Unsupervised cellular phenotypic hierarchy enables spatial intratumor heterogeneity characterization, recurrence-associated microdomains discovery, and harnesses network biology from hyperplexed in-situ fluorescence images of colorectal carcinoma

Samantha A. Furman[1], Andrew M. Stern[1,2], Shikhar Uttam[1,+], D. Lansing Taylor[1,2], Filippo Pullara[3], S. Chakra Chennubhotla[1,2,3,*]

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

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

[3]SpIntellx, Inc., 2425 Sidney St, Pittsburgh, PA 15203

*Corresponding Author Email: chakra@spintellx.com (S. Chakra Chennubhotla)

[+]New affiliation: UPMC Hillman Cancer Center, Pittsburgh, PA, 15213, USA

Abstract

LEAPH is an unsupervised machine **le**arning **a**lgorithm for characterizing in situ **p**henotypic **h**eterogeneity in tissue samples. LEAPH builds a phenotypic hierarchy of cell types, cell states and their spatial configurations. The recursive modeling steps involve determining cell types with low-ranked mixtures of factor analyzers and optimizing cell states with spatial regularization. We applied LEAPH to hyperplexed (51 biomarkers) immunofluorescence images of colorectal carcinoma primary tumors (N=213). LEAPH, combined with pointwise mutual information (PMI), enables the discovery of phenotypically distinct *microdomains,* composed of spatially configured computational phenotypes. LEAPH identified a subset of microdomains whose network biology is consistent with a tumor mediated immunosuppression program driven by cancer stem cells through an enhanced Wnt signaling pathway. The LEAPH framework, when combined with microdomain discovery and microdomain-specific network biology, has the potential to provide insights into pathophysiological mechanisms, identify novel drug targets and inform therapeutic strategies for individual patients.

Introduction

Tumor cells and their stromal cell counterparts that comprise the tumor microenvironment (TME) reciprocally coevolve to generate heterocellular communication networks. A distinctive characteristic of the functional organization of this continuously evolving ecosystem is spatial intratumoral heterogeneity (ITH), a key determinant of disease progression landmarks in multiple carcinomas that include colorectal cancer [1-10]. Therefore, to optimize diagnosis, prognosis, therapeutic strategies and to identify novel therapeutic targets it is important to define spatial ITH in the tumors of individual patients and determine the mechanistic underpinnings of its relationship to metastatic potential, immune evasion, recurrence, therapeutic response and drug resistance. [1, 3, 11, 12].

The first step in investigating spatial ITH is to identify the cell phenotypes within the TME. This, however, is a challenging task owing not only to the diversity of well-defined cell types within the TME but also the intrinsic plasticity of many of these cell types in response to the selection pressures within their particular confines (i.e., microdomains) [13]. Several single cell approaches are well suited to capture this emergent phenotypic continuum, but at the expense of preserving spatial context. Thus, complementary experimental and computational approaches are needed to further define spatial ITH. The choice of construction of a cellular phenotyping algorithm is based on the type of data generated due to the steep tradeoff between cellular resolution, spatial context, and the number of genomic/proteomic probes employed [14-16]. There has been a recent explosion of hyperplexed [17] (> 9 fluorescence or mass spec-based) biomarker labeling and imaging modalities utilizing various reagent technologies to probe the same tissue sections with several dozens of biomarkers at cellular and subcellular resolutions [18-20]. The challenge now is to accurately characterize the complex spatial and high-dimensional output of these hyperplexed techniques.

Single-cell phenotyping analysis exists in various forms, each with advantages and drawbacks. With high-dimensional data, the analysis usually involves dimensionality reduction but observations have been made that cellular subpopulations distinct in multidimensional space overlap visually in the reduced two-dimensional space [21-24]. Common drawbacks of single-cell clustering algorithms include: 1) reliance on supervised methods leading to user-defined cellular phenotypes as opposed to data-driven cellular phenotypes, 2) lack of a spatial context due to the limitations of data generation, 3) use of hard clustering (a cell belongs to one and only one cellular phenotype) which leaves no room to identify cells along a phenotypic

continuum. In Table S1, we detail the current state-of-the-art cellular phenotyping methods and analysis studies within the past decade and specify the incorporation of spatial context, soft clustering, hierarchical component, and unsupervised techniques. To our knowledge, there does not exist a cellular phenotyping method and spatial analysis framework that is unsupervised, harnesses the spatial context, and promotes the phenotypic continuum with sensitivity for identifying transitional state cell populations from hyperplexed in-situ fluorescence images.

In this paper, we propose LEAPH, an unsupervised machine **le**arning **a**lgorithm for characterizing cellular **p**henotypic **h**eterogeneity in a tumor microenvironment (Figure 1). LEAPH builds a phenotypic hierarchy of cell types, cell states and their spatial configurations. We applied LEAPH on hyperplexed (51 biomarkers) immunofluorescence images of colorectal carcinoma (CRC) primary tumor tissue samples (N=213). The biomarkers were selected with the interest of sampling a range of CRC and cancer biology TME properties (see Data section).

The data-driven and computationally unbiased approach of LEAPH captured a phenotypic hierarchy comprised of *specialized*, *transitional*, and *multi-transitional* cell states that are intrinsic to the architecturally complex and reciprocally co-evolving TME of CRC. Each LEAPH derived computational phenotype (CP) is further characterized by a unique biomarker signature for ease of interpretation. With these biomarker signatures, we conclude properties demonstrating the heterogeneity of immune cells, cancer stem cells, tumor cells, and cancer associated fibroblasts. In addition, we performed a virtual simulation where the tissue is assumed to be labeled with a subset of key biomarkers (as opposed to the 51 biomarkers we started with) to demonstrate the process to obtain a rank-order subset of biomarkers for characterizing phenotypic diversity.

Using our previous work utilizing pointwise mutual information (PMI) to characterize spatial ITH [25], we show how LEAPH enables the discovery of *microdomains*, which are spatial configurations of computational phenotypes driving disease progression dynamics in the TME of CRC. With a microdomain-specific network biology approach using partial correlation network analysis described in our previous work [26], we find microdomains where the cancer stem cells are driving a Wnt-signaling based immunosuppressive program in patients that will experience recurrence within 3 years post-surgical resection. The LEAPH framework, when combined with microdomain discovery and spatial network biology, has the potential to provide mechanistic insights into disease biology, inform therapeutic strategies for individual patients, identify novel targets for drug discovery and support prognostic/diagnostic predictive analyses.

Results

Hyperplexed fluorescence colorectal cancer (CRC) data

The primary source of data for this study, is a cohort of 213 CRC tissue samples (TMA core size = 0.6mm, 1 sample per patient) hyperplexed using Cell DIVE (GE Life Sciences, Issaquah, WA) [18] (Figure 1a). Currently, there exists multiple hyperplexing technologies, each with unique properties. Hyperplex imaging techniques can be non-destructive with the methodology of antibody labeling being either iterative [18] or performed in a batch prior to imaging [19]. There exists other hyperplexing technologies unrelated to fluorescence imaging such as mass cytometry which are both destructive and perform biomarker batch labeling [20].

For this study, our data is imaged using Cell DIVE (GE Life Sciences, Issaquah, WA) [18] which involves non-destructive cyclical immunofluorescence labeling with two or three antibodies labeled with distinct fluorescent probes, followed by imaging and subsequent quenching of the fluorescence. This process is repeated to capture all the required antibodies (biomarkers). The data consists of image stacks taken at each region of interest and overall image stack consists of several images for each of several imaging rounds. Each round includes a nuclear (DAPI) image that is used as a reference for registering all the images from all the rounds. Quantitation of images in each round includes the fluorescence intensity of each measured biomarker. Images are also acquired after quenching rounds for the purpose of autofluorescence removal [18]. Validation of this process shows robustness and preservation of biomarker stability and biological integrity [18, 27-31]. Processing of Cell DIVE images includes correction for uneven illumination across the field of view, removal of autofluorescence, registration, and automated quality control (QC) detection of several categories of defects, including failed registration, blurred or saturated images, and other imaging issues.

The images and data undergo a series of tissue and cell quality checks, log2 transformation and normalization steps. To integrate data from batch processing each biomarker is normalized to a control median. Cellular segmentation is done using a collection of structural biomarkers: NaKATPase (cell membrane, cell border), S6 (cytoplasm), and DAPI (nucleus) (Figure S1). Cells are filtered using individual QC scores generated for each cell (scores less than 0.7-0.8 will not be included indicating inaccurate registration, misalignment, or tissue loss) and based on number of pixels per segmented subcellular compartment. See Methods for more information on data generation and pre-processing steps.

The biomarkers chosen are protein markers for specific cell lineages, oncogenes, tumor suppressors, and post-translational protein modifications indicative of cellular activation states [18] (Figure S2, Table S2). A subset of the biomarkers was selected specifically based on their known association with CRC. In addition to the tissue samples, the CRC patient data set also includes clinical information for each patient regarding sex, age, chemotherapy treatment, and days until recurrence post-surgical resection (patient statistics in Table S3).

LEAPH builds a phenotypic hierarchy of cell types and states through recursive steps of probabilistic clustering and spatial regularization

LEAPH can process a range of input data from a single sample (~2K cells) to a cohort of samples (tested on 213 tissue samples, ~500K cells total). LEAPH applies recursive steps of modeling cell types with low-ranked mixtures of factor analyzers (MFA) and cell states with spatial regularization to account for the spatial context around each cell (Figure 1b). The *ownership probabilities* for any given cell are constructed recursively by parsing through the hierarchy. Our stopping criterion for recursion is to apply thresholds on the angle between the subspaces created by the mixture models and the fraction of cells with a majority ownership probability in each cluster to ensure that the hyperplexed dataset is not being overfitted (see Methods). The terminal nodes of the tree, i.e., *leaves*, signify distinct data-driven CPs discovered in the input dataset and form the components of the final MFA model determined by the recursive decomposition (Figure 1c).

We define cell states as being *specialized* or *non-specialized* based on the CP ownership probabilities for each cell. A specialized cell state is defined by a strong propensity towards a single CP (ownership probability >0.95). For convenience, we further group the non-specialized cell state as being a *transitional* (ownership probabilities spread between two CPs) or *multi-transitional* (ownership probabilities spread across more than two CPs). We can visualize the spatial distribution of the resulting CP's within a tissue sample by assigning each cell to the CP with the highest ownership probability (Figure 1c).

