table S1. Hyperplexed imaging is implemented via iterative label–image–dye-inactivation immunofluorescence cycle (see Methods and Fig. S1). **(c-e)** Dissection of the TMA spot into three spatial-domains (epithelial, stromal, and epithelial-stromal domains) identified and segmented using structural biomarkers (see Methods and Fig. S2). **(f)** For each of the three spatial domains both expressions of the 55 biomarkers and their Kendall rank-correlations both within and across the cells together defined the domain-specific features. L1-norm based penalized Cox regression was used for model selection, while L2-penalty was used for final model parameter (coefficients) estimation. The stability of the model was tested at the 90% concordance level, and the parameters were reevaluated for final construction of the SpAn spatial model.

Figure 2: **SpAn domain-specific feature selection. (a)** Individual mean protein-expression intensity profile depicted as a vector and pairwise Kendall rank-correlations between protein expressions – visualized as a matrix for each of the three spatial domains. The protein expressions are shown in log scale. To prevent inclusion of false-positive protein expression, only intensities above the 85th percentile were considered as expressions and used to compute the correlations. (See Methods.) **(b)** Features, including both expressions and correlations, selected by SpAn based on L1-penalized Cox regression used for model selection. **(c)** Stability analysis of the selected features, with only those features from the ones selected in **(b)** that maintain their sign in 90% of the 500 bootstrap runs are included as input into the SpAn spatial model. These features are visualized in black in the three bar graphs. Features selected in **(b)** that did not meet this criterion are shown in red.

Figure 3. Performance of SpAn platform for predicting risk of 5-year CRC recurrence. (a) SpAn receiver operating characteristics (ROC) curves for predicting risk of 5-year CRC recurrence in patients with resected CRC primary tumor. The plot shows ROC curves for 500 bootstrap runs with independent training and validation sets. Mean area under the ROC curve is 88.5% with a standard error of 0.1%. **(b)** Boxplots of the sensitivity and specificity values for the same 500 bootstrap runs as in **(a)** with the operating point on the ROCs chosen by minimizing overall misdiagnosis rate. The mean sensitivity and specificity values respectively are 80.3% (standard error of 0.4%) and 85.1% (standard error of 0.3%). **(c)** Kaplan-Meier recurrence-free survival curves for each of the 500 bootstrap runs for patients identified by SpAn at low and high-risk of five-year CRC recurrence. **(d)** Boxplots of AUC values showing the statistically significant improvement in SpAn performance over the null model, which considers the TMA spot as a single whole, and unlike SpAn, does not dissect the TMA spot into epithelial and stromal spatial domains. Significance was computed using the 500 bootstraps. **(e)** Boxplots of stage-based area under the 500 bootstrapped ROC curves showing the stable stage-based clinical performance of SpAn. **(f)** Stable temporal performance of SpAn illustrated by the time-dependent AUC values plotted as a function of time in years. The 95%

confidence interval computed using the 500 bootstraps is also shown by the yellow shaded area around the mean time-dependent AUC values depicted by the solid black line.

Figure 4. SpAn based computational and systems derived spatial domain networks. (a) Epithelial domain, **(b)** Stromal domain, and **(c)** Epithelial-stromal domain Jensen-Shannon divergence matrices that show the information distance of partial correlations (computed for biomarkers selected by recurrence-guided SpAn feature-selection and stability analysis) between patients in the no-evidence-of-CRC and CRC-recurrence cohorts. **(d)** Epithelial domain, **(e)** Stromal domain, and **(f)** Epithelial-stromal domain spatial domain networks obtained by thresholding the corresponding spatial domain information distance matrices at the 99th percentile, which identify differential change most significant for CRC recurrence prognosis.

Figure 5. CRC recurrence-specific network biology inferred by SpAn. Domain-specific biomarkers identified by the spatial domain networks are used to interrogate the KEGG and STRING databases to identify domain-specific pathways enriched by the biomarkers. The epithelial, stromal and epithelial-stromal domain are respectively shown in green, red and blue, with pathways unique to those domains also coded with the same colors. Pathways that are enriched in more than one domain are coded with a color combination of those respective domains. For example, the PI3K-AKT signaling pathway is enriched in all three spatial-domains, and therefore, has a boundary box color-coded with all three colors. On the other hand, the mismatch repair pathway is enriched in the epithelial and epithelial-stromal domains, and is therefore, color-coded by red and blue colors.

Figure 6. Workflow of spatial analytics (SpAn) computational and systems pathology platform. (a-c) SpAn utilizes images of resected primary tumors from TMAs or whole slide images based on hyperplexed fluorescence and other imaging modality platforms,[18–21] to **(d)** perform spatial domain analysis for **(e)** patient diagnoses and prognoses and **(f)** infer recurrence-specific spatial domain networks to **(g)** potentially inform therapeutic strategies.

