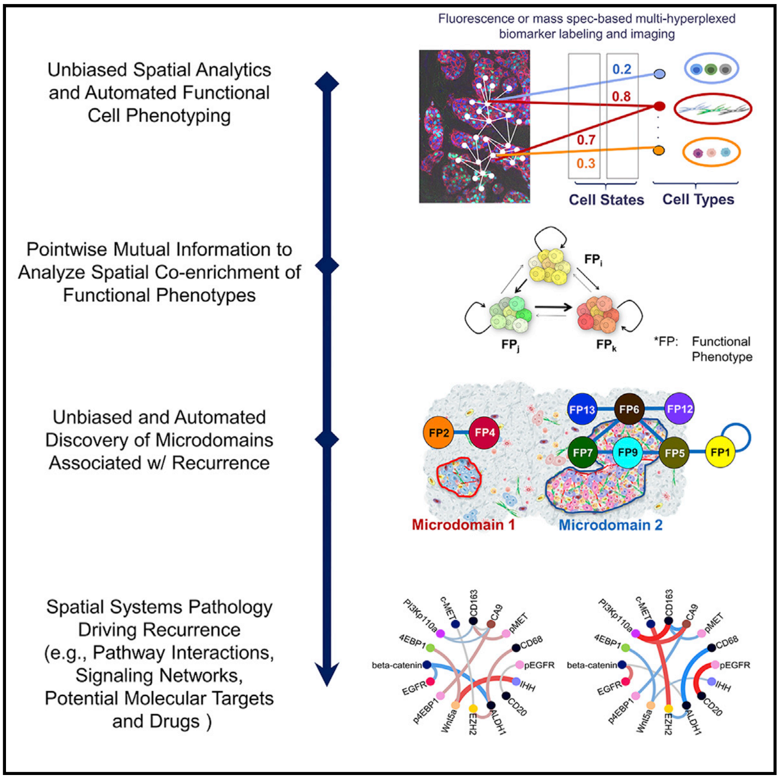
Cell Reports
Methods

Article

In situ functional cell phenotyping reveals microdomain networks in colorectal cancer recurrence

Graphical abstract


Unbiased Spatial Analytics and Automated Functional Cell Phenotyping
Fluorescence or mass spec-based multi-hyperplexed biomarker labeling and imaging
0.2
0.8
0.7
0.3
Cell States
Cell Types
Pointwise Mutual Information to Analyze Spatial Co-enrichment of Functional Phenotypes
FP_i
FP_j
FP_k
*FP: Functional Phenotype
Unbiased and Automated Discovery of Microdomains Associated w/ Recurrence
FP2
FP4
FP13
FP6
FP12
FP7
FP9
FP5
FP1
Microdomain 1
Microdomain 2
Spatial Systems Pathology Driving Recurrence (e.g., Pathway Interactions, Signaling Networks, Potential Molecular Targets and Drugs)
PI3Kp110a
c-MET
CD163
CA9
pMET
4EBP1
CD68
beta-catenin
pEGFR
EGFR
IHH
p4EBP1
CD20
Wnt5a
EZH2
ALDH1

Highlights

- LEAPH derives functional cell phenotypes on a continuum from spatial biomarker data
- Pointwise mutual information captures spatial co-enrichment of functional phenotypes
- Identifies recurrence-associated, spatial configurations of functional phenotypes
- Derives microdomain-specific signaling networks driving recurrence

Authors

Samantha A. Furman, Andrew M. Stern, Shikhar Uttam, D. Lansing Taylor, Filippo Pullara, S. Chakra Chennubhotla

Correspondence

sternam@pitt.edu (A.M.S.),
chakra@spintellx.com (S.C.C.)

In brief

Furman et al. present an unbiased spatial analytics approach, LEAPH, to derive functional cell phenotypes on a continuum from spatial proteomics data. LEAPH was applied to a hyperplexed colorectal cancer image dataset to discover microdomains and signaling networks associated with, and potentially driving, colorectal cancer recurrence.