To illustrate the process of discovering phenotypic hierarchy, we generated synthetic biomarker data (Figure 2a) to reflect the distribution of biomarker values in the hyperplexed CRC data. In particular, there is a varying spread of biomarker values likely due to the individual biomarker response sensitivity (Figure S4) and the observation of dominant phenotypes (e.g., epithelial and stromal cells) with various nested cell subtypes (e.g., immune, cancer stem cells). To automate the process of phenotypic discovery, we use a recursive probabilistic approach with a

low-dimensional latent space mixture of factor analyzers (MFA) (See Methods for more details). Each mixture component is a factor analyzer in a two-dimensional latent space, as we observed this is sufficient to capture the input variance (Figure S5). On the synthetic data (Figure 2a top-left), we instantiate a one-tier MFA model with four mixture components (Figure 2a top-right) and our proposed recursive decomposition where each level of the hierarchy identifies two dominant mixture components (Figure 2a bottom). Comparing the results to the ground truth synthetic data, the one-tier MFA model is unsuccessful at discovering the four phenotypes. The recursive decomposition method separates the larger broad clusters in the first level and the finer sub-clusters at the second level of the hierarchy.

The MFA model parameters are learned through random initializations over a hundred different runs and a consensus set of model parameters are inferred to capture the optimal subspace representation (Figure S6). The recursive probabilistic clustering when applied along results in a large and unrealistic population of non-specialized cells (typically 25% in a single tissue sample). The soft clustering is agnostic to the spatial complexity of the TME, a key component driving ITH. Based on properties of spatial ITH and the spatial tissue architecture of a tumor, we expect neighborhoods of cells to be spatially coherent (e.g., epithelial/tumor cells to be surrounded by, or spatially proximal to, other epithelial/tumor cells, but making allowance for the presence of tumor-infiltrating lymphocytes and other stromal cells) [32]. To filter false-positive non-specialized cells, we add a spatial regularization component to the recursive probabilistic clustering.

The novel objective function for the spatial regularization component consists of two terms: one to promoting ownership probability confidence and one to promote spatial coherence within a small neighborhood. We assume the two terms hold equal weight and calculate the tuning parameter accordingly. All cells are tested under this objective function but only the cells classified as non-specialized initially (ownership probability 0.05-0.95) are further optimized. Cells initially classified as specialized are fixed (remain invariant during the optimization) to avoid altering the ownerships of cells such as tumor-infiltrating lymphocytes which will *not* maintain spatial coherence. On a single tissue sample, we track the ownership probabilities of the cells undergoing spatial regularization. We observe that most of the cells converge to one cluster, but a subset of cells remains non-specialized (1-4% in a single tissue sample) (Figure 3b). To gain a deeper spatial understanding, we demonstrate the transformation of ownership probabilities of two cells in this simulation. One cell conforms to the surrounding homogenous neighborhood and one cell remains non-specialized between the two clusters due to the

heterogeneous nature of the surrounding neighborhood (Figure 3c). A more detailed description of LEAPH can be found in the Methods section.

LEAPH captures a diverse set of computational phenotypes having distinct biomarker signatures comprised of specialized, transitional, and multi-transitional cell states and enables the visualization of their spatial configurations

We applied LEAPH on the CRC hyperplexed data (~500K cells) and identified a cellular phenotypic heterogeneity tree consisting of 13 distinct CP's (leaves) (Figure 3a). We characterize each CP with a unique biomarker signature to describe the intrinsic biological properties (Figure 3b). For each CP, the biomarkers are classified as either "ON", "OFF", or "Non-Specific" ("NS") (see Methods, Figure S7). With the biomarker signatures, we identify a heterogeneous population of epithelial (CP1-7) and stromal (CP8-13) cell types. We can further classify some phenotypes as tumor cells (CP1-3), cancer stem cells (CP5), cancer associated fibroblasts (CP9, CP12), macrophages (CP8), and immune cells (CP10, 13) based on specific indicative biomarkers (Table S2). A small subset of non-specialized cells (7% of ~500K) are dispersed across the 13 CP's identified.

To visualize the spatial complexity, we assign cells a CP identify based on the highest ownership probability and systematically select representative patients from each outcome-stage based cohort (e.g., NED-Stage I, NED-Stage II, … etc – see Methods for selection process) (Figure 3c). As expected, each tissue sample is comprised of a heterogeneous population of CP's. In addition to applying LEAPH to the entire patient cohort, we can apply LEAPH to sub-cohorts of patients (particularly based on diagnosis: Stage I, II, III). We do not find a relationship between each stage-specific set of CP's and recurrence-outcome, but we did notice a difference visually in the spatial distributions. In the next section, we further pursue the spatial relationships between the 13 LEAPH-derived CP's within each tissue sample to time-to-recurrence and find that the spatial distributions of CP-pairs are statistically significant in relation to time-to-recurrence.

We demonstrate a heterogeneous neighborhood of cells with examples of the ownership probabilities for specialized, transitional, and multi-transitional cells (Figure 3e). The transitional cells in this neighborhood hold ownership probabilities spread across an epithelial phenotype (CP6) and tumor cell phenotype (CP2) which could indicate a group of epithelial cells gaining tumor cell properties as the disease progresses. The multi-transitional cell shown in this example holds ownership probabilities spread across an epithelial phenotype (CP6) and two

tumor cell phenotypes (CP2, CP3). The multi-transitional cells could be further classified in the future by exploiting the cyclical, non-destructive process of labeling, imaging, and quenching with additional biomarkers on the same tissue section (see also Figure 5).

LEAPH derived computational phenotypes enable the discovery of recurrence-associated microdomains with pointwise mutual information and determine the microdomain-specific network biology

Visually, each tissue sample is composed of a unique pattern of computational phenotypes (from LEAPH). We hypothesize that the patient cohorts with the greatest differences in outcomes will also have distinct spatial heterogeneity patterns. To discover these spatial patterns, we pooled two different cohorts: 1) patients who exhibit recurrence under 3 years (REC-3yrs, N=46) and 2) patients who exhibit no-evidence-of-disease beyond 8 years (NED-8yrs, N=45) (Table S3, see Methods). We use pointwise mutual information (PMI) to quantify the spatial co-occurrence of each CP-pair within each tissue sample. The resulting PMI values are either negative, zero, or positive indicating if a CP-pair spatially co-occurs below, the same, or above average compared to a random background distribution (Figure 4a, see Methods for details).

For each CP-pair, we aggregate the PMI values to form NED-8yrs and REC-3yrs distributions (Figure S8). We use the Wilcoxon rank sum test and Kendall-rank correlations (and p-values) to determine which CP-pairs have a significant difference in spatial co-occurrence between the NED-8yrs and REC-3yrs patients (Table S4, Methods). We found 5 CP-pairs whose PMI distributions are statistically different between the NED-8yrs and REC-3yrs patient cohorts (Figure 4b). The 5 CP-pairs consist of 6 CP's and form two microdomains (Figure 4c). We found two microdomains more prevalent in the REC-3yrs cohort (Figure 4c). Microdomain 1 ($\mathcal{M}_1$) consists of a self-paired tumor cell phenotype (CP1) and microdomain 2 ($\mathcal{M}_2$) is comprised of an interconnected network between an immune cell phenotype (CP13), cancer stem cell phenotype (CP5), cancer associated fibroblasts (CP9), and two epithelial phenotypes (CP6, CP7) (Figure 4c).

We report the Kendall-rank correlation between the PMI values and time-to-recurrence in Figure 4c as well as the p-value and results of the Wilcoxon rank sum test in Table S4. In addition, we perform the same statistical analysis using the clinical variables (stage, grade, sex, age) and CP fractions per tissue sample. We found that the spatial distributions (PMI covariates) are

statistically more significant in relation to time-to-recurrence than the CP fractions per tissue sample and come close to the correlation between stage/grade with time-to-recurrence.

We further utilize partial correlation network analysis to investigate the microdomain-specific network biology driving the spatial contingency within each microdomain to recurrence outcome (see Methods). We report the 99th percentile of absolute partial correlation values over all biomarker pairs (see Methods, Table S5). In Figure 4d, we visualize the outcome-specific significant biomarker networks, highlighting the biomarker pairwise interactions unique to NED-8yrs (blue) and REC-3yrs (red) for each microdomain. Additionally, the biomarkers are color-coordinated based on the properties and functions in Table S2 and Figure S2.

The differential expression of microdomains, $\mathcal{M}_1$ and $\mathcal{M}_2$, between cohorts, NED-8yrs and REC-3yrs, and the biomarker pairwise interactions unique to the REC-3yrs cohort from both microdomains ($\mathcal{M}_1$ – EZH2::MSH2, pERK::CD31, pEGFR::CD20, PTEN::CD68, Wnt5a::CD79, CollagenIV::SMA, $\mathcal{M}_2$ – E-cadherin::beta-catenin, EZH2::MSH2, p53::ALDH1, among others in Table S5) is consistent with the existence of a tumor mediated immunosuppression program driven by cancer stem cells through an enhanced Wnt signaling pathway [33-45] (Figure 4d). The crucial oncogenic Wnt signaling pathway plays a vital role in cancer stem cell maintenance and is tightly coupled with immunosuppression in CRC and other cancer types [33, 34, 36, 37, 40-42, 46, 47]. Our unsupervised approach to computational cellular phenotyping and spatial analytics derived microdomains leads to results that independently corroborate with previous experimental studies.

Optimal selection of biomarkers for reproducing phenotypic diversity

For hyperplexing platforms that are non-destructive of the tissue and iterative in labeling biomarkers, such as the Cell DIVE platform, there is an option of bringing in biomarkers on demand. In other words, instead of using 51 biomarkers upfront, can we select an optimal list of biomarkers much smaller in number that we can apply iteratively to reveal the phenotypic diversity?

To test this hypothesis, we performed a virtual simulation of systematically introducing biomarkers into the data and applying LEAPH at the first (epithelial-stromal dissection) and second (epithelial- and stromal- subtyping dissection) levels of the hierarchy (see Methods). For comparison, we use the maximum ownership probability from the ALL-DATA results with 51 biomarkers (Figure 3) as the ground truth CP identity. We measure the accuracy as the percentage of cells with a matching CP identity to the ground truth.

At the first level, only 2 biomarkers (pck26, E-cadherin) are needed to reproduce the cellular phenotypic assignments with an accuracy above 97% (Figure S9). Increasing the number of biomarkers further increases the accuracy to almost 99% (Figure S9). This analysis demonstrates that the epithelial-stromal dissection is a low-dimensional dissection, as suspected.

At the second level of the hierarchy, both the epithelial and stromal subtyping dissections require a larger set of biomarkers than level one (Figure S9). The epithelial subtyping dissection reaches 89% accuracy with 4 biomarkers (4EBP1, pS6, pMAPKAPK2, PCNA) (Figure S9). An increase in biomarkers does not make an overall substantial difference in the reproduction accuracy. The stromal subtyping dissection reaches above 93% accuracy with 4 biomarkers (Lamin A/C, Claudin 1, Akt, 4EBP1). Contrasting from the epithelial subtyping dissection, the addition of more biomarkers leads to a convergence to almost 97% accuracy with 8 or more biomarkers (Figure S9).