Figure 1

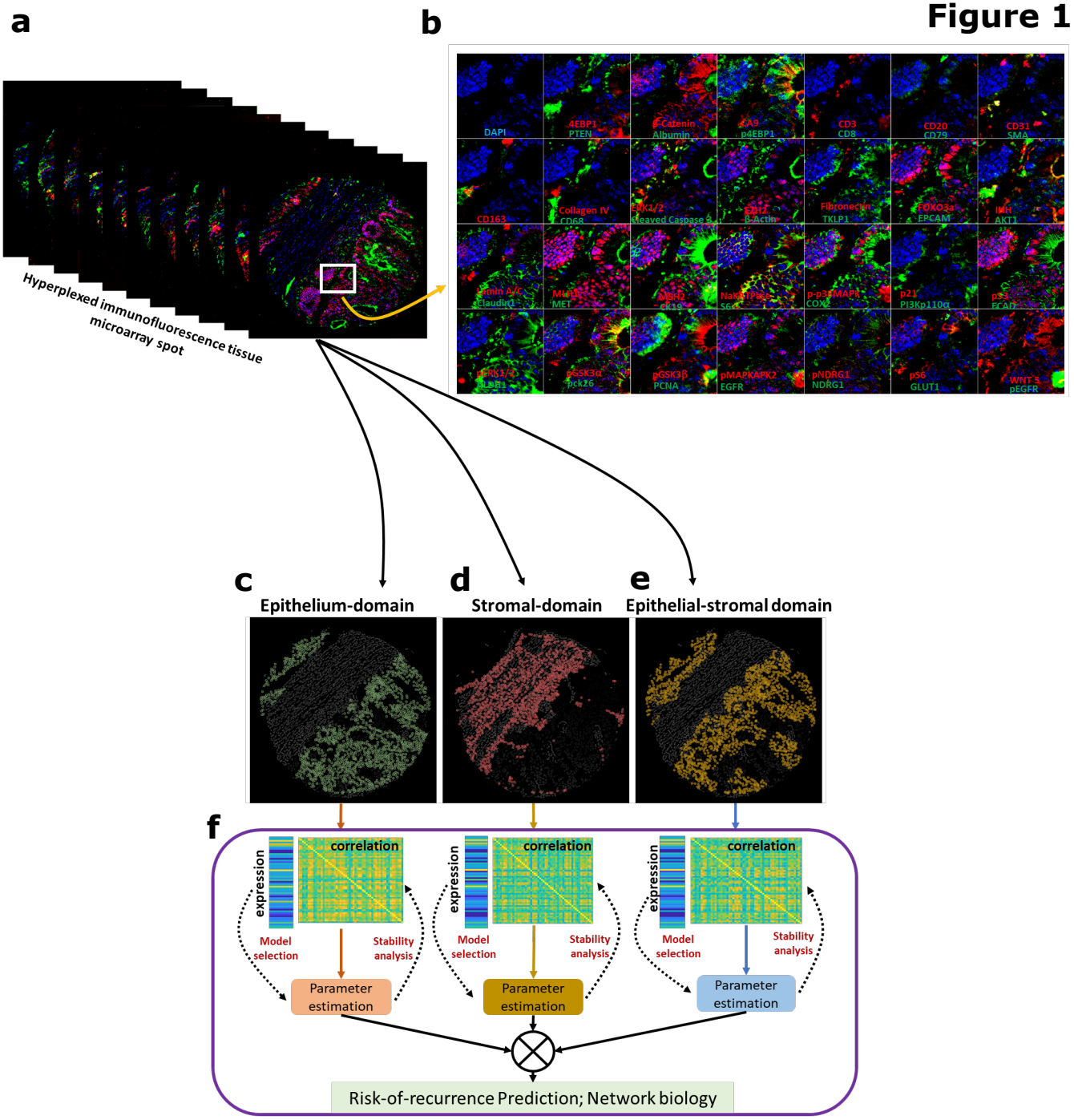

a
b
Hyperplexed immunofluorescence tissue microarray spot
DAPI
4EBP1
PTEN
β-Catenin
Albumin
CA9
p4EBP1
CD3
CD8
CD20
CD79
CD31
SMA
CD163
Collagen IV
CD68
ERK1/2
Cleaved Caspase 3
EZH2
β-Actin
Fibronectin
TKLP1
FOXO3a
EPCAM
IHH
AKT1
Lamin A/C
Claudin1
MLH1
MET
MSH2
CK19
NaKATPase
S6
p-p38MAPK
COX2
p21
PI3Kp110α
p53
ECAD
pERK1/2
LDH1
pGSK3α
pck26
pGSK3β
PCNA
pMAPKAPK2
EGFR
pNDRG1
NDRG1
pS6
GLUT1
WNT 5
pEGFR
c
Epithelium-domain
d
Stromal-domain
e
Epithelial-stromal domain
f
expression
correlation
Model selection
Stability analysis
Parameter estimation
Risk-of-recurrence Prediction; Network biology

Figure 2

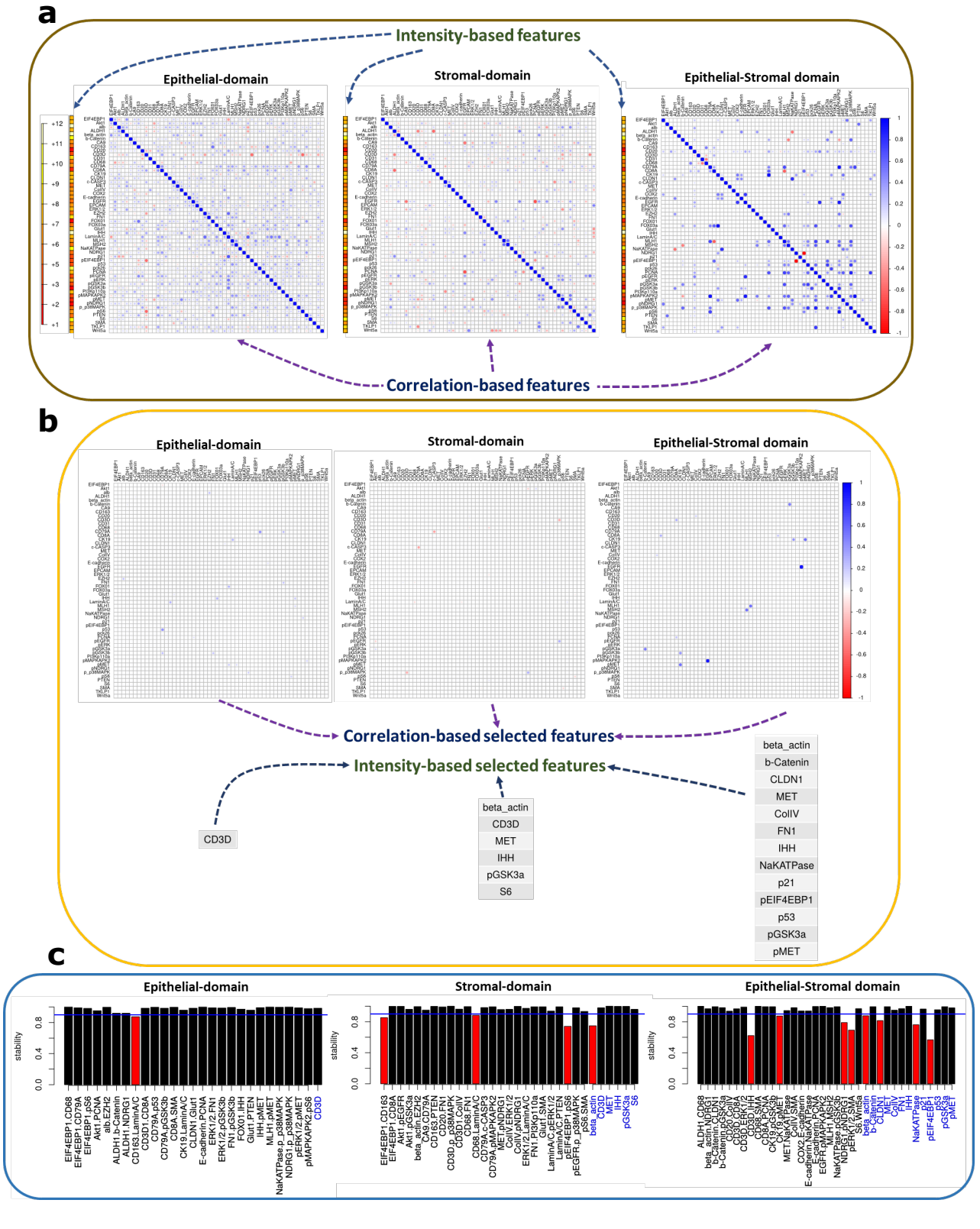

a
Intensity-based features
Epithelial-domain
Stromal-domain
Epithelial-Stromal domain
Correlation-based features
b
Epithelial-domain
Stromal-domain
Epithelial-Stromal domain
Correlation-based selected features
Intensity-based selected features
CD3D
beta_actin
CD3D
MET
IHH
pGSK3a
S6
beta_actin
b-Catenin
CLDN1
MET
CollV
FN1
IHH
NaKATPase
p21
pEIF4EBP1
p53
pGSK3a
pMET
c
Epithelial-domain
Stromal-domain
Epithelial-Stromal domain
stability