Furman et al., 2021, Cell Reports Methods *1*, 100072
September 27, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.crmeth.2021.100072

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

In situ functional cell phenotyping reveals microdomain networks in colorectal cancer recurrence

Samantha A. Furman,[1] Andrew M. Stern,[1,2,*] Shikhar Uttam,[1] D. Lansing Taylor,[1,2,3] Filippo Pullara,[3] and S. Chakra Chennubhotla[1,2,3,4,*]

[1]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260, USA
[2]University of Pittsburgh Drug Discovery Institute, University of Pittsburgh, Pittsburgh, PA 15261, USA
[3]SpIntellx, Inc., 2425 Sidney Street, Pittsburgh, PA 15203, USA
[4]Lead contact
*Correspondence: sternam@pitt.edu (A.M.S.), chakra@spintellx.com (S.C.C.)

https://doi.org/10.1016/j.crmeth.2021.100072

MOTIVATION The functional response of a cell is determined not only by its internal state, but also by its interactions with its neighbors and the stimulus it receives from its local environment ("microdomain"). Predefined cell types based on dichotomous cell states fail to capture the highly plastic functional phenotypic continuum of cells in a complex setting such as the tumor microenvironment. To address this challenge, Furman et al. present an unbiased spatial analytics approach to characterize cell states on a continuum from hyperplexed datasets and discover microdomains and signaling networks associated and potentially driving colorectal cancer recurrence.

SUMMARY

Tumors are dynamic ecosystems comprising localized niches (microdomains), possessing distinct compositions and spatial configurations of cancer and non-cancer cell populations. Microdomain-specific network signaling coevolves with a continuum of cell states and functional plasticity associated with disease progression and therapeutic responses. We present LEAPH, an unsupervised machine learning algorithm for identifying cell phenotypes, which applies recursive steps of probabilistic clustering and spatial regularization to derive functional phenotypes (FPs) along a continuum. Combining LEAPH with pointwise mutual information and network biology analyses enables the discovery of outcome-associated microdomains visualized as distinct spatial configurations of heterogeneous FPs. Utilization of an immunofluorescence-based (51 biomarkers) image dataset of colorectal carcinoma primary tumors (n = 213) revealed microdomain-specific network dysregulation supporting cancer stem cell maintenance and immunosuppression that associated selectively with the recurrence phenotype. LEAPH enables an explainable artificial intelligence platform providing insights into pathophysiological mechanisms and novel drug targets to inform personalized therapeutic strategies.

INTRODUCTION

Tumors are dynamic and complex ecosystems. Tumor cells and their stromal 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 carcinoma (CRC) (Almendro et al., 2013; Bussard et al., 2016; Caswell and Swanton, 2017; DeGregori, 2017; Hanahan and Weinberg, 2011; Roerink et al., 2018; Shia et al., 2017; Tape, 2017; Tauriello et al., 2017; Tauriello and Batlle, 2016). 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 (Almendro et al., 2013; Balkwill et al., 2012; Caswell and Swanton, 2017; Junttila and de Sauvage, 2013).

The functional response of a cell is determined not only by its internal state, but also by its interactions with its neighbors, and the stimulus it receives from its local environment ("microdomain")

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

(Smith and Hodges, 2019; Vitale et al., 2021). Although the cell states are typically considered to be dichotomous, such as macrophages existing in known M1-like and M2-like states, an emerging view is the existence of a continuum of phenotypic states and the consequent emergence of functional plasticity in responding to perturbations (Aoki et al., 2020; Azizi et al., 2018; Lambrechts et al., 2018; Neftel et al., 2019; Vitale et al., 2021). The heterogeneity in the composition of the microdomains and resulting spatial intercellular communication patterns within each microdomain has been linked to the spatial variability in disease progression and treatment response (Azizi et al., 2018; Bartoschek et al., 2018; Binnewies et al., 2018; Maynard et al., 2020; Sautès-Fridman et al., 2019; Smith and Hodges, 2019; Suda et al., 2018; Thorsson et al., 2018; Vitale et al., 2021; Zhang et al., 2018). In addition, previous work has described the spatial landscape of individual tumors as hot, cold, or excluded on the basis of the immune-stromal microdomain patterns (Bonaventura et al., 2019). The basic numerical quantification of these categories is an oversimplification of spatial ITH and does not consider the evolutionary pressures of the microenvironment (e.g., immunosuppressive signals confining CD8+ cytotoxic T lymphocytes to the tumor periphery) (Vitale et al., 2021). In fact, the microdomain-specific signaling reciprocally supports the viewpoint of a highly plastic functional phenotypic continuum with complex intermediary cell types and cell states shaping the TME (Azizi et al., 2018; Smith and Hodges, 2019; Vitale et al., 2021). Characterizing the phenotypic continuum can identify emergent cell dependencies to improve diagnosis/prognosis and inform new therapeutic opportunities (Smith and Hodges, 2019; Vitale et al., 2021).