We conclude that there exists an optimal subset of biomarkers at each LEAPH dissection. This virtual simulation demonstrates the capabilities of LEAPH to exploit the non-destructive and iterative nature of the Cell DIVE platform with a parsimonious selection of biomarkers to re-generate identical computational phenotypes (Figure 5).

Discussion

We have proposed LEAPH, an unsupervised, spatially informed, probabilistic recursive clustering method, to exploit the spatial ITH properties prominent in many cancer types and other diseases. The unsupervised and spatially informed approach using biomarker data enables the tumor architecture to drive the discovery of computational phenotypes. Additionally, rather than limiting cells to identify with a single cell type, we exploit the cellular phenotypic continuum with probabilistic clustering and spatial regularization to further classify cell states as specialized or non-specialized. Biologically, the non-specialized cells (transitional and multi-transitional) represent cells undergoing a transformation (e.g., epithelial-mesenchymal-transition, cell-fusion) between CPs. LEAPH does not computationally identify the direction of the transition.

LEAPH is built as a generalizable model amenable to other data sources, soft clustering algorithms, and spatial regularization objective functions/optimization methods. The input criteria for LEAPH is single cell data vectors with corresponding spatial information per sample. Other sources for spatial multiparameter cellular and sub-cellular imaging data include transmitted

light (H&E and IHC), fluorescence, immunofluorescence, live cell biomarkers, mass spectrometry, spatial transcriptomics, electron microscopy etc. The soft clustering step of LEAPH is amenable to other probabilistic mixture models such as but not limited to gaussian mixture models [48] and mixtures of probabilistic principal component analysis [49]. An argument could be made that a low-dimensional model will be scalable with a larger number of biomarkers. In addition to the choice of a low-dimensional model, there is room for improvements to better estimate the noise model based on the data generation method. For the spatial regularization step, the objective function in place is amenable to other optimization methods (e.g., gradient descent) and the objective function is adjustable to other sources of input data.

Beyond LEAPH, we have also laid out a framework utilizing the LEAPH derived CPs to investigate the systems biology of TMEs in relation to disease progression. We use PMI to compute the pairwise relative likelihood of spatial co-occurrence between CPs and subsequently form microdomains based on the recurrence-specific CP pairs. We hypothesize that capturing the higher order spatial relationships between groups of phenotypes may provide additional information on the spatial configurations of the microdomains. Further, there are many graphical models [50-52] that could also be used in the microdomain-specific network biology analyses.

In this study, we applied LEAPH on hyperplexed immunofluorescence images of CRC primary tumor tissue samples and identify 13 data-driven CP's comprised of specialized, transitional, and multi-transitional cell states. We characterized each CP with a biomarker signature and concluded properties demonstrating the heterogeneity of epithelial and stromal cell types. We further used PMI to discover two microdomains with spatial distributions significantly different between patients in the NED-8yrs and REC-3yrs cohorts. The microdomains found suggest that the heterotypic cellular interaction drives the aggressiveness of the tumor. Specifically, microdomain 2 is an interconnected network consisting of phenotypes with properties of immune cells, cancer associated fibroblasts, cancer stem cells, and heterogeneous epithelial cells. Further, with a microdomain-specific network biology approach, we investigated the biomarker interactions within the microdomains and find an emergence of a Wnt-signaling based immunosuppressive program driven by cancer stem in the REC-3yrs patient cohort [33-45]. Our computational findings corroborate with previous experimental studies thus validating our unsupervised approach to computational phenotyping and spatial analytics.

With this novel framework, our results have shown evidence consistent with known properties a heterocellular cross talk between cancer stem cells, cancer associated fibroblasts, and immune cell subtypes (T-cells, B-cells, macrophages) driven by an enhanced Wnt signaling [33-45]. To further investigate our findings, we can exploit the non-destructive property of Cell DIVE™ like imaging platforms to label, image, and quench additional biomarkers, such as T-cell specific CD4 and CD8 to study the T-cell exclusion mechanism, on the same tissue section (Figure 5). Specifically, there is an increase in evidence that suggests the inhibition of Wnt signaling that could reinstate immunosurveillance and provide a favorable therapeutic outcome [35, 40, 53].

Here, we denoted the leaves of the tree generated by LEAPH as our CPs. However, information at any level of the LEAPH hierarchy can be used for the spatial analysis described in this paper. In fact, in our previous work, we have used only the first level of the LEAPH hierarchy (epithelial and stromal domains) to successfully predict the risk of 5-year recurrence in CRC [26]. Indeed, the same analysis can be repeated by using the entire library of CPs derived from LEAPH. We expect the predictive analysis will be further enhanced with the additional phenotypes and the derived microdomains.

The selection of biomarkers is a key factor in determining the resulting phenotypes that will emerge. We demonstrated the existence of an optimal set of biomarkers capable of reproducing the CPs derived from LEAPH with the entire biomarker set. Our virtual simulation provides evidence for performing iterative cycles of imaging and computational analysis with an optimal biomarker set to fully exploit the capabilities of LEAPH in combination with a non-destructive multi to hyperplexed imaging platform.

LEAPH bridges an important knowledge gap in the analytical frameworks that we previously proposed [25, 26]. In particular, LEAPH builds s statistical framework for the application of PMI [25] in the unsupervised discovery of data-driven microdomains. With this advancement, we can fully exploit the spatial network biology based analysis from our previous work [26] to provide mechanistic insights into disease biology such as the generation of outcome-specific new pathways, new cellular phenotypes, and iterative experimental testing of additional biomarkers of the same microdomains (Figure 5). We can generate hypothesis for future experimental and computational studies to further investigate the signaling and cross talk of this immunosuppressive program across the TME landscape to understand the biology of CRC recurrence and possible therapeutic strategies. Systematically deciphering the microdomain-specific network biology will allow us to not only a priori predict recurrence and its aggressiveness resulting in personalized patient surveillance but also potentially select optimal

therapeutic interventions by identifying dominant risk-associated microdomains (Figure 5). This framework forms the basis of an explainable AI platform [54] with applications probing and modulating tumor environment including prognostics, diagnostics, patient stratification for clinical trials, drug target identification, patient-specific therapeutic strategies including immunotherapy, as well as animal toxicology studies (Figure 5).

Acknowledgements

This work was supported in part by the grants NIH/NCI U01CA204836 (SCC, DLT); NIH/NIBIB 5T32EB009403 and 1NIH/NCI 1F31CA254332 (SF); NIH P30CA047904, and PA DHS 4100054875 (DLT); NIH/NHGRI U54HG008540, and UPMC Center for Commercial Applications of Healthcare Data 711077 (SCC). The authors are very grateful to Dr. Fiona Ginty and her team members at GE Global Research, including Christopher J. Sevinsky, Yousef Al-Kofahi and Brion Sarachan, for their collaborative support and for sharing the hyperplexed colorectal cancer dataset generated on Cell DIVE™ from their previous study [18].

Competing Interests:

S.C. C. and D.L.T. have ownership interest in SpIntellx, Inc., a computational and systems pathology company. S.C.C is on a leave of absence from the University of Pittsburgh. The other authors disclose no potential conflicts of interest. The computational and systems pathology intellectual property is owned by the University of Pittsburgh and is exclusively licensed to SpIntellx Inc., Pittsburgh, PA.

Author contributions

Conceived and designed the study – S.A.F., A.M.S., D.L.T., F.P., S.C.C.

Wrote the paper – S.A.F., A.M.S., D.L.T., F.P., S.C.C.

Developed computational platform – S.A.F., F.P., S.C.C.

Performed formal analysis – S.A.F., F.P., S.C.C.

Developed systems biology framework – S.A.F., A.M.S., F.P., S.C.C.

Data quality control and providing analysis support – S.A.F., S.U., F.P., S.C.C.

Technical issues and edited and reviewed the paper – S.A.F., A.M.S., S.U., D.L.T., F.P., S.C.C.

References

1. Almendro, V., A. Marusyk, and K. Polyak, *Cellular heterogeneity and molecular evolution in cancer.* Annu Rev Pathol, 2013. **8**: p. 277-302.
2. Bussard, K.M., et al., *Tumor-associated stromal cells as key contributors to the tumor microenvironment.* Breast Cancer Res, 2016. **18**(1): p. 84.
3. Caswell, D.R. and C. Swanton, *The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome.* BMC Med, 2017. **15**(1): p. 133.
4. DeGregori, J., *Connecting Cancer to Its Causes Requires Incorporation of Effects on Tissue Microenvironments.* Cancer Res, 2017. **77**(22): p. 6065-6068.
5. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
6. Roerink, S.F., et al., *Intra-tumour diversification in colorectal cancer at the single-cell level.* Nature, 2018. **556**(7702): p. 457-462.
7. Shia, J., et al., *Morphological characterization of colorectal cancers in The Cancer Genome Atlas reveals distinct morphology-molecular associations: clinical and biological implications.* Mod Pathol, 2017. **30**(4): p. 599-609.
8. Tape, C.J., *The Heterocellular Emergence of Colorectal Cancer.* Trends Cancer, 2017. **3**(2): p. 79-88.
9. Tauriello, D.V.F. and E. Batlle, *Targeting the Microenvironment in Advanced Colorectal Cancer.* Trends Cancer, 2016. **2**(9): p. 495-504.
10. Tauriello, D.V., et al., *Determinants of metastatic competency in colorectal cancer.* Mol Oncol, 2017. **11**(1): p. 97-119.
11. Balkwill, F.R., M. Capasso, and T. Hagemann, *The tumor microenvironment at a glance*. 2012, The Company of Biologists Ltd.
12. Junttila, M.R. and F.J. de Sauvage, *Influence of tumour micro-environment heterogeneity on therapeutic response.* Nature, 2013. **501**(7467): p. 346-354.
13. Smith, E.A. and H.C. Hodges, *The Spatial and Genomic Hierarchy of Tumor Ecosystems Revealed by Single-Cell Technologies.* Trends in cancer, 2019.
14. Chen, H., et al., *Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM.* Nat Commun, 2019. **10**(1): p. 1903.
15. Efremova, M. and S. Teichmann, *Computational methods for single-cell omics across modalities.* Nature methods, 2020. **17**(1): p. 14.
16. Lundberg, E. and G.H.H. Borner, *Spatial proteomics: a powerful discovery tool for cell biology.* Nat Rev Mol Cell Biol, 2019. **20**(5): p. 285-302.