Figure 3

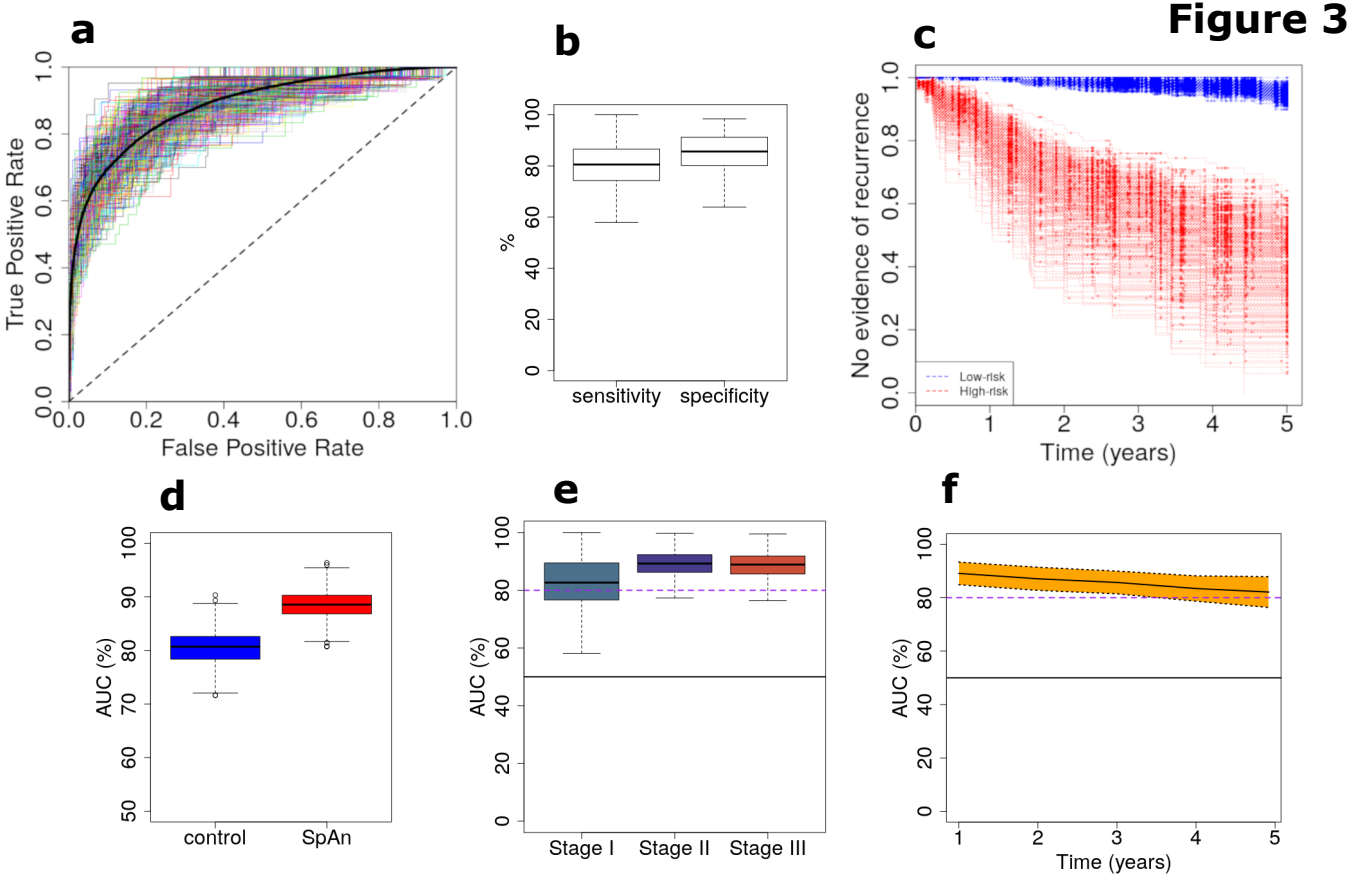

a
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
b
%
0
20
40
60
80
100
sensitivity
specificity
c
No evidence of recurrence
Time (years)
0
1
2
3
4
5
Low-risk
High-risk
d
AUC (%)
50
60
70
80
90
100
control
SpAn
e
AUC (%)
Stage I
Stage II
Stage III
f
AUC (%)
Time (years)

Figure 4

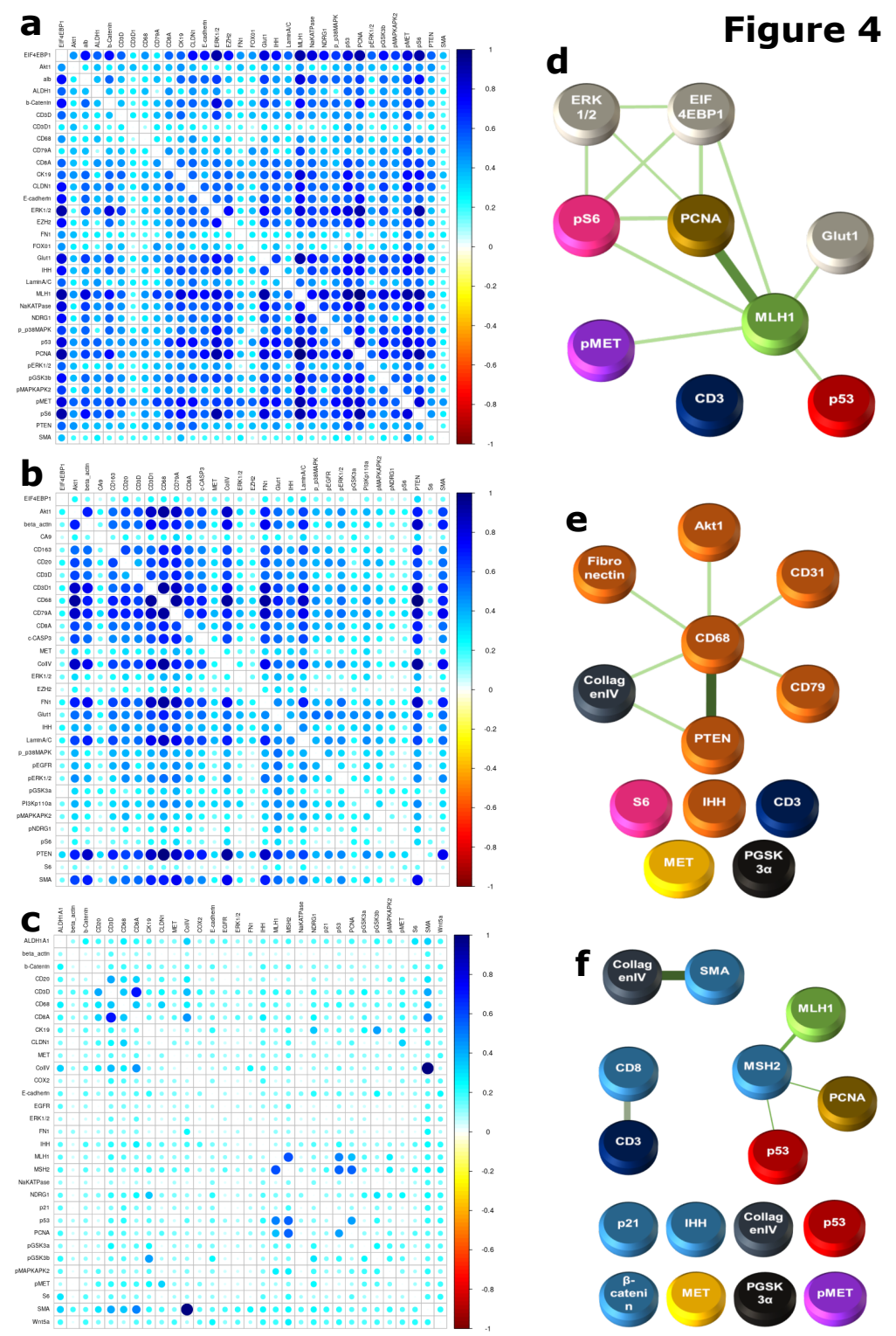

a
b
c
d
e
f
ERK 1/2
EIF 4EBP1
pS6
PCNA
Glut1
MLH1
pMET
CD3
p53
Akt1
Fibro nectin
CD31
CD68
Collag enIV
CD79
PTEN
S6
IHH
CD3
MET
PGSK 3α
SMA
CD8
MSH2
PCNA
CD3
p53
p21
IHH
Collag enIV
p53
β-cateni n
MET
PGSK 3α
pMET

Figure 5

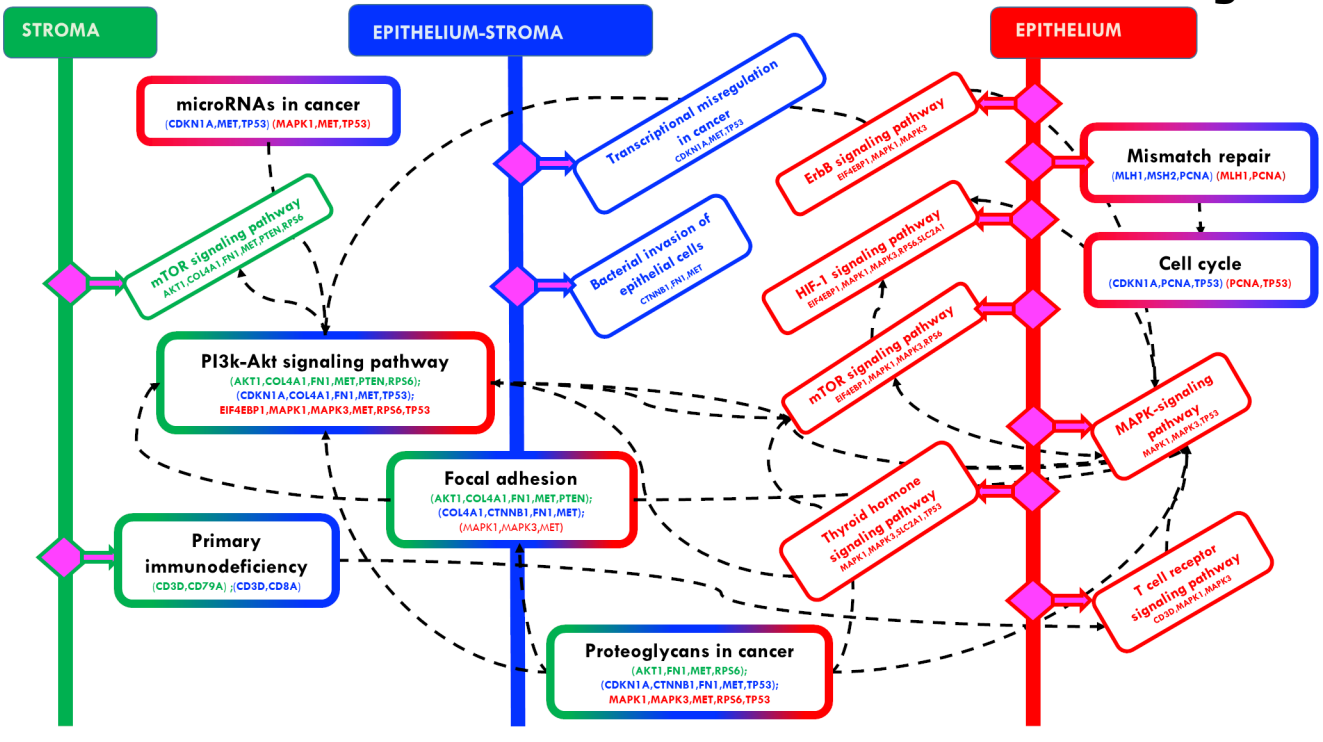
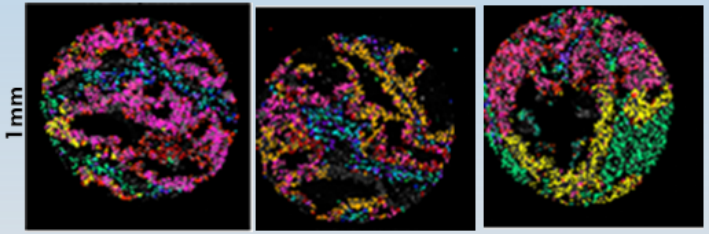