Single-cell approaches, which include a "large" set of biomarkers (including DNA, RNA, and proteins), are well suited to capture this emergent phenotypic continuum (Nachmanson et al., 2021; Nirmal et al., 2021). However, preserving the spatial context is equally crucial to capture the various functional states that cells might emerge through their neighborhood interactions (Vitale et al., 2021; Zanotelli et al., 2020). Spatial transcriptomics is a recent method that provides transcriptome-wide information but is currently limited to a resolution of 10 μm (Burgess, 2019; *Nature Methods*, 2021; Stickels et al., 2021). Here, we focus on recent high-dimensional single-cell imaging methods, such as multiplexed and hyperplexed immunofluorescence imaging (Gerdes et al., 2013; Goltsev et al., 2018; Lin et al, 2016, 2018), imaging mass cytometry (IMC) (Giesen et al., 2014), spatial transcriptomics, and probabilistic cell typing by *in situ* sequencing (pciSeq) (Qian et al., 2020).

The challenge we tackle in this study is how to utilize the complex spatial and high-dimensional output of these hyperplexed technologies to characterize cell types along a phenotypic continuum. We recognize that there are various cell-phenotyping methods available, each with its own advantages and disadvantages. First, many studies that have spatial data available do not utilize the spatial context for cell phenotyping. Commonly, the spatial information is applied as a post-cell typing step to describe the differences in the spatial composition of cell types across samples (Chen et al., 2020; Jackson et al., 2020; McKinley et al., 2017; Schürch et al., 2020; Menietti et al., 2020). Second, to capture the transitional cell states along a phenotypic continuum, it is advantageous to use probabilistic clustering as opposed to hard clustering, which assumes each cell belongs to only one cell type. Methods such as pciSeq (Qian et al., 2020) and Harmony (Korsunsky et al., 2019) utilize probabilistic clustering but each with a different motivation. pciSeq uses probabilistic clustering to estimate the assignments of each single-cell RNA sequencing (scRNA-seq) read to each *in situ* cell and each cell to a cell type. Harmony uses probabilistic clustering to estimate cell types across multiple datasets with dataset-specific conditions. Third, utilizing a hierarchical component is advantageous to dissect nested cell types and states (e.g., immune cells, macrophages, specific macrophage states). A frequently used tool, CellEngine (www.cellEngine.com), includes the option to dissect cell types in a hierarchical structure but the cell types defined in this tool are supervised and, therefore, user defined. In another study, a hierarchical structure is used to tease apart specifically the immune cell subclusters (Santamaria-Pang et al., 2017). In this paper, we explore a cellular phenotyping method that utilizes a spatial, probabilistic, and hierarchical component.

Specifically, we describe LEAPH, an unsupervised machine **le**arning **a**lgorithm for identifying cell **ph**enotypes. LEAPH builds a tree-structured hierarchy of cell types and cell states, which we refer to as functional phenotypes (FPs), on a continuum using recursive probabilistic clustering and spatial regularization steps. LEAPH learns the FPs with no ground truth or tagged data, hence our use of the word unsupervised. However, the number of cell clusters derived by LEAPH is tied to a small set of free parameters within the stopping criteria (e.g., using a minimum cell fraction of 1%). We applied LEAPH in combination with pointwise mutual information (PMI) to hyperplexed (51 biomarkers) immunofluorescence images of CRC primary tumors (n = 213) with corresponding clinical data. The biomarkers were selected with the interest of sampling a range of CRC and cancer biology TME properties (see STAR Methods for more details).