17. Gough, A., et al., *High content analysis with cellular and tissue systems biology: a bridge between cancer cell biology and tissue-based diagnostics.* The molecular basis of cancer, 2014. **4**.
18. Gerdes, M.J., et al., *Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue.* Proc Natl Acad Sci U S A, 2013. **110**(29): p. 11982-7.
19. Goltsev, Y., et al., *Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging.* Cell, 2018. **174**(4): p. 968-981 e15.
20. Giesen, C., et al., *Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry.* Nature methods, 2014. **11**(4): p. 417-422.
21. Amir, E.-a.D., et al., *viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia.* Nature biotechnology, 2013. **31**(6): p. 545-552.
22. DiGiuseppe, J.A., et al., *PhenoGraph and viSNE facilitate the identification of abnormal T-cell populations in routine clinical flow cytometric data: PHENOGRAPH AND VISNE FACILITATE THE IDENTIFICATION OF ABNORMAL T CELLS.* Cytometry Part B: Clinical Cytometry, 2018. **94**(5): p. 744-757.
23. Chen, H., et al., *Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline.* PLoS computational biology, 2016. **12**(9): p. e1005112.
24. van der Maaten, L.J.P. and G.E. Hinton, *Visualizing High-Dimensional Data Using t-SNE.* Journal of Machine Learning Research, 2008. **9**(nov): p. 2579-2605.
25. Spagnolo, D.M., et al., *Pointwise mutual information quantifies intratumor heterogeneity in tissue sections labeled with multiple fluorescent biomarkers.* J Pathol Inform, 2016. **7**: p. 47.
26. Uttam, S., et al., *Spatial domain analysis predicts risk of colorectal cancer recurrence and infers associated tumor microenvironment networks.* Nature communications, 2020. **11**(1): p. 1-14.
27. Gerdes, M.J., et al., *Emerging understanding of multiscale tumor heterogeneity.* Frontiers in oncology, 2014. **4**: p. 366.
28. Graf, J.F. and M.I. Zavodszky, *Characterizing the heterogeneity of tumor tissues from spatially resolved molecular measures.* PLoS One, 2017. **12**(11): p. e0188878.
29. McKinley, E.T., et al., *Optimized multiplex immunofluorescence single-cell analysis reveals tuft cell heterogeneity.* JCI insight, 2017. **2**(11).

30. Nelson, D.A., et al., *Quantitative single cell analysis of cell population dynamics during submandibular salivary gland development and differentiation.* Biology open, 2013. **2**(5): p. 439-447.
31. Santamaria-Pang, A., et al. *Robust single cell quantification of immune cell subtypes in histological samples*. in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE.
32. Chen, Z., et al., *Modeling Multiplexed Images with Spatial-LDA Reveals Novel Tissue Microenvironments.* Journal of Computational Biology, 2020.
33. Augustin, I., et al., *Immune cell recruitment in teratomas is impaired by increased Wnt secretion.* Stem cell research, 2016. **17**(3): p. 607-615.
34. Bienz, M. and H. Clevers, *Linking colorectal cancer to Wnt signaling.* Cell, 2000. **103**(2): p. 311-320.
35. De Jaeghere, E.A., H.G. Denys, and O. De Wever, *Fibroblasts Fuel Immune Escape in the Tumor Microenvironment.* Trends in cancer, 2019.
36. Galluzzi, L., et al., *WNT signaling in cancer immunosurveillance.* Trends in cell biology, 2019. **29**(1): p. 44-65.
37. Grasso, C.S., et al., *Genetic mechanisms of immune evasion in colorectal cancer.* Cancer discovery, 2018. **8**(6): p. 730-749.
38. Kraman, M., et al., *Suppression of antitumor immunity by stromal cells expressing fibroblast activation protein–α.* Science, 2010. **330**(6005): p. 827-830.
39. Naito, Y., et al., *CD8+ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer.* Cancer research, 1998. **58**(16): p. 3491-3494.
40. Pai, S.G., et al., *Wnt/beta-catenin pathway: modulating anticancer immune response.* Journal of hematology & oncology, 2017. **10**(1): p. 101.
41. Spranger, S. and T.F. Gajewski, *A new paradigm for tumor immune escape: β-catenin-driven immune exclusion.* Journal for immunotherapy of cancer, 2015. **3**(1): p. 1-3.
42. Spranger, S. and T.F. Gajewski, *Impact of oncogenic pathways on evasion of antitumour immune responses.* Nature Reviews Cancer, 2018. **18**(3): p. 139-147.
43. Turley, S.J., V. Cremasco, and J.L. Astarita, *Immunological hallmarks of stromal cells in the tumour microenvironment.* Nature reviews immunology, 2015. **15**(11): p. 669-682.
44. Wong, P.F., et al., *Multiplex quantitative analysis of cancer-associated fibroblasts and immunotherapy outcome in metastatic melanoma.* Journal for immunotherapy of cancer, 2019. **7**(1): p. 194.

45. Yaeger, R., et al., *Clinical sequencing defines the genomic landscape of metastatic colorectal cancer.* Cancer cell, 2018. **33**(1): p. 125-136. e3.
46. Fu, C., et al., *β-Catenin in dendritic cells exerts opposite functions in cross-priming and maintenance of CD8+ T cells through regulation of IL-10.* Proceedings of the National Academy of Sciences, 2015. **112**(9): p. 2823-2828.
47. Gattinoni, L., et al., *Wnt signaling arrests effector T cell differentiation and generates CD8+ memory stem cells.* Nature medicine, 2009. **15**(7): p. 808-813.
48. Reynolds, D.A., *Gaussian Mixture Models.* Encyclopedia of biometrics, 2009. **741**.
49. Tipping, M.E. and C.M. Bishop, *Mixtures of probabilistic principal component analyzers.* Neural computation, 1999. **11**(2): p. 443-482.
50. Barber, R.F. and M. Drton, *High-dimensional Ising model selection with Bayesian information criteria.* Electronic Journal of Statistics, 2015. **9**(1): p. 567-607.
51. Foygel, R. and M. Drton. *Extended Bayesian information criteria for Gaussian graphical models*. in *Advances in neural information processing systems*. 2010.
52. Meinshausen, N. and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso.* The annals of statistics, 2006. **34**(3): p. 1436-1462.
53. Chen, D.S. and I. Mellman, *Oncology meets immunology: the cancer-immunity cycle.* Immunity, 2013. **39**(1): p. 1-10.
54. Tosun, A.B., et al., *Explainable AI (xAI) for Anatomic Pathology.* Advances in Anatomic Pathology, 2020. **27**(4): p. 241-250.
55. https://www.gelifesciences.com/en/us/shop/cell-imaging-and-analysis/high-content-analysis-systems/instruments/cell-dive-multiplexed-imaging-p-11361.
56. Ghahramani, Z. and G.E. Hinton, *The EM algorithm for mixtures of factor analyzers*. 1996, Technical Report CRG-TR-96-1, University of Toronto.
57. Francis, K. and B.O. Palsson, *Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion.* Proceedings of the National Academy of Sciences, 1997. **94**(23): p. 12258-12262.
58. Hallac, D., J. Leskovec, and S. Boyd. *Network lasso: Clustering and optimization in large graphs*. in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015. ACM.
59. Leinster, T. and C.A. Cobbold, *Measuring diversity: the importance of species similarity.* Ecology, 2012. **93**(3): p. 477-489.
60. Wasserman, L., *All of nonparametric statistics*. 2006: Springer Science & Business Media.

61. Efron, B. and T. Hastie, *Computer age statistical inference*. Vol. 5. 2016: Cambridge University Press.
62. Zeng, X., Y. Xia, and H. Tong, *Jackknife approach to the estimation of mutual information.* Proc Natl Acad Sci U S A, 2018. **115**(40): p. 9956-9961.
63. Vickovic, S., et al., *High-definition spatial transcriptomics for in situ tissue profiling.* Nature methods, 2019. **16**(10): p. 987-990.

Methods

Hyperplexed data

The data for this study consists of 747 colorectal carcinoma (CRC) tissue samples hyperplexed using cell DIVE with 56 biomarkers measured in protein expressions plus DAPI nuclear counterstain. Cell Dive (GE Life Sciences, Issaquah, WA) [18] involves non-destructive cyclical immunofluorescence labeling with two or three antibodies labeled with distinct fluorescent probes, imaging and subsequent quenching of the fluorescence. This process is repeated to capture all the required antibodies (biomarkers). The data consists of image stacks taken at each region of interest and overall image stack consists of several images for each of several imaging rounds. Each round includes a nuclear (DAPI) image that is used as a reference for registering all the images from all the rounds. Quantitation of images in each round includes the fluorescence intensity of each measured biomarker. Images are also acquired after quenching rounds for the purpose of autofluorescence removal [18, 55]. Processing of Cell DIVE images includes correction for uneven illumination across the field of view, removal of autofluorescence, registration, and automated quality control (QC) detection of several categories of defects, including failed registration, blurred or saturated images, and other imaging issues. The images and data undergo a series of tissue and cell quality checks, log2 transformation and normalization steps. To integrate data from batch processing each biomarker is normalized to a control median. Validation of this process shows robustness and preservation of biomarker stability and biological integrity [18, 27-31]. Images are acquired in TIFF format, while image metadata is captured in files having a simple structure that captures the provenance of which images were derived from which slides and characteristics of the acquisition [18, 55]. The biomarkers chosen are protein markers for specific cell lineages, oncogenes, tumor suppressors, and post-translational protein modifications indicative of cellular activation states (Table S2, Figure S2). The data also includes corresponding clinical information including the histological tumor grade, cancer stage, gender, age, and follow up monitoring for 10 years (Table S3).

Data pre-processing

<u>Cell quantitation</u>

Cellular segmentation is done using a collection of structural biomarkers: NaKATPase (cell membrane, border), S6 (cytoplasm), and DAPI (nucleus) (Figure S1). Protein expression and standard deviation were quantified by the median biomarker intensity value within each cell

mask and transformed to the log2 scale [18]. Cells are filtered using individual QC scores generated for each cell (scores less than 0.7-0.8 will not be included indicating inaccurate registration, misalignment, or tissue loss) and based on number of pixels per segmented subcellular compartment.

Patient selection

For this study, based on the clinical data, we limit the patient data set to deceased patients with recurrence within 5 years (postsurgical resection) and alive patients with no evidence of recurrence within 5 years (post-surgical resection). Further, we eliminate tissue samples with less than a threshold of 1000 cells to limit the potential adverse effects of hyperplex imaging (i.e. damaged, folded, or lost tissue). This cell threshold is computed based on the 20th percentile of number of cells per tissue sample shown in Figure S3**.** The final data set used is composed of 213 TMA spots (Table S3). The alternative strategies could have been to filter out cells within the damaged areas of the TMA's by examining the nuclear-to-cytosolic ratio of structural biomarkers. We chose to use the cell threshold method to preserve the unsupervised nature of LEAPH.