STROMA
EPITHELIUM-STROMA
EPITHELIUM
microRNAs in cancer
(CDKN1A,MET,TP53) (MAPK1,MET,TP53)
mTOR signaling pathway
AKT1,COL4A1,FN1,MET,PTEN,RPS6
Transcriptional misregulation in cancer
CDKN1A,MET,TP53
Bacterial invasion of epithelial cells
CTNNB1,FN1,MET
ErbB signaling pathway
EIF4EBP1,MAPK1,MAPK3
HIF-1 signaling pathway
EIF4EBP1,MAPK1,MAPK3,RPS6,SLC2A1
mTOR signaling pathway
EIF4EBP1,MAPK1,MAPK3,RPS6
Thyroid hormone signaling pathway
MAPK1,MAPK3,SLC2A1,TP53
Mismatch repair
(MLH1,MSH2,PCNA) (MLH1,PCNA)
Cell cycle
(CDKN1A,PCNA,TP53) (PCNA,TP53)
MAPK-signaling pathway
MAPK1,MAPK3,TP53
T cell receptor signaling pathway
CD3D,MAPK1,MAPK3
PI3k-Akt signaling pathway
(AKT1,COL4A1,FN1,MET,PTEN,RPS6);
(CDKN1A,COL4A1,FN1,MET,TP53);
EIF4EBP1,MAPK1,MAPK3,MET,RPS6,TP53
Focal adhesion
(AKT1,COL4A1,FN1,MET,PTEN);
(COL4A1,CTNNB1,FN1,MET);
(MAPK1,MAPK3,MET)
Primary immunodeficiency
(CD3D,CD79A) ;(CD3D,CD8A)
Proteoglycans in cancer
(AKT1,FN1,MET,RPS6);
(CDKN1A,CTNNB1,FN1,MET,TP53);
MAPK1,MAPK3,MET,RPS6,TP53

Figure 6

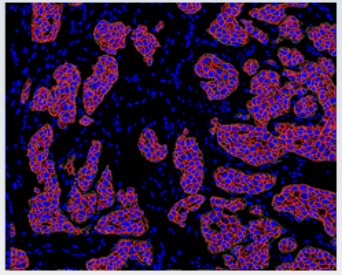
a. Patient cohort

b. Hyperplexed imaging

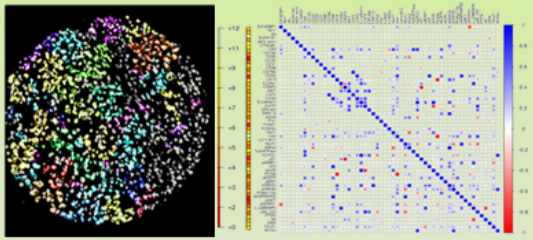


TMA/whole slide images

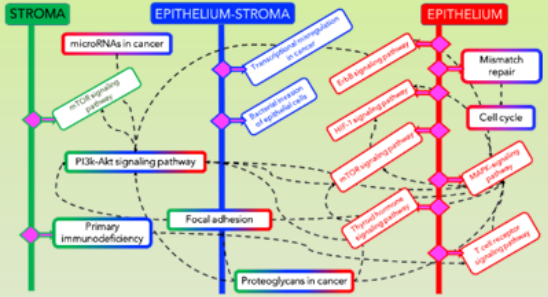
c. Image pre-processing



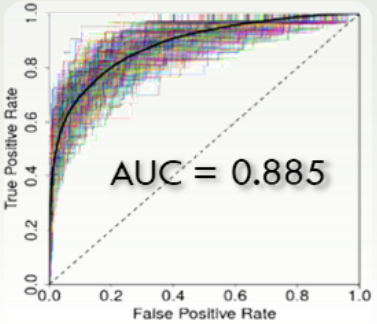
d. Perform spatial domain analysis



f. Infer recurrence-specific network biology of spatial domains



e. Perform diagnostics/prognostics



CRC Cohort
Stages I – III
N = 432
Chemo-naïve

g. Inform therapeutic strategies

cancer.gov

Table S1: Biomarkers and their relation to colon cancer

Table S2: Patient Cohort and Clinical Properties

Figure S1. Cell DIVE hyperplexed immunofluorescence imaging and processing scheme. For the Kth iteration, autofluorescence image of the tissue is acquired prior to labeling it with 2-3 fluorescent dye-conjugated primary antibodies and DAPI counterstain. Fluorescent-labeled tissue images are then acquired, followed by inactivation of the dyes and start of the next iteration.[20]

Figure S2. Tissue and cell segmentation. TMA spot, visualized here through a virtual Hematoxylin and Eosin image, is segmented into epithelial and stromal regions using E-cadherin to identify epithelial region from the stroma, and is followed by individual cell identification using Na^+K^+ATPase (cell membrane marker), ribosomal protein S6 (cytoplasmic marker) and DAPI (nuclear counterstain).

Figure S3. Validity of the proportional hazard assumption in penalized Cox regression. The p-values (shown in log scale) indicate the validity of the proportional hazard assumption for the overall Cox regression for all the 500 bootstraps and for each of the three domains. It can be seen that the overall global test is not statistically significant at the 95% confidence level (indicated by the dashed line) for almost all Cox models generated for the 500 bootstrap runs, demonstrating that the proportional hazard assumption is consistently valid.

Figure S4. Rationale for choosing 90% concordance rate. Plot of concordance of the penalized Cox regression model as a function of a threshold function that identifies the biomarker features most consistently selected by L1 penalization at the concordance level corresponding to the threshold. The larger the threshold the more stringent the consistency requirement on feature selection, and smaller the number of selected features. As shown in the plot, for low threshold values, the concordance value is saturated, and therefore, in this region injective correspondence between threshold value and concordance does not exist. In the monotonic decay region such a correspondence can be identified. The 90% concordance level identifies such a correspondence for all three spatial domains without compromising performance.

Figure S5. Coefficients for the recurrence-guided and domain-specific penalized Cox regression models of SpAn. Boxplots for coefficients that control the contribution of the selected features (obtained using L1-penalty) to each of the recurrence-guided and domain-specific penalized Cox regression under L2 regularization. The coefficients were computed for all 500 bootstrap runs and the boxplots capture the spread of values. It is worth noting that for all bootstraps the coefficients maintain their sign, which quantifies the nature of their contribution. A positive coefficient implies worse prognosis for increase in the corresponding feature value, while negative coefficient implies the converse.

Figure S6. Performance of SpAn platform for predicting risk of 5-year CRC recurrence for intensity-based features. SpAn ROC curves for predicting risk of 5-year CRC recurrence in

patients with resected CRC primary tumor using only biomarker expressions. The plot shows ROC curves for 500 bootstrap runs with independent training and validation sets. Mean area under the ROC curve is 72% with a standard error of 0.2%.