The data-driven and computationally unbiased approach of LEAPH captured a phenotypic continuum comprised of *specialized*, *transitional*, and *multi-transitional* cell states that are intrinsic to the architecturally complex and reciprocally coevolving TME of CRC. Each LEAPH-derived FP is further characterized by a unique biomarker-positive (+) signature for ease of interpretation. With these biomarker signatures, we conclude properties demonstrating the heterogeneity of tumor cells, cancer stem cells (CSCs), macrophages, cancer-associated fibroblasts (CAFs), immune cells, and hybrid tumor-macrophage cells. 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. This simulation demonstrates the potential of performing iterative cycles of imaging and computational analysis with an optimal biomarker set to fully exploit the capabilities of LEAPH.

Using our previous work utilizing PMI to characterize spatial ITH (Spagnolo et al., 2016), we show how LEAPH enables the discovery of *microdomains*, which are spatial configurations of FPs driving disease progression dynamics in the TME of CRC. We discovered recurrence-associated microdomains and hypothesize

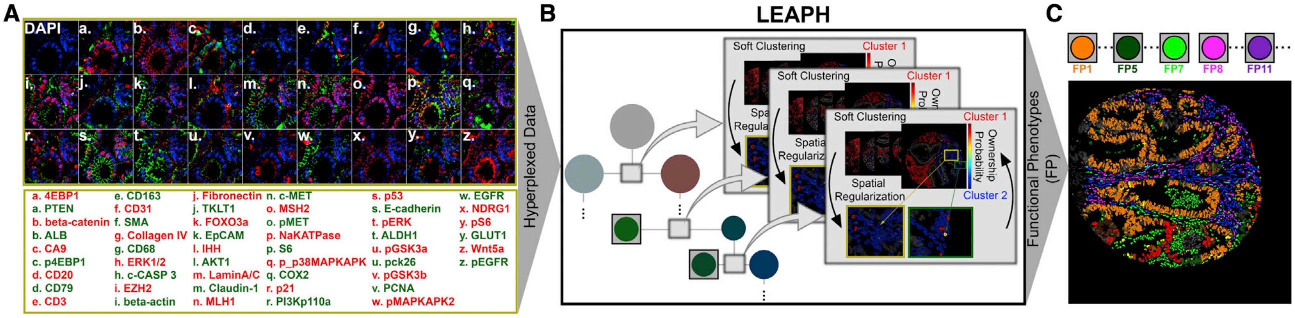


Figure 1. Unsupervised machine learning algorithm for identifying functional phenotypes (LEAPH) on hyperplexed CRC tissue microarray dataset

(A) A zoomed in region of a stage II CRC tissue sample (ca. 0.6 mm 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]). See also Figure S1.

(B) Workflow schematic of the LEAPH algorithm, which performs recursive steps of probabilistic clustering and spatial regularization.

(C) The terminal nodes of the tree (leaves) signify distinct FPs.

the microdomains to hold either tumor-promoting or -suppressing properties. Furthermore, co-evolving with the TME, microdomain-specific network dysregulation was observed supporting CSC maintenance and immunosuppression, which appeared necessary for driving the recurrence phenotype. 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.

RESULTS

Hyperplexed fluorescence CRC tissue microarray image dataset

Hyperplex imaging techniques can be non-destructive with the methodology of antibody labeling being either iterative (Gerdes et al, 2013, 2014; Graf and Zavodszky, 2017; Lin et al, 2016, 2018; McKinley et al., 2017; Nelson et al., 2013; Santamaria-Pang et al., 2017) or performed in a batch before imaging (Goltsev et al., 2018). Hyperplexing technologies unrelated to fluorescence imaging are also available, such as mass cytometry, which are both destructive and perform biomarker batch labeling (Giesen et al., 2014).

In this study, the primary source of data is a cohort of 213 CRC tissue samples (tumor microarray core size = 0.6 mm, one sample per patient) hyperplexed by using Cell DIVE (GE Life Sciences, Issaquah, WA) (Gerdes et al., 2013) (Figure 1A). 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 by using a collection of structural biomarkers: NaKATPase (cell membrane, cell border), S6 (cytoplasm), and DAPI (nucleus) (Figure S1A). Cells are filtered by using individual quality control scores generated for each cell (scores less than 0.7–0.8 will not be included indicating inaccurate registration, misalignment, or tissue loss) and on the basis of number of pixels per segmented subcellular compartment. See STAR 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 (Gerdes et al., 2013) (Table S1). A subset of the biomarkers was selected specifically on the basis of their known association with CRC. In addition to the tissue samples, the CRC patient dataset also includes clinical information for each patient regarding sex, age, chemotherapy treatment, and days until recurrence post-surgical resection (patient statistics in Table S2).