Biomarker selection and distributions

We removed biomarkers showing batch affects resulting in a selection of 51 biomarkers for this analysis. The distribution of each biomarker (log2 scale) across all cells in the patient cohort is shown in Figure S4. The Kurtosis values measures the skewness of the distribution. Comparing the kurtosis values of the biomarker distributions to the kurtosis of a univariate normal distribution (kurtosis = 3), many but not all these distributions can be considered to have a Gaussian shape.

LEAPH: unsupervised learning algorithm for cellular phenotypic heterogeneity

We will describe the hyperplexed dataset in a high-dimensional space, where each cell $\vec{x}\ (p \times 1)$ is described by a p dimensional vector of biomarker expressions quantitated appropriately. Further, we assume that the hyperplexed dataset has an intrinsic low-dimensional representation. We will use a mixture of factor analyzers described by low-dimensional factor loadings ($\Lambda\ (p \times k)$), latent variables ($\vec{z}\ (k \times 1)$), mean vector ($\vec{\mu}\ (p \times 1)$), and noise term ($\vec{\nu}\ (p \times 1)$): $\vec{x} = \Lambda\vec{z} + \vec{\mu} + \vec{\nu}$, where p is the number of biomarkers and k is the low-dimension latent space [56]. The latent factors, $\vec{z}$, are generated from zero-mean, unit-variance Normal distribution $N(0, \mathrm{I})$, and the noise term, $\vec{\nu}$ is sampled from $N(0, \Psi)$. I is the unit variance and Ψ is

assumed to be a diagonal matrix. With this construction, $\vec{x}$ is distributed with zero mean and covariance $\Lambda\Lambda^T + \Psi$ [56].

Soft clustering

Typically, cellular phenotyping methods are constructed under the assumption that each cell belongs to one and only one cluster (hard clustering) leaving no room to identify specific cells that may belong to more than one phenotype due to an existing phenotypic continuum. With a Mixture of Factor Analyzers (MFA), we model the cells as M components (clusters) with the parameters $\left(\{\pi_j, \vec{\mu}_j, \Lambda_j\}_{j=1}^{M}, \Psi\right)$ where π_j is the component weight: $p(\vec{x}) = \sum_{j=1}^{M} N\left(\vec{x} \middle| \vec{\mu}_j, \Lambda_j\Lambda_j^T + \Psi\right)$. We chose a two-dimensional latent space for each component in the MFA model, as we observed this is enough to capture the input variance (Figure S5). The expectation-minimization (EM) algorithm is utilized to estimate the model parameters [56]. The EM algorithm is initialized with a random set of parameters and the EM algorithm is not guaranteed to converge to a globally optimal solution. To account for this and ensure stability, we perform a hundred different EM optimizations, each initialized randomly. Each optimization yields an MFA model with a set of model parameters. We compute the biomarker ranking for each set of model parameters (see discriminative biomarkers section) and aggregate all biomarker rankings to compute their mean ranking. The model with a biomarker ranking closest (Euclidean distance) to the mean ranking is selected as the consensus model and deemed to provide an optimal subspace representation (Figure S6). The MFA model results in soft clustering probabilities (*ownership probabilities)* – each cell, x_c, holds a unique probability of belonging to each cluster j, denoted as Ω_{cj}.

Spatial regularization

The soft clustering is agnostic to the spatial complexity of the TME, a key component driving ITH. Based on properties of spatial ITH and the spatial tissue architecture of a tumor, we expect neighborhoods of cells to be spatially coherent (e.g., epithelial/tumor cells to be surrounded by, or spatially proximal to, other epithelial/tumor cells, but making allowance for the presence of tumor-infiltrating lymphocytes and other stromal cells for example). To promote specialization in cells, we add a spatial regularization component to optimize the ownership probabilities of non-specialized cells. The spatial regularization step optimizes the objective function which consists of two terms: ownership confidence and spatial coherence given by : $\min_{\Omega}\left(-\sum_{i=1}^{N}\sum_{j=1}^{M}\Omega_{cj}\log_2\left(\Omega_{cj}\right) + \lambda\sum_{\{j,k\}} w_{jk}\left\|\Omega_j - \Omega_k\right\|_2\right)$. The first term minimizes the entropy

of the ownership probabilities promoting specialization in cells. The second term promotes spatial coherence where w_{jk} is the weight between cell i and cell j and is computed as the reciprocal of the distance between two cells: $w_{jk} = \frac{1}{dist(cell_i, cell_j)}$. We use a distance threshold (100 pixels at 0.5 μm/pixel) to eliminate an influence between cells that are too far apart to communicate [57].

The objective function is optimized using Alternating Directions Method of Multipliers (ADMM) [58]. We assume that the probability ownership confidence (term 1) and spatial coherence (term 2) should hold equal weight and therefore compute the tuning parameter, λ, to scale term 2 to the range of term 1: $\lambda = \frac{N_{opt} \times \text{maxEntropy}}{\sum_{\{j,k\}} w_{jk}}$ where N_{opt} is the number of cells being optimized and maxEntropy is the maximum value of the entropy function (=1). Relaxing the assumption that spatial coherence and ownership confidence should hold equal weight in the objective function would lead to a larger parameter space. A higher weight for spatial coherence results in homogenous neighborhoods and a larger set of non-specialized cells. On the contrary, a larger weight for ownership confidence results in the abolishment of all non-specialized cells. We have found stable and consistent results when the tuning parameter represents an equal weighting. Cells can only have neighbors within the same tissue sample and therefore to increase computational speed and efficiency, spatial regularization is performed on each tissue sample independently.

Recursive decomposition

To automate the process of phenotypic discovery, we propose a recursive probabilistic approach where each step dissects the most dominant clusters with $M = 2$ components. At each recursive step, the soft clustering step utilizes a low-dimensional latent space MFA. Subsequently, spatial regularization optimizes the resulting per-cell ownership probabilities to filter false-positive non-specialized cells by promoting ownership confidence and spatial coherence. The resulting parameters (ownership probabilities, Ω_j, mean vector, $\vec{\mu}_j$, factor loadings, Λ_j) for each cluster, j, are passed to the next recursive step to decompose each cluster into further sub-clusters. This process is continued until an attempted cluster split invalidates any one of the following stopping criteria: 1) a resulting cluster takes ownership of $< 1\%$ of cells, or 2) the angle between the mean vectors and factor loading space are both below a given threshold.

Discriminative biomarker order

Each LEAPH split results in two clusters with high dimensional mean vectors $(\vec{\mu}_1, \vec{\mu}_2)$. To determine the discriminative ordering of the biomarkers, we compute and sort the proportional difference for each biomarker j: $\Delta_j = |\vec{\mu}_{1j} - \vec{\mu}_{2j}| * \max(\vec{\mu}_{1j}, \vec{\mu}_{2j})$. The absolute difference of the mean vectors may bias the selection of biomarkers with high biomarker value ranges and therefore, we opt for a proportional difference to place the biomarkers on an even level for comparison.

Interpreting Computational Phenotypes

We characterize each CP with a unique biomarker signature for ease of interpretation. For each biomarker and CP pair two histograms are computed: 1) histogram of biomarker values for all cells (h_A) and 2) histogram of biomarker values for all specialized cells in the CP (h_B). We compare the 20th and 80th percentiles of each histogram to determine the CP-biomarker classification. Each classification is as follows:

Classification	Operator	Interpretation
"ON"	$h_B^{80th} > h_A^{80th}$ AND $h_B^{20th} > h_A^{20th}$	h_B is skewed to the right compared to h_A
"OFF"	$h_B^{20th} < h_A^{20th}$ AND $h_B^{80th} < h_A^{80th}$	h_B is skewed to the left compared to h_A
"NS" (Non-specific)	$h_B^{80th} > h_A^{80th}$ AND $h_B^{20th} < h_A^{20th}$ OR $h_B^{80th} < h_A^{80th}$ AND $h_B^{20th} > h_A^{20th}$	h_B spans across h_A or is confined within h_A

Figure S7 depicts each distribution and 20th/80th percentiles for each CP-biomarker pair to demonstrate how the biomarkers are classified in Figure 3b.

Selecting representative patients for visualizing computational phenotypes

Post LEAPH, we assign cells a CP identity based on the highest ownership probability value. For C CP's, we compute the fraction of each phenotype found in each patient, i, forming a vector $\vec{f_i}$ $(C \times 1)$. When computing this fraction, we only consider the specialized cells (ownership probability >0.95) to avoid the transitional and rare cells biasing our results. For each outcome-stage-based group (e.g., NED-Stage I, NED-Stage II, NED-Stage III), we compute the average phenotype fraction vector, $\vec{f}_{avg}$, and choose the patient closest to the mean to display in Figure 3c (Euclidean distance).

Biomarker selection virtual simulation

For comparison, we assume the LEAPH derived CP's is the ground truth. Based on the rank-ordered biomarkers from each LEAPH recursion step, we re-run LEAPH with the first 2, 4, …, 20 biomarkers at the first and second level of the hierarchy (Table S6). Assigning each cell to one CP based on the highest ownership probability (cell-label), we compute the accuracy of the cell-labels derived from each sub-set of biomarkers to the assumed ground truth (Figure S9). This virtual simulation demonstrates the reproducibility and capabilities to integrate LEAPH into an integrated iterative imaging and computational platform.

Discover recurrence-associated microdomains with pointwise mutual information and determine microdomain-specific network biology

Step 1: Select patient sub-cohorts with recurrence under 3 years and with no evidence of recurrence in 8 years

To study the differences between patients at each extremum and balance the NED/REC cohorts, we prune the patient cohort further based on the time to recurrence. We consider a group of patients with no evidence of disease over 8 years (NED-8yrs, N=45) and a group of patients that exhibit recurrence under 3 years (REC-3yrs, N=46). See Table S3 for the patient statistics.

Step 2: Construct a spatial network of interacting cells in each patient sample

In each tissue sample, we build a spatial network where each node is a cell and the edges connect cells, say m and n, to each other with a weight w_{mn}= 1 if their spatial distance d_{mn} is within a threshold (100 pixels/50 μm, which is the same threshold used in LEAPH spatial regularization), or $w_{mn} = 0$ otherwise. To remove possible artifacts, we prune each network to include only cells within the largest connected component.

Step 3: Use Pointwise Mutual Information (PMI) to quantify spatial co-occurrence between CP-pairs

Pairwise association statistics, specifically pointwise mutual information (PMI), identifies microdomains consisting of one, two, or groups of phenotypes which spatially co-occur more frequently than a given background distribution. PMI maps can characterize associations at the cohort, patient, and tissue sample level and can describe both the local and global spatial heterogeneity scenarios that cannot be captured by well-known methods such as quadratic entropy [59].