Figure S7. Time-dependent AUCs for domain-specific temporal performance of SpAn. Temporal performance of SpAn for the three spatial-domains illustrated by the time-dependent AUC values plotted as a function of time in years. The 95% confidence interval computed using the 500 bootstraps for each of the three spatial-domains is also shown by the yellow shaded area around the mean time-dependent AUC values depicted by the solid black line.

Table S1: Colon cancer related biomarkers

p4EBP1	Eukaryotic initiation factor 4E binding protein one: inhibits mRNA translation intitiation/phosphorylation relieves inhibition of mRNA translation initiation https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4845917/
AKT	Signal transduction:cell survival/anabolic metabolism https://www.ncbi.nlm.nih.gov/pubmed/31010234#
Albumin	Extracellular transport: binds drugs and small metabolites https://www.ncbi.nlm.nih.gov/pubmed/30662523
ALDH1	Cytosolic aldehyde dehydrogenase / retinoic acid metabolism / alcohol Metabolism https://www.ncbi.nlm.nih.gov/pubmed/22101256
BetaActin	Microfilament protein: structure and motility https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4133152/
BetaCatenin	Adherens junctions/Wnt signaling https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5464927/
CA9	Membrane associated carbonic anhydrase: Hypoxia response/pH modulation https://www.ncbi.nlm.nih.gov/pubmed/28554753
CD20	B-lymphocyte protein of unknown function https://www.ncbi.nlm.nih.gov/pubmed/26215222
CD31	Endothelial cell:cell junctions/immune cell transendothelial migration https://www.ncbi.nlm.nih.gov/pubmed/29110224
CD68	Tissue macrophage scavenger receptor / endosomal-lysosomal glycoprotein /selectin-dependent migration https://www.ncbi.nlm.nih.gov/pubmed/23868006
CD79	B-lymphocyte antigen receptor complex - mediates antigen dependent B-lymphocyte activation and signal transduction https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3637070/
CD8a	Antigen recognition/class I MHC binding/T-lymphocyte mediated killing https://www.ncbi.nlm.nih.gov/pubmed/30042763
CK1,5,6,8 (pck26)	Pan-cytokeratin: recognizes most basal and luminal epithelial cell subtypes https://www.ncbi.nlm.nih.gov/pubmed/23902688
CK19	Cytokeratin 19: unpaired with basic cytokeratin Useful for identification of epithelium and epithelial malignancies including adenocarcinomas of colon, stomach, pancreas, biliary tract, liver, breast, and thyroid carcinoma of the papillary type https://www.ncbi.nlm.nih.gov/pubmed/19956064

Claudin1	Epithelial and endothelial tight junction protein/epithelial barrier function/homo and heteropolymers and zona occludins protein binding. Neu- trophil expression inCRC. Colocalizes withneutrophil specific elastase https://www.ncbi.nlm.nih.gov/pubmed/29504115
Cleaved Caspase 3	Cysteine peptidase - active form: activation of apoptosis/possible role in lipid metabolism through activation of SREBP https://www.ncbi.nlm.nih.gov/pubmed/23591201
E-cadherin	Epithelial specific homotypic adherens junctions/Wnt signaling/epithelial-mesenchymal transition/calcium dependent cell-celladhesion/E7 integrin ligand https://www.ncbi.nlm.nih.gov/pubmed/30904616
EGFR	Epidermal growth factor receptor: receptor tyrosine kinase/ binds epithelial growth factor/ homo or heterodimerization/ activates signal transduction https://www.ncbi.nlm.nih.gov/pubmed/20857619
EPCAM	Epithelial cell adhesion molecule: homotypic calcium independent cell adhesion molecule/ cell cycle modulation (myc, cyclin A and E) https://www.ncbi.nlm.nih.gov/pubmed/20606680
ERK	Mitogen activated protein kinase 1: central hub of mitogenic-progrowth signaling https://www.ncbi.nlm.nih.gov/pubmed/29988110
EZH2	Polycomb repressor complex two methyltransferase: inactivates gene expression through histone acetylation on H3K9 and H3K27 https://www.nature.com/articles/s41598-017-13670-z
Fibronectin	Secreted cell adhesion protein involved in basement membrane function https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5112592/
FOXO1	Forkhead Box-O1: transcription factor involved in regulating intermediary metabolic enzymes and cellsurvival https://www.ncbi.nlm.nih.gov/pubmed/30518405 https://www.ncbi.nlm.nih.gov/pubmed/29867200
FOXO3a	Forkhead Box-O3a: transcription factor involved in regulating intermediary metabolic enzymes and cell survival https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3721407/
GLUT1	Facilitated glucose/aldose uptake: upregulated in cancer cells/aerobic glycolysis/Warburg effect https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5593607/ https://www.ncbi.nlm.nih.gov/pubmed/21786248
Indian Hedge Hog	Hedgehog/patched/smoothened signaling: involved in regulating bone metabolism (ossification). https://www.ncbi.nlm.nih.gov/pubmed/30926792
Lamin A/C	Nuclear lamin protein: involved in organizing nuclear pores and chromatin https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2496895/