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 in total). LEAPH applies recursive steps of modeling cell types with probabilistic clustering and refining cell states with spatial regularization (Figure 1B). The *ownership probabilities* for any given cell are constructed recursively by parsing through the hierarchy (sum to 1 for each cell). To avoid overfitting, we apply stopping criterion on the basis of thresholds on the angle between the cluster subspaces and the fraction of cells with majority ownership probability in each cluster (see the STAR Methods). The terminal nodes of the tree, i.e., *leaves*, signify distinct data-driven FPs discovered in the input dataset determined by the recursive decomposition (Figure 1C).

We define cell states as being *specialized* or *non-specialized* on the basis of the FP ownership probabilities for each cell. A specialized cell state is defined by a strong propensity toward a single FP (ownership probability >0.95). For convenience, we further group the non-specialized cell state as being a *transitional* (ownership probabilities spread between two FPs) or *multi-transitional* (ownership probabilities spread across more than two FPs). We visualize the spatial distribution of the LEAPH-derived FPs by assigning each cell to the FP with the highest ownership probability (Figure 1C).

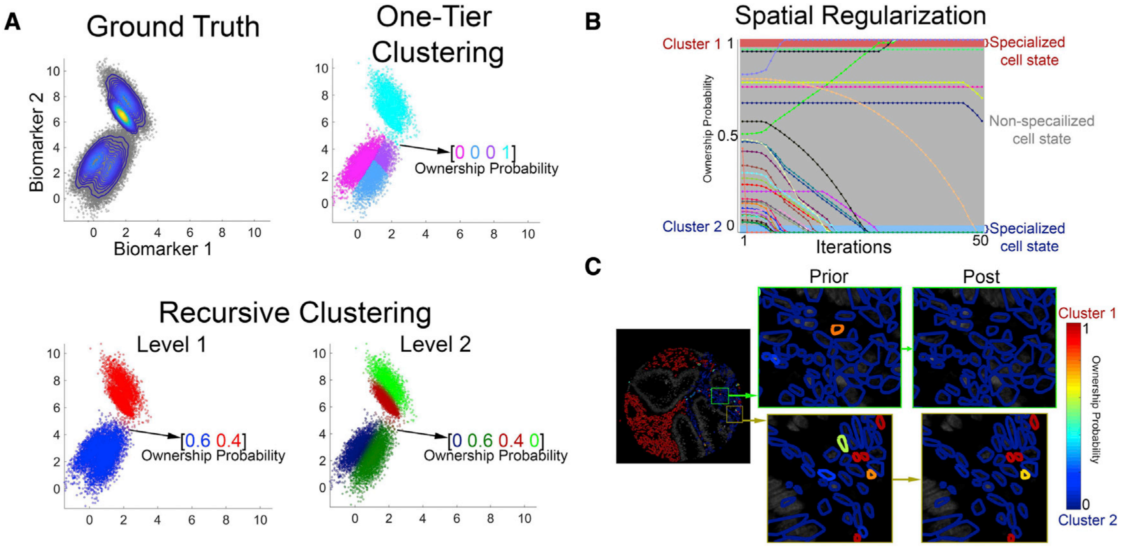


Figure 2. Illustrating the recursive steps of probabilistic clustering and spatial regularization in LEAPH

(A)**Top left**: synthetic data generated to reflect the statistics of the CRC biomarker data. **Top right**: one-tier non-recursive probabilistic model 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 on the basis of respective cluster.

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

(C) Tissue sample where each cell is outlined with the ownership probability before and after spatial regularization. We specifically point to two cells, each with an ownership probability of 0.8 before spatial regularization. After regularization, one cell converges to a specialized FP (top zoomed box) and one cell remains non-specialized (bottom zoomed box).