Each cell, c, holds an ownership probability vector over the M CP's, $\vec{\Omega}_c$ $(M \times 1)$, such that $\sum_{j=1}^{M} \vec{\Omega}_{cj} = 1$. The pointwise mutual information between two phenotypes, (f_i, f_j), for a given network or network set, s, is defined as:

$$\mathrm{PMI}_s(f_i, f_j) = \log_2\left(\frac{p(f_i^s, f_j^s)}{p(f_i^t)p(f_j^t)}\right)$$

where $p(f_i^s)$ is the probability of phenotype i occurring in a network set s and $p(f_i^t)$ is the background probability distribution of phenotype i (this can be an ensemble of networks or individual networks, see below). For a single patient, the joint probability is computed as:

$$p(f_i^s, f_j^s) = \frac{1}{z}\left(\sum_{(m,n)\in\Phi} w_{mn}\left(\vec{\Omega}_{mf_i}\vec{\Omega}_{mf_j} + \vec{\Omega}_{mf_j}\vec{\Omega}_{nf_i}\right)\right)$$

where ϕ is the set of edges, w_{mn} is the edge weight between cells m and n, and Z is the normalization factor given by:

$$Z = \sum_{i=1}^{M}\sum_{j=i}^{M}\left(\sum_{(m,n)\in\Phi} w_{mn}\left(\vec{\Omega}_{mf_i}\vec{\Omega}_{nf_j} + \vec{\Omega}_{mf_j}\vec{\Omega}_{nf_i}\right)\right)$$

PMI computation results in values with direct implication to spatial co-occurrence. The PMI values are either negative implicating that a phenotype-pair spatially co-occurs less than the background distribution, positive implicating that a phenotype-pair spatially co-occurs more than the background distribution, or zero implicating that a phenotype-pair spatially co-occurs the same as the background distribution.

Step 3a: Select a background distribution

We choose to use a random background distribution to depict the probability each CP-pair spatially co-occurs more, less, or the same as random. To construct the random distribution, we set the probability of each CP to the probability of each CP over all cells: $p(f_i^t) = \frac{1}{N}\sum_{c=1}^{N} \Omega_{ci}$, where N is the total number of cells.

Step 3b: Perform Jackknife estimation

A jackknife estimation is commonly used to remove bias when measuring the dependence between two random variables (e.g., Pearson's correlation coefficient, mutual information) [60-62]. For a given estimator, PMI: $T_n = T(X_1, \ldots, X_n)$ for n samples. Let $T_{(-i)}$ denote the statistic

where the i-th patient is removed. The jackknife bias estimate is defined as: $b_{jack} = (n-1)(\bar{T}_n - T_n)$, where $\bar{T}_n = \frac{1}{n}\sum_i T_{(-i)}$. The bias-corrected estimator is

$$T_{jack} = T_n - b_{jack} = (n)T_n - \left(\frac{n-1}{n}\right)\sum_i T_{(-i)}.$$

The jackknife estimate of standard error is: $\widehat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_i \left(T_{(-i)} - \frac{1}{n}\sum_i T_{(-i)}\right)^2}$.

Step 3c: Filter unstable PMI values

We observed instability in the PMI estimates (including high jackknife estimated standard error) for any given CP-pair, (f_i, f_j) when the number of cells involved in the pairing is low. To remove this bias, we compute the effective number of cells (total ownership probability) in the CP-pair for a patient, p, as the summation of ownership probabilities, $N^p_{(f_i,f_j)}$, over all cells in the network: $N^p_{(f_i,f_j)} = \sum_c^{N^p} \vec{\Omega}_{cf_i} + \vec{\Omega}_{cf_j}$. Aggregating all patients, we use the 25th percentile as a cutoff value and remove all patients with $N^p_{(f_i,f_j)} <$ cutoff. We find that removing these patients also removes the patients with a relatively high standard error.

<u>Step 4: Identify significant CP-pairs (and microdomains) by comparing PMI distributions between recurring and non-recurring cohorts</u>

For each CP-pair, (f_i, f_j), we compute the Kendall rank correlation, τ, (and p-value, p_τ) between $\mathrm{PMI}(f_i, f_j)$ and time to recurrence (reported in days). In addition, we perform a Wilcoxon rank-sum (WRS) test (does not reject the null hypothesis, $h_{WRS} = 0$, or does reject the null hypothesis, $h_{WRS} = 1$ and corresponding p-value p_{WRS}) to determine if the NED-8yrs and REC-3yrs $\mathrm{PMI}(f_i, f_j)$ values are from equivalent distributions. We define a phenotype pair as significant if $|\tau| > prctile(\tau, 99th)$, $p_\tau \leq 0.05$, $h_{WRS} = 1$, and $p_{WRS} \leq 0.05$. Each CP-pair distribution is shown in Figure S8. The Kendall rank correlation and Wilcoxon rank-sum test results are reported for CP-pair (and clinical variables) in Table S4. The 5 CP-pairs identified as significant by our analysis form two microdomains: $\mathcal{M}_1$: CP1-CP1, $\mathcal{M}_2$: CP13-CP9-CP7-CP6-CP5.

<u>Step 5: Perform partial correlation analysis to construct outcome-specific biomarker networks within each microdomain</u>

For each microdomain, we aim to identify a biomarker network representative of the NED-8yrs and REC-3yrs patient groups. Within a single microdomain and each patient group, we gather

all specialized cells and compute the rank partial correlation between each biomarker-pair, controlling for the remaining biomarkers. In addition, a p-value is also computed for testing the hypothesis of no partial correlation against the alternative that there is a nonzero partial correlation. We define a biomarker-pair (edge) as significant if the absolute value of the partial correlation value is above the 99th percentile of all absolute value partial correlation values and the p-value is below 0.01. The partial correlation values (and p-values) for each edge defined as significant in at least one network is reported in Table S5.

Figure captions

Figure 1 - Unsupervised learning algorithm for cellular phenotypic heterogeneity (LEAPH) on hyperplexed colorectal carcinoma tissue microarray dataset

a. A zoomed in region of a Stage II CRC tissue samples (ca. 0.6mm cores). Each subpanel is pseudo-colored with DAPI (blue), and pairs of biomarkers indicated in the subpanel (e.g., subpanel a is pseudo-colored with 4EBP1 (red) and PTEN (green)). We selected a large set of biomarkers (p=51) to test the feasibility of cyclical labeling, imaging, and quenching process of the hyperplexing technology and to provide insights into the CRC biology.

b. LEAPH can perform cellular phenotyping on hyperplexed single sample or a cohort of tissue samples (currently tested up to 213 tissue samples, ~500K cells). LEAPH applies recursive steps of probabilistic clustering with mixtures of factor analyzer (MFA) models and spatial regularization to construct data-driven computational phenotypic trees. We apply a stopping criterion on the recursion, by applying a threshold on the angle between subspaces created by the mixture models, to ensure that the resulting phenotypes are distinct and to avoid overfitting of the input dataset by LEAPH.

c. The terminal nodes of the tree (leaves) signify distinct computational phenotypes (CP) discovered in the input dataset and form the mixture components of the final MFA model determined by the recursive decomposition. We visualize the spatial distribution of phenotypes within a tissue sample by assigning each cell to the phenotype with the highest ownership probability.

Figure 2 – Illustrating the recursive steps of probabilistic clustering with mixtures of factor analyzer models and spatial regularization in LEAPH

a. ***Top left:*** Synthetic data generated to reflect the statistics of the CRC biomarker data. We find that CRC biomarker values have a wide spread likely due to the individual biomarker response sensitivity (e.g., log(ALDH1) values range between 0-12, while log(CD20) values range between 0-4, see Figure S3). Consequently, we observe dominant phenotypes in the form of epithelial and stromal cells and parsing further we discover the various cell subtypes (e.g., immune, STEM cells). To reflect these properties, the synthetic data has two dominant phenotypes each with two additional subtypes and hence four different phenotypes nested in total. ***Top right:*** One tier non-recursive MFA clustering fails at segmenting the four modes in the ground truth synthetic data. ***Bottom left:*** Level 1 of the recursive split identifies the two dominant clusters in

the synthetic data. ***Bottom right:*** Level 2 of the recursive clustering splits each dominant cluster from Level 1 to segment the four clusters from the synthetic data. An example ownership probability vector is shown in each clustering attempt, color coded based on respective cluster. This simulation demonstrates the advantage of performing recursive clustering with MFA rather than a one-tier approach. On the more complex and high-dimensional CRC data, each recursive step of fitting MFA model is repeated 100 times and a consensus model is chosen to eliminate any spurious results that arise from the random cluster initialization.

b. We perform spatial regularization on a single patient sample and track the ownership probabilities for each cell undergoing regularization at each iteration. Majority of cells converge to a specialized cell identity (ownership probabilities >0.95 or <0.05), but a small subset of cells remains in the non-specialized range (ownership probabilities between 0.05 and 0.95)

c. Tissue sample where each cell is outlined with the ownership probability prior- and post-spatial regularization. We specifically point to two cells, each with an ownership probability of 0.8 prior to spatial regularization. Post-regularization, one cell converges to a specialized CP2 (blue) because of the dominance of this CP in its neighborhood (top zoomed box). The second cell remains non-specialized because of the heterogeneity in its neighborhood (bottom zoomed box).

Figure 3 – LEAPH reveals the hierarchical nature of the cellular phenotypic heterogeneity and identifies specialized, transitional, and rare cell populations in the hyperplexed colorectal carcinoma dataset

a. Cellular phenotypic hierarchy derived by applying LEAPH on the entire CRC cohort. The size of each node is proportional to the fraction of cells with majority ownership to that computational phenotype (CP) (size key). Each leaf node represents a data-driven CP determined by the stopping criteria. The set of all leaf nodes, 13 CPs in total, form the components of the *final* mixture of factor analyzers model for the CRC patient cohort.

b. For the ease of interpretation, each CP is characterized by a unique biomarker signature, where the biomarkers are classified as being “ON”, “OFF”, or “Non-specific (NS)” (see Methods, Figure S7). The biomarker signatures enable identifying heterogeneous populations of known cell types: epithelial cells (CP6, CP7), tumor cells (CP1, CP2, CP3), cancer stem cells (CP5), cancer associated fibroblasts (CP9, CP12), stromal cells (CP4), macrophages (CP8), and immune cells (CP10, CP13).

c. Illustrative tissue samples from clinically relevant patient sub-cohorts. Cell boundaries are outlined with the colors of the assigned phenotype (panel a) based on highest ownership probability values.

d. Example of a neighborhood of non-specialized cells. We point to multiple cells within the neighborhood with varying ownership probabilities. This neighborhood depicts a *rare* cell (defined as ownership probabilities spread across more than 2 CPs), *transitional* cells (ownership probabilities spread across 2 CPs) compared to a specialized cell (ownership probability greater than 0.95 for a single CP).