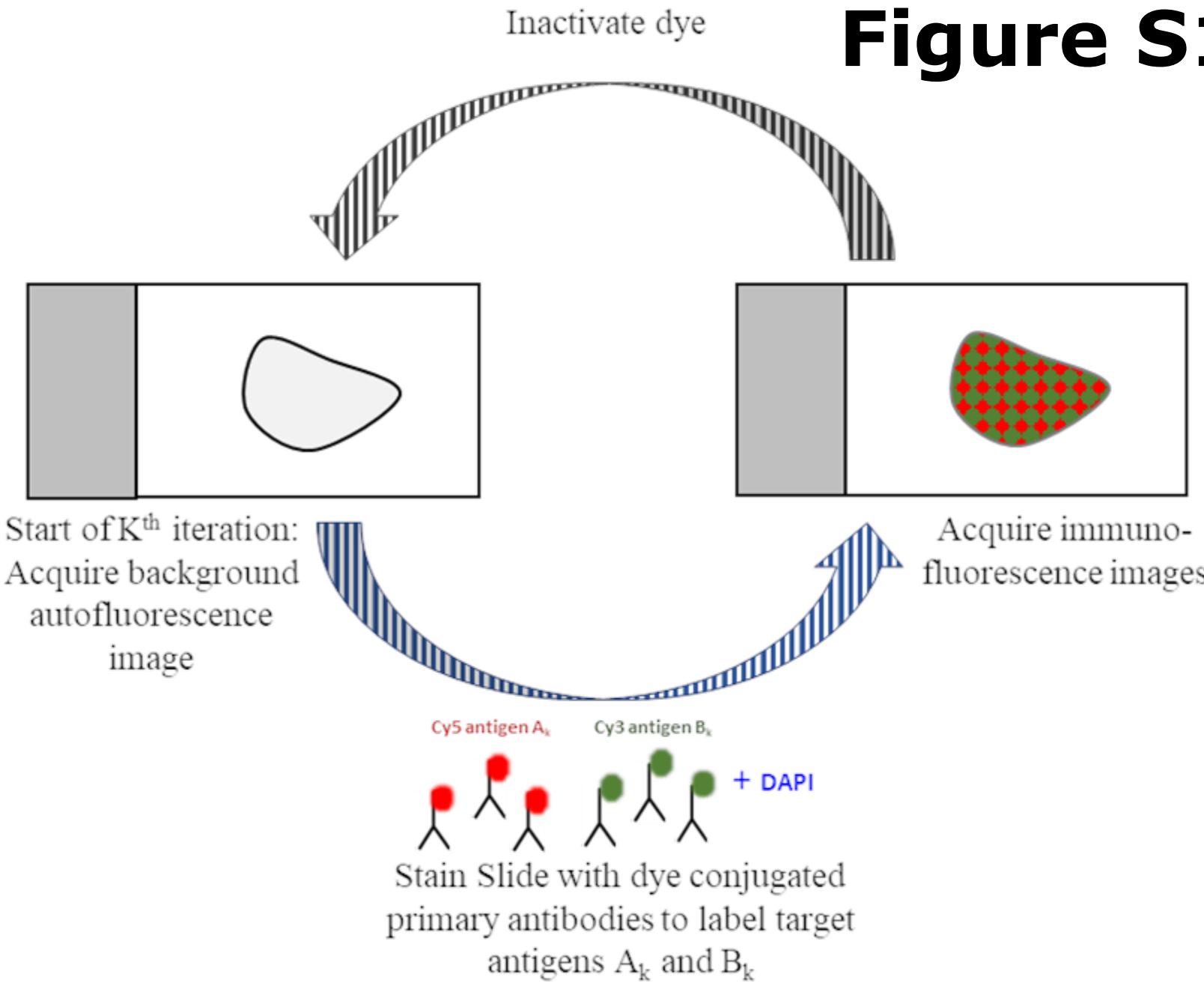
MLH1	MutL homolog one: mismatch repair enzyme Used for differential identification of colorectal carcinoma. Deficiency of MLH-1 is associated with the onset of HNPCC https://www.nature.com/articles/6600565
Na^+/K^+-ATPase	Sodium potassium antiporter - ATP dependent: regulates membrane potential polarization/constitutive expression in all known cell types https://www.ncbi.nlm.nih.gov/pubmed/9450495
NDRG1	nMyc Downstream Regulated One: involved in cytoskeletal dynamics/adherens junctions/metabolism/poorly characterized https://www.ncbi.nlm.nih.gov/pubmed/30407715 https://www.ncbi.nlm.nih.gov/pubmed/28346422
p-p38MAPK	Mitogen-Activated Protein Kinase 14: integrates stress signals from environmental and cytokine stimuli/phosphorylation indicates the kinase is active. https://www.ncbi.nlm.nih.gov/pubmed/29687844
p21	Cyclin dependent kinase inhibitor. https://www.ncbi.nlm.nih.gov/pubmed/30343528
pEGFR	Epidermal growth factor receptor: receptor tyrosine kinase/ binds epithelial growth factor/ homo or heterodimerization/ activates signal transduction https://www.ncbi.nlm.nih.gov/pubmed/26171935
pERK 1/2	Mitogen activated protein kinase 1/2: central hub of mitogenic-progrowth signaling/this phosphorylation on the residues targeted here is indicative of kinase activity Reference: https://www.ncbi.nlm.nih.gov/pubmed/22108866
pGSK3a	Glycogen sythase kinase alpha: probable component of the destruction com-plex (-catenin degradation)/phsophorylation on the residue targeted here in- activates the kinase though the creation of a pseudosubstrate https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270937/
pGSK3beta	Glycogen sythase kinase beta: known component of thedestruction complex (-catenin degradation)/phsophorylation on the residue targeted here inactivates the kinase though the creation of a pseudosubstrate https://www.ncbi.nlm.nih.gov/pubmed/16043125
pMAPKAPK2	Mitogen-Activated Protein Kinase-Activated Protein Kinase 2: p38MAPK substrate motif/ active MAPKAPK2 stabilizes TNF and IL6 mRNA/destabilizes HSP27complexes https://www.ncbi.nlm.nih.gov/pubmed/15774489
pMET	Met Proto-Oncogene (Hepatocyte Growth Factor Receptor): instigates signaling through prosurvival and EMT signaling pathways/expression is required for gastrulation andderegulated overexpression iscommon in many cancers/Y1349 phosphorylation serves as a Gab1 binding site (a scaffold for activation of PI3K, PLC, and SHP2) https://www.nature.com/articles/1209859 https://www.ncbi.nlm.nih.gov/pubmed/18258742

pNDRG1	nMyc Downstream Regulated One: involved in cytoskeletal dynamics/adherens junctions/metabolism/poorly characterized/phosphorylation of this residue is downstream of mTORC2 https://www.nature.com/articles/onc201774
pS6	Ribosomal Protein S6: phosphorylation is associatedwith mitogens and growth factors and may regulate selective translation of particular classes of mRNAs defined by consensus sequences in untranslated regions. https://www.ncbi.nlm.nih.gov/pubmed/17175097 https://www.ncbi.nlm.nih.gov/pubmed/29526493
PTEN	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN: tumor suppressor/opposes PI3K function by de- phosphorylating the insoitol 3'OH group/emerging nuclear functions related to DNA repair and apoptosis https://www.ncbi.nlm.nih.gov/pubmed/30983885 https://www.ncbi.nlm.nih.gov/pubmed/30118842
S6	Ribosomal Protein S6: phosphorylation is associatedwith mitogens and growth factors and may regulate selective translation of particular classes of mRNAs defined by consensus sequences in untranslated regions https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2474683/
SMA	Smooth muscle actin alpha: cytoskeletal protein of smooth muscle and vascular pericytes https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5876461/
TKLP1	Transketolase-like protein 1: transfers 2 carbon ketol groups to aldose acceptor molecules (TKTL1) https://www.ncbi.nlm.nih.gov/pubmed/26650256
WNT5a	Wnt signaling secreted glycoprotein 5a: signals through canonical and non-canonical Wnt pathways/may affect cell motility and metastasis http://clincancerres.aacrjournals.org/content/14/1/55

Table S2

	All	StageI	StageII	StageIII
Patients	432	175	175	175
Age (median)	75	67	73	70
Age (std)	6.7	11.7	11.4	13.4
Gender	220	88	101	31
Gender (%)	50.9	50.3	53.2	46.3
No. of Recurrences	65	10	30	25
Recurrence (%)	100	15.4	46.2	38.5
Recurrence Time (median)	1651	1899	1574	1098
Survival Time (median)	1695.5	1899	1658	1280

Figure S1


Inactivate dye
Start of Kth iteration: Acquire background autofluorescence image
Acquire immuno-fluorescence images
Cy5 antigen A_k
Cy3 antigen B_k
+ DAPI
Stain Slide with dye conjugated primary antibodies to label target antigens A_k and B_k

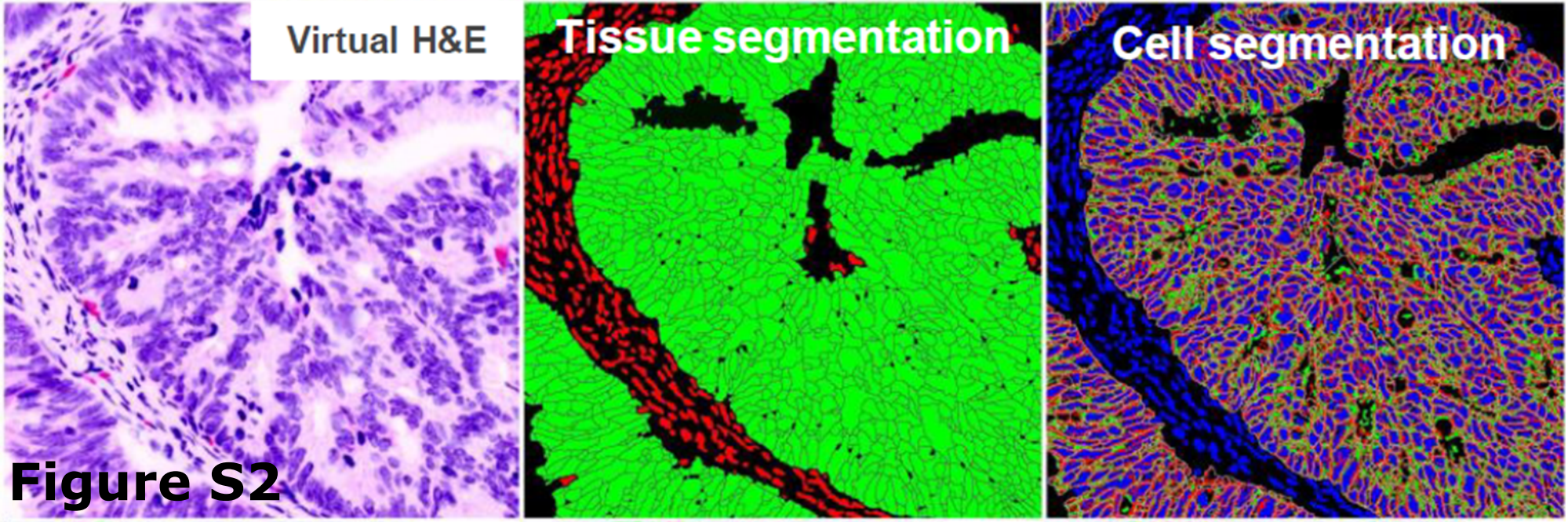
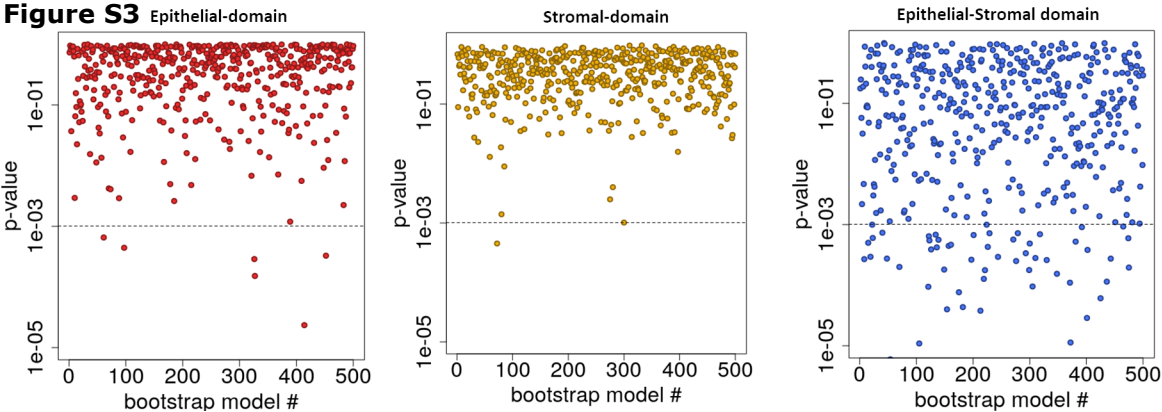