We generate a synthetic biomarker dataset (Figure 2A) to reflect the distribution of biomarker values in the hyperplexed CRC data. There is a varying spread of biomarker values likely due to the individual biomarker response sensitivity (Figure S1D) and the observation of dominant phenotypes (e.g., epithelial and stromal cells) with various nested cell subtypes (e.g., immune, CSCs). To account for these observations and automate the process of cellular phenotype discovery, we utilize a recursive probabilistic clustering approach where each step attempts to dissect the two most dominant clusters. A probabilistic clustering approach allows each cell to discover an identity along the phenotypic continuum by claiming “ownership” to more than one cluster. This is different from hard clustering approaches, which result in discrete cellular identities with binary ON/OFF states.

For the probabilistic clustering, we use a mixture of factor analyzers (MFA) model. In this model, 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 S1C). 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 with 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 subclusters at the second level of the hierarchy.

The MFA model parameters are learned through random initializations over 100 different runs and a consensus set of model parameters are inferred to capture the optimal subspace representation (Figure S2A). The recursive probabilistic clustering when applied alone results in a large and unrealistic population of non-specialized cells (typically 25% in a single tissue sample). The probabilistic clustering is agnostic to the spatial complexity of the TME, a key component driving ITH. On the basis of 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) (Chen et al., 2020). 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 promote ownership probability confidence and one to promote spatial coherence within a small neighborhood. We assume that 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. Within a single tissue sample, 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 2B). 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 homogeneous neighborhood and one cell remains non-specialized between the two clusters because of the heterogeneous nature of the surrounding neighborhood (Figure 2C). A more detailed description of LEAPH can be found in the STAR Methods.

LEAPH captures a diverse set of FPs comprised of specialized, transitional, and multi-transitional cell states

We applied LEAPH on the CRC hyperplexed data (~500K cells) and identified a cellular phenotypic heterogeneity tree consisting of 13 distinct FPs (leaves) (Figure 3A). A small subset of non-specialized cells (4%) is dispersed across the 13 FPs identified. We show the total number of cells with a shared ownership probability between each FP pair (Figure 3B) and observe the greatest number of non-specialized cells share ownership probabilities between FP2 and FP4. Interestingly, we find a bigger population of non-specialized cells between FP2 and FP4 in the no evidence of disease beyond 8 years (NED-8yr) patient cohort than the disease recurrence within 3 years (REC-3yr) patient cohort (Figure S2B). As a visual example, we show a small neighborhood of cells classified as specialized and non-specialized cell states (Figure 3C).

On the basis of known discriminative biomarkers (E-cadherin, NaKATPase), we identify FPs as either epithelial (FP1-7) or stromal (FP8-13). For ease of further interpretation, we identify FPs with heterogeneous properties of tumor cells (PCK26+), CSCs (EZH2+), macrophages (CD68+, CD163+), CAFs (SMA+), immune cells (CD3+, CD20+, CD31+, CD79+), and hybrid tumor-macrophage cells (Figure 3D) on the basis of FP-specific biomarker-positive (+) signatures (Figure 3E). Interestingly, on the basis of the biomarker-positive (+) signature, FP4 shows macrophage (CD163+) and tumor cell (PCK26+) properties and, therefore, we hypothesize that FP4 has identified a phenotypic cluster of protumorigenic hybrid epithelial-macrophage cells derived through cell fusion (Gast et al., 2018; Pinto et al., 2019).

To visualize the spatial complexity, we assign cells an FP identity on the basis of the highest ownership probability and randomly select patients from the extrema of the outcome-based cohorts: NED-8yrs and REC-3yrs. As expected, each tissue sample is comprised of a heterogeneous population of FPs comprised of specialized, transitional, and multi-transitional cell states (Figure 3F). In addition to applying LEAPH to the entire patient cohort, we can apply LEAPH to sub-cohorts of patients (particularly on the basis of diagnosis: stage I, II, III). We did not find a relationship between each stage-specific set of FPs and recurrence outcome. However, we did notice a difference visually in the spatial distributions and found specific properties, such as the population change of non-specialized cells between FP2 and FP4. In the next section, we further pursue the spatial relationships between the 13 LEAPH-derived FPs within each tissue sample to time to recurrence and find that the spatial distributions of FP pairs are statistically significant in relation to time-to-recurrence.