Figure 4 – Discovery of recurrence-associated microdomains with pointwise mutual information and determine the microdomain-specific network biology

a. Pointwise mutual information (PMI) maps are computed for each tissue sample to quantify the spatial co-occurrence of each CP-pair compared to a random background distribution (see Methods). A PMI value above 0 indicates that the CP-pair spatially co-occurs more often than random and below 0 indicates that the CP-pair spatially co-occurs less often than random. The PMI values are normalized to the range -1 to 1 for visualization.

b. The PMI values for each CP-pair are grouped by the outcome data: no-evidence of recurrence in 8 years (NED-8yrs) and recurrence within 3 years (REC-3yrs). We compared the PMI distributions using the Kendall rank correlation coefficients and the Wilcoxon rank sum test. We found 5 CP-pairs to be significant (see Methods, Table S4, Figure S8). Each significant CP-pair has a distribution skewed higher for the REC-3yr group implicating a greater level of spatial co-occurrence compared to a random background distribution.

c. Two microdomains emerge from the significance analysis: $\mathcal{M}_1$– CP1::CP1 microdomain 1 consisting of only CP1 and microdomain: $\mathcal{M}_2$ consisting of a network between CP13, CP5, CP9, CP6, and CP7. In the microdomain visualizations, we report the Kendall rank correlation coefficients on each CP-pair edge (p-values for Kendall rank correlation coefficients and Wilcoxon rank sum test are found in Table S4).

d. Within each spatial microdomain, we identify a recurrence-associated biomarker network through partial correlation analysis (see Methods, Table S5). The biomarkers are grouped based on their presumed cellular functions/processes and edges connecting each biomarker pair is highlighted to signify a common, NED-8yrs unique, or REC-3yrs unique edge (see legend)

Figure 5 – Analytical platform that can be applied to any multi to hyperplexed image data sets.

LEAPH enables the application of a spatial analytics and microdomain-specific network biology-based early discovery and development platform to a multi to hyperplexed imaging platform such as Cell DIVE™ used in this study, as well as other hyperplexed technologies [19, 20, 63]. The pipeline begins with the preparation of patient tissue samples for pathology including labeling with multi to hyperplexed biomarkers, imaging based on multiple imaging modalities (e.g. transmitted light, fluorescence, and mass spec [20]). Basic image processing and basic image quantitation is then followed by data normalization in preparation of applying the analytical platform. The hyperplexed imaging process can be initiated with a limited set of biomarkers from which LEAPH builds data-driven, computationally unbiased phenotypic hierarchy of cell types, cell states capturing a phenotypic continuum, and their spatial configurations. LEAPH in combination with pointwise mutual information (PMI) discovers outcome-specific microdomains composed of spatially configured computational phenotypes. LEAPH combined with microdomain-specific network biology provides mechanistic insights into disease biology. This analysis will further suggest outcome-specific new pathways, new cellular phenotypes and additional biomarkers which can be optimally tested through iterative probing of the same microdomains on non-destructive imaging platforms such as Cell DIVE™ [18]. The spatial analytics and microdomain-specific network biology features form the core components of the explainable AI platform [54] that can drive a wide range of applications probing and modulating tumor environment including prognostics, diagnostics, patient stratification for clinical trials, drug target identification, patient-specific therapeutic strategies (e.g. immunotherapy), and animal toxicology studies.

Figure 1

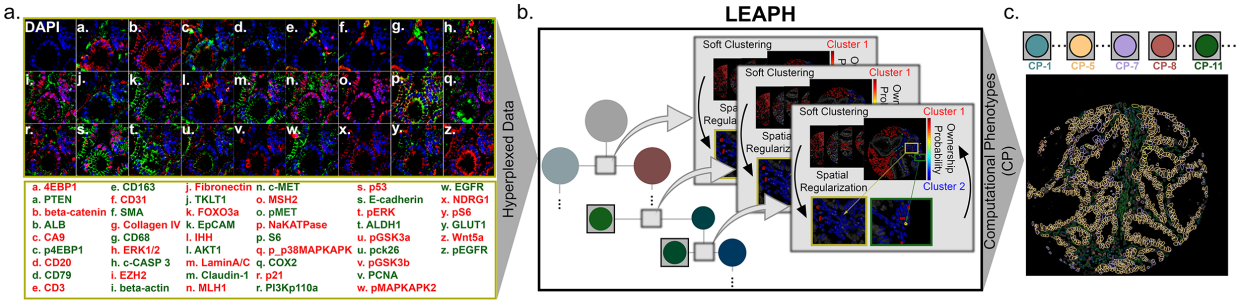


Figure 2

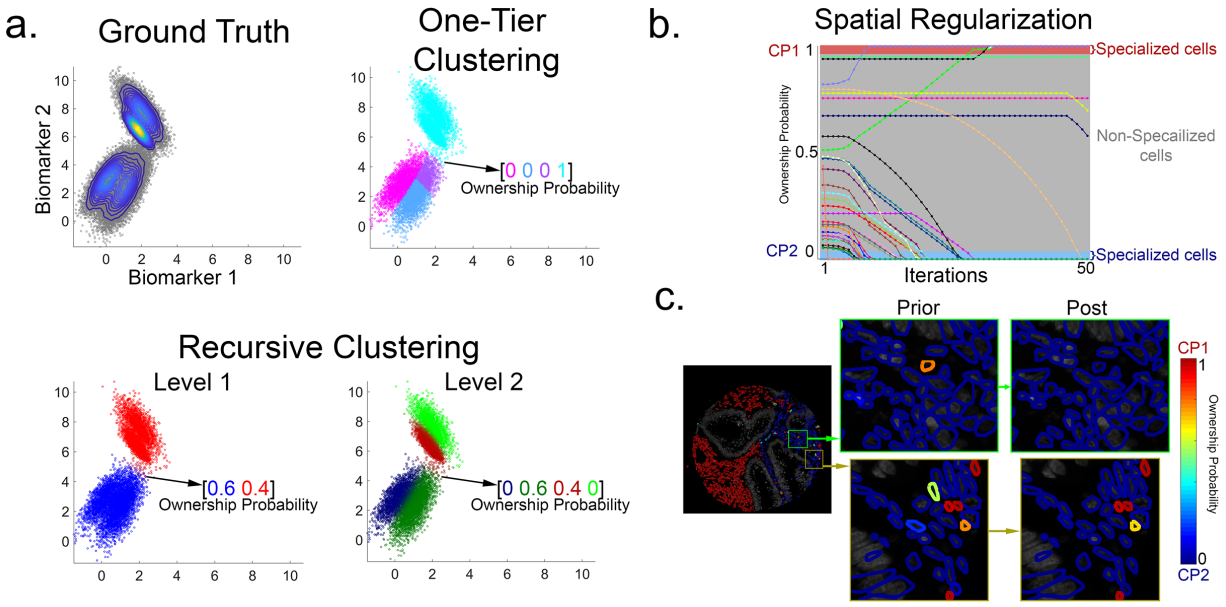


Figure 3

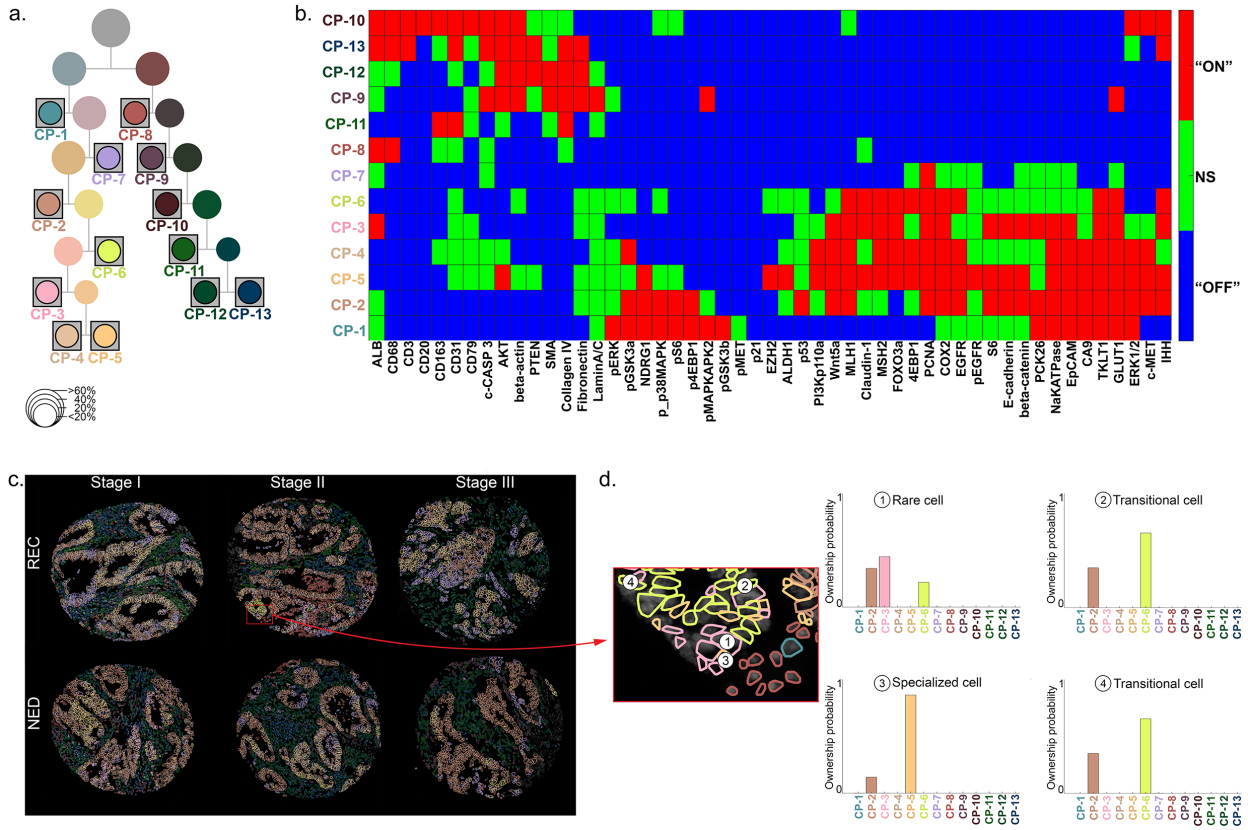

a.
CP-1
CP-8
CP-7
CP-9
CP-2
CP-10
CP-6
CP-11
CP-3
CP-12
CP-13
CP-4
CP-5
>60%
40%
20%
<20%
b.
CP-10
CP-13
CP-12
CP-9
CP-11
CP-8
CP-7
CP-6
CP-3
CP-4
CP-5
CP-2
CP-1
ALB
CD68
CD3
CD20
CD163
CD31
CD79
c-CASP 3
AKT
beta-actin
PTEN
SMA
Collagen IV
Fibronectin
LaminA/C
pERK
pGSK3a
NDRG1
p_p38MAPK
pS6
p4EBP1
pMAPKAPK2
pGSK3b
pMET
p21
EZH2
ALDH1
p53
PI3Kp110a
Wnt5a
MLH1
Claudin-1
MSH2
FOXO3a
4EBP1
PCNA
COX2
EGFR
pEGFR
S6
E-cadherin
beta-catenin
PCK26
NaKATPase
EpCAM
CA9
TKLT1
GLUT1
ERK1/2
c-MET
IHH
"ON"
NS
"OFF"
c.
Stage I
Stage II
Stage III
REC
NED
d.
1 Rare cell
2 Transitional cell
3 Specialized cell
4 Transitional cell
Ownership probability
1
0
CP-1
CP-2
CP-3
CP-4
CP-5
CP-6
CP-7
CP-8
CP-9
CP-10
CP-11
CP-12
CP-13