Figure S2

Figure S3


Epithelial-domain
Stromal-domain
Epithelial-Stromal domain
p-value
1e-01
1e-03
1e-05
0
100
200
300
400
500
bootstrap model #

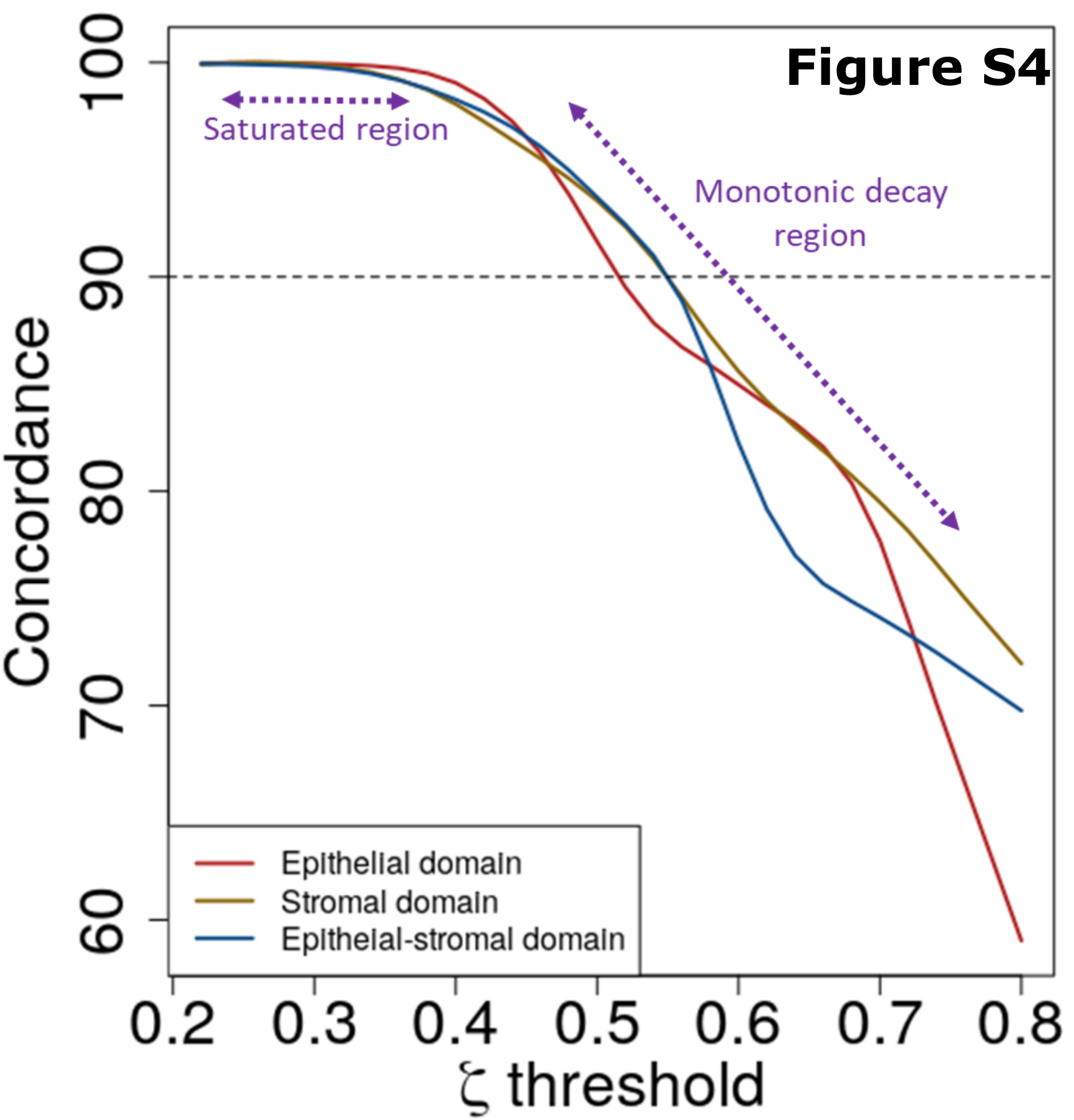
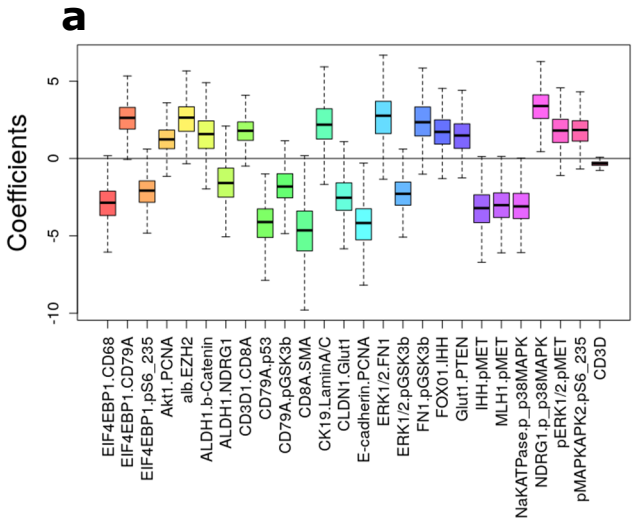
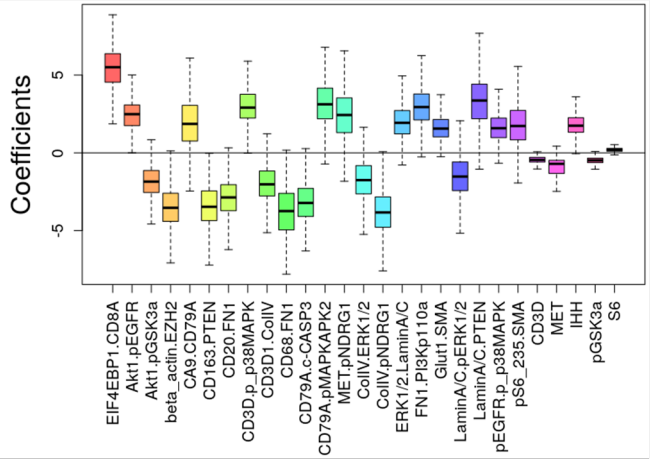


Figure S4
Saturated region
Monotonic decay region
Concordance
100
90
80
70
60
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ζ threshold
Epithelial domain
Stromal domain
Epithelial-stromal domain