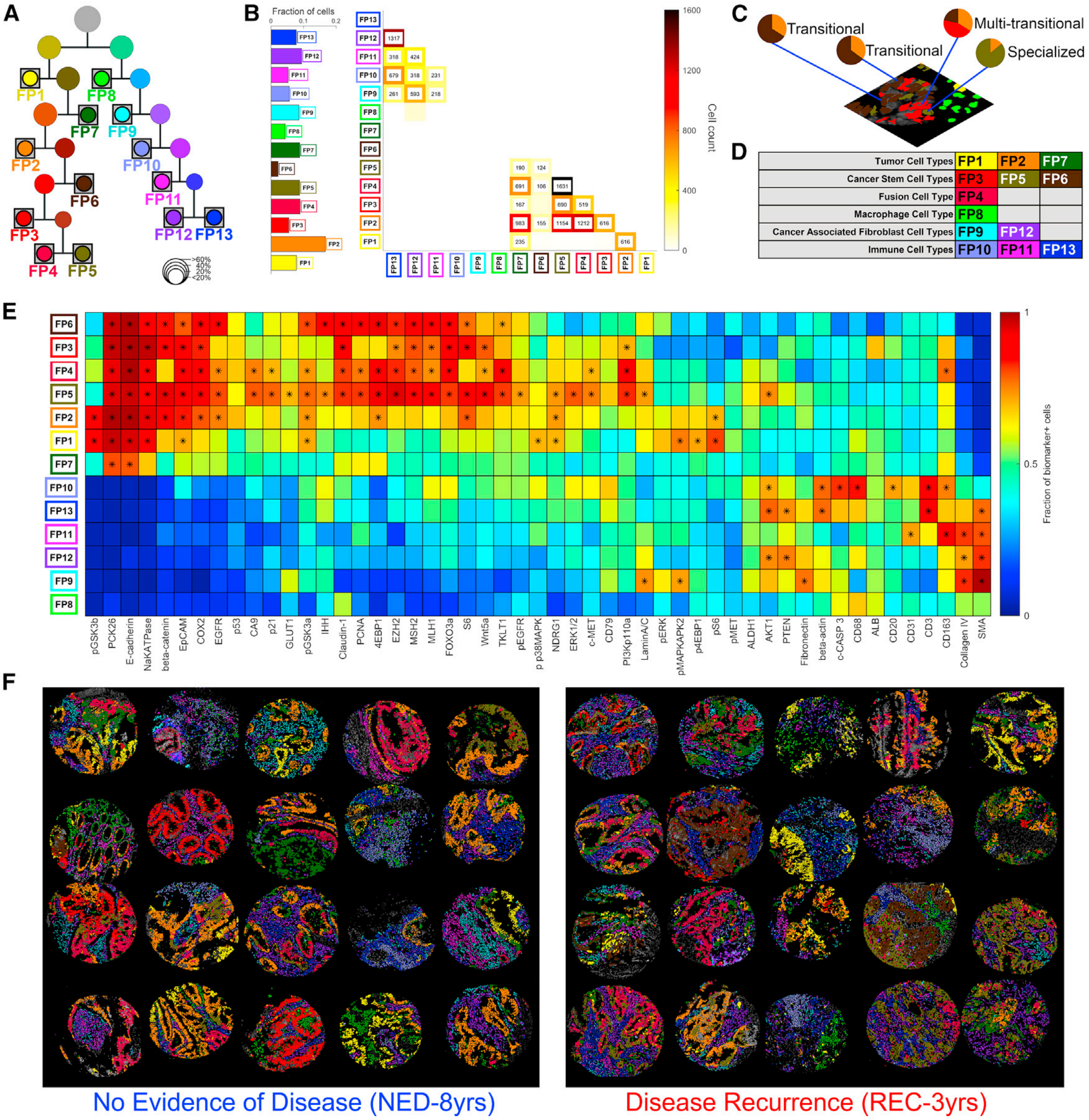
LEAPH FPs enable the discovery of microdomains driving a recurrence phenotype with PMI and spatial network biology

Visually, each tissue sample is composed of a unique pattern of LEAPH-derived FPs (Figure 3F). 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: NED-8yrs (n = 45) and REC-3yrs (n = 46) (Table S2, see the STAR