Figure 4

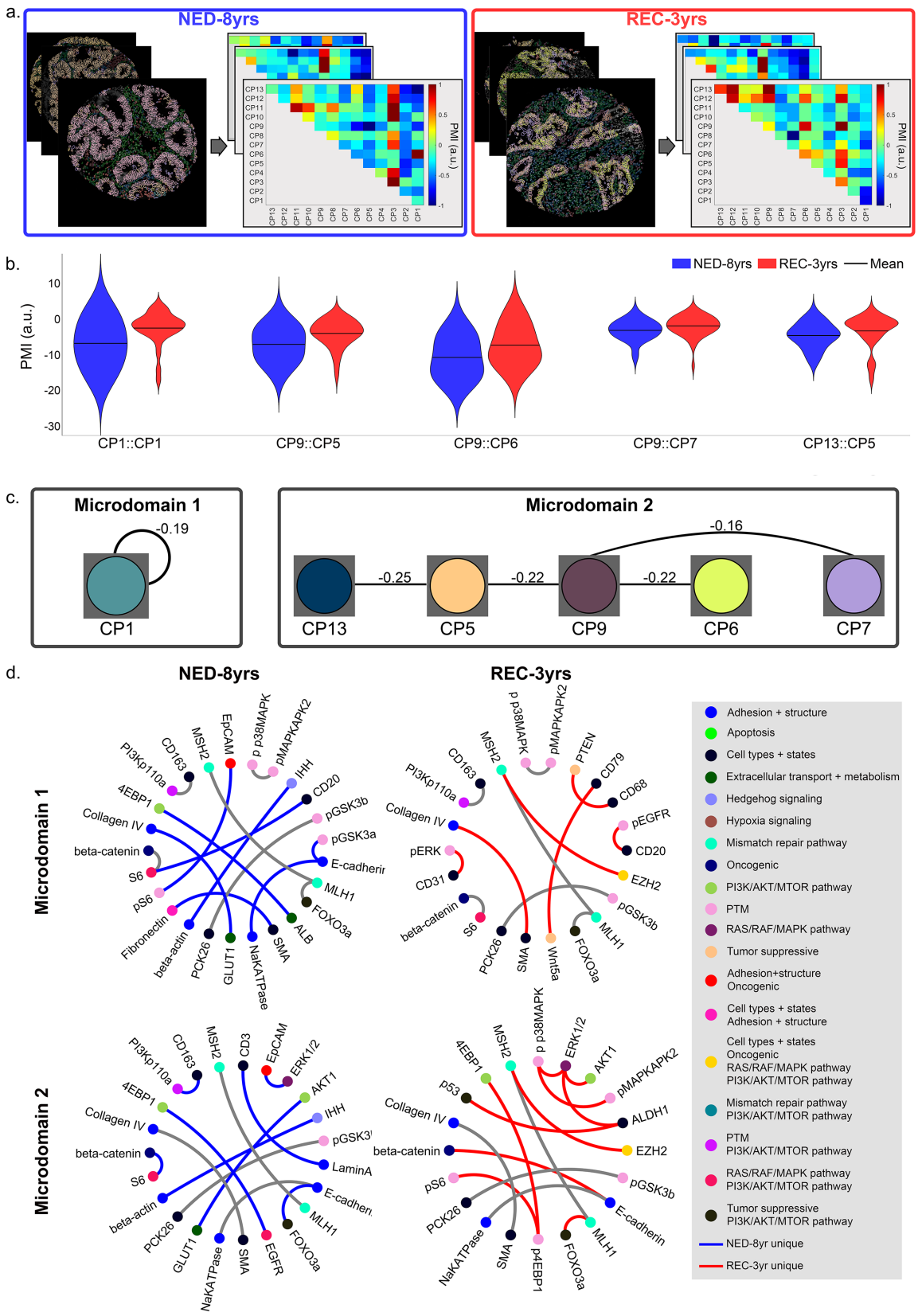
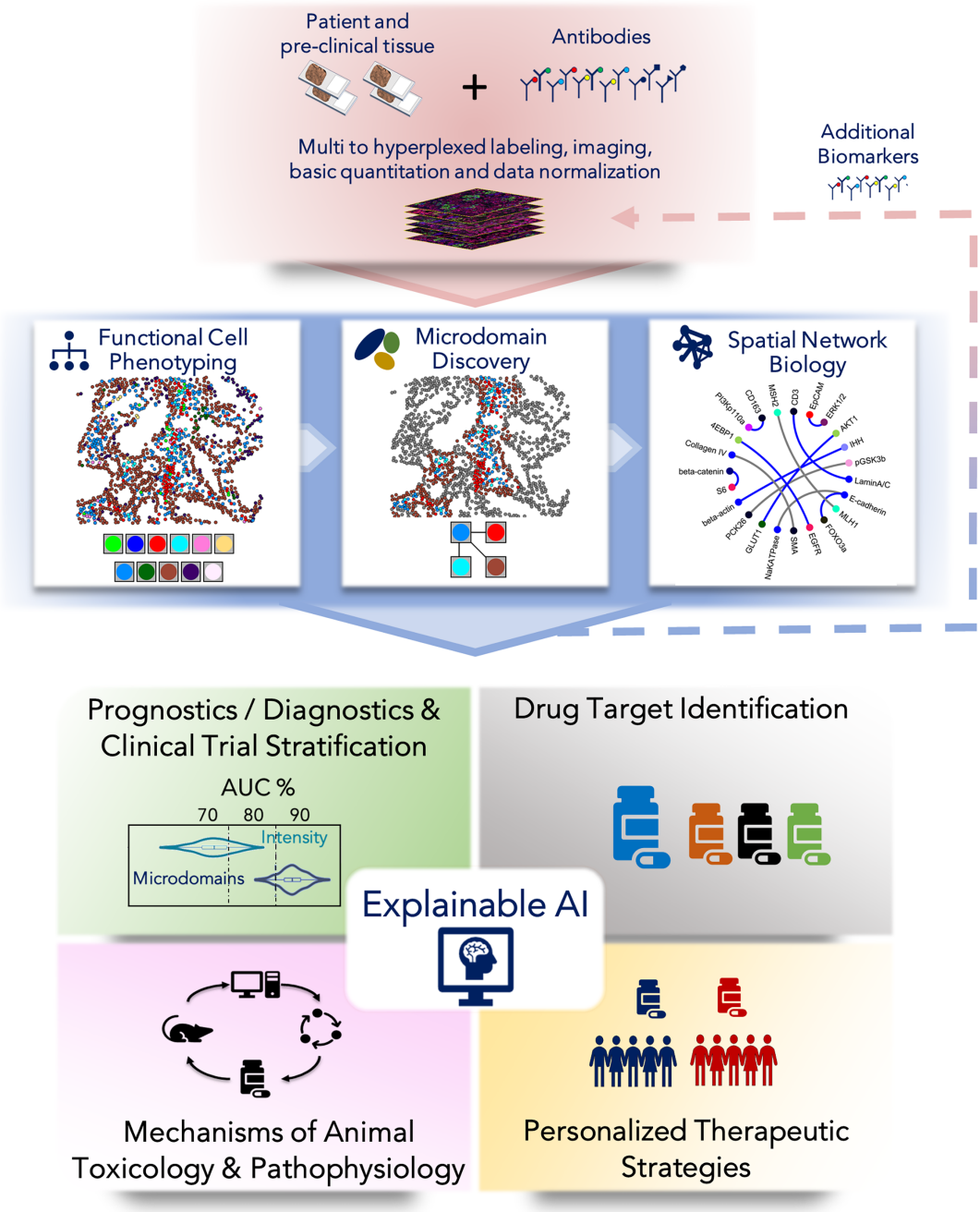

a.
NED-8yrs
REC-3yrs
PMI (a.u.)
b.
NED-8yrs
REC-3yrs
Mean
PMI (a.u.)
CP1::CP1
CP9::CP5
CP9::CP6
CP9::CP7
CP13::CP5
c.
Microdomain 1
-0.19
CP1
Microdomain 2
-0.16
-0.25
-0.22
-0.22
CP13
CP5
CP9
CP6
CP7
d.
NED-8yrs
REC-3yrs
Microdomain 1
Microdomain 2
Adhesion + structure
Apoptosis
Cell types + states
Extracellular transport + metabolism
Hedgehog signaling
Hypoxia signaling
Mismatch repair pathway
Oncogenic
PI3K/AKT/MTOR pathway
PTM
RAS/RAF/MAPK pathway
Tumor suppressive
Adhesion+structure Oncogenic
Cell types + states Adhesion + structure
Cell types + states Oncogenic RAS/RAF/MAPK pathway PI3K/AKT/MTOR pathway
Mismatch repair pathway PI3K/AKT/MTOR pathway
PTM PI3K/AKT/MTOR pathway
RAS/RAF/MAPK pathway PI3K/AKT/MTOR pathway
Tumor suppressive PI3K/AKT/MTOR pathway
NED-8yr unique
REC-3yr unique

Figure 5


Patient and
pre-clinical tissue
Antibodies
Multi to hyperplexed labeling, imaging,
basic quantitation and data normalization
Additional
Biomarkers
Functional Cell
Phenotyping
Microdomain
Discovery
Spatial Network
Biology
PI3Kp110a
CD163
MSH2
CD3
EpCAM
ERK1/2
4EBP1
AKT1
Collagen IV
IrhH
pGSK3b
beta-catenin
LaminA/C
S6
E-cadherin
beta-actin
MLH1
PCKCS
FOXO3a
GLUT1
EGFR
NaKATPase
SMA
Prognostics / Diagnostics &
Clinical Trial Stratification
AUC %
70
80
90
Intensity
Microdomains
Drug Target Identification
Explainable AI
Mechanisms of Animal
Toxicology & Pathophysiology
Personalized Therapeutic
Strategies