Figure S5


a
Coefficients
5
0
-5
-10
EIF4EBP1.CD68
EIF4EBP1.CD79A
EIF4EBP1.pS6_235
Akt1.PCNA
alb.EZH2
ALDH1.b-Catenin
ALDH1.NDRG1
CD3D1.CD8A
CD79A.p53
CD79A.pGSK3b
CD8A.SMA
CK19.LaminA/C
CLDN1.Glut1
E-cadherin.PCNA
ERK1/2.FN1
ERK1/2.pGSK3b
FN1.pGSK3b
FOX01.IHH
Glut1.PTEN
IHH.pMET
MLH1.pMET
NaKATPase.p_p38MAPK
NDRG1.p_p38MAPK
pERK1/2.pMET
pMAPKAPK2.pS6_235
CD3D


b
Coefficients
5
0
-5
EIF4EBP1.CD8A
Akt1.pEGFR
Akt1.pGSK3a
beta_actin.EZH2
CA9.CD79A
CD163.PTEN
CD20.FN1
CD3D.p_p38MAPK
CD3D1.ColIV
CD68.FN1
CD79A.c-CASP3
CD79A.pMAPKAPK2
MET.pNDRG1
ColIV.ERK1/2
ColIV.pNDRG1
ERK1/2.LaminA/C
FN1.PI3Kp110a
Glut1.SMA
LaminA/C.pERK1/2
LaminA/C.PTEN
pEGFR.p_p38MAPK
pS6_235.SMA
CD3D
MET
IHH
pGSK3a
S6

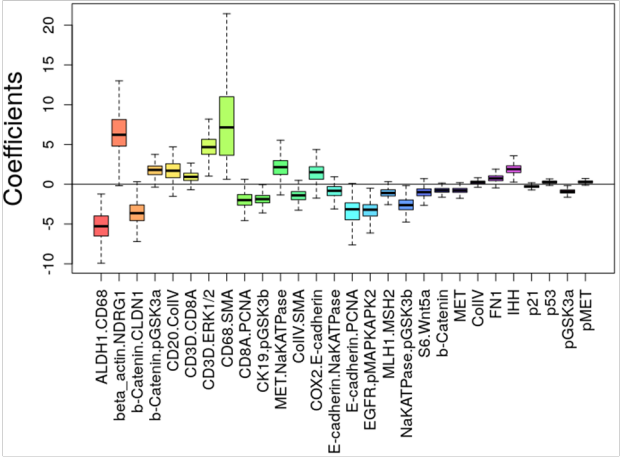

c
Coefficients
20
15
10
5
0
-5
-10
ALDH1.CD68
beta_actin.NDRG1
b-Catenin.CLDN1
b-Catenin.pGSK3a
CD20.ColIV
CD3D.CD8A
CD3D.ERK1/2
CD68.SMA
CD8A.PCNA
CK19.pGSK3b
MET.NaKATPase
ColIV.SMA
COX2.E-cadherin
E-cadherin.NaKATPase
E-cadherin.PCNA
EGFR.pMAPKAPK2
MLH1.MSH2
NaKATPase.pGSK3b
S6.Wnt5a
b-Catenin
MET
ColIV
FN1
IHH
p21
p53
pGSK3a
pMET

Figure S6

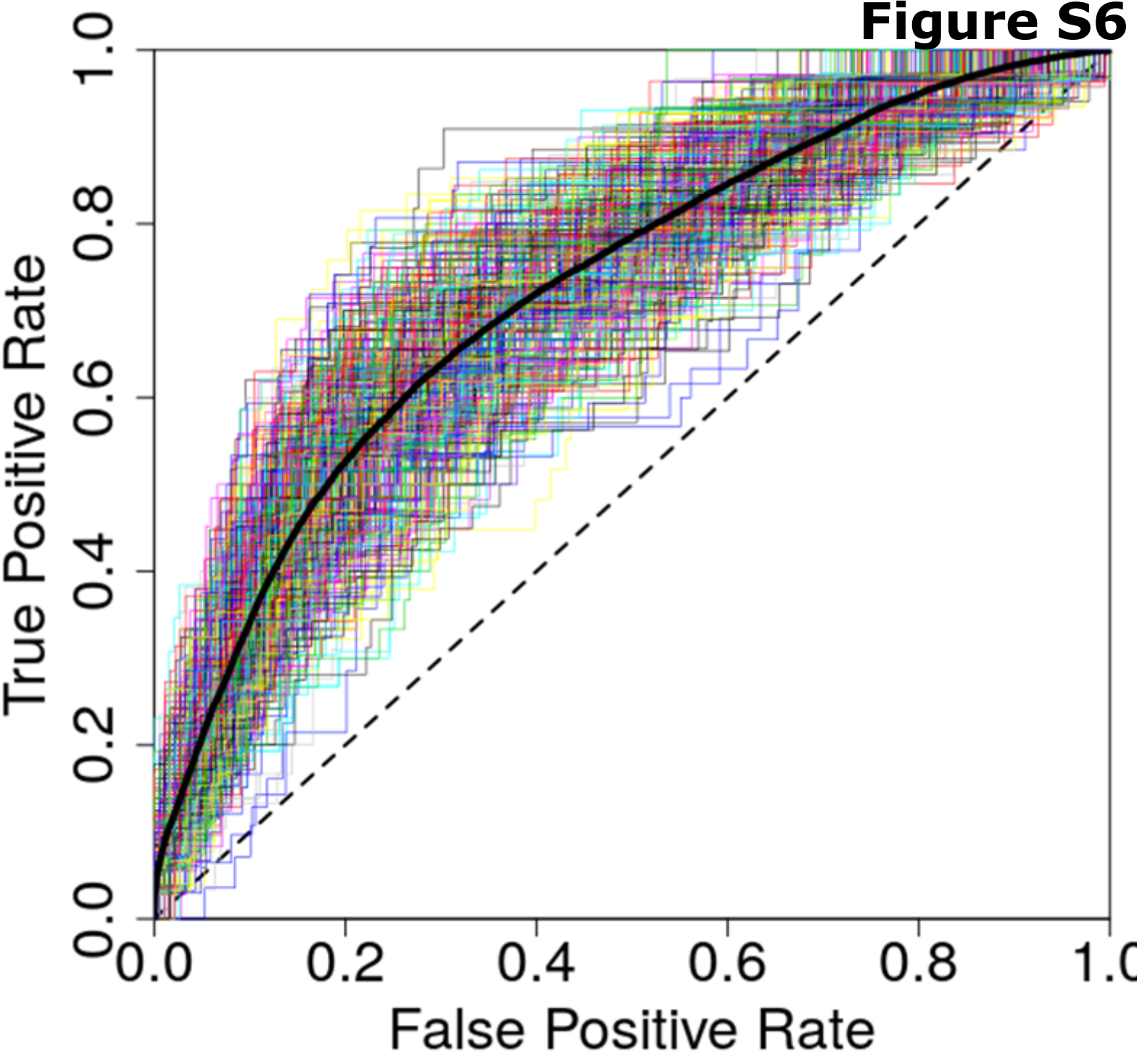
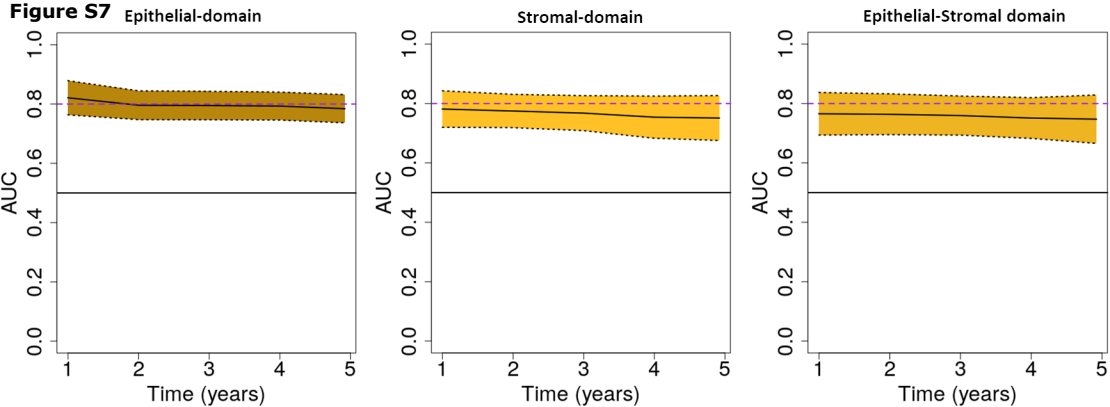


Figure S7


Epithelial-domain
Stromal-domain
Epithelial-Stromal domain
AUC
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Time (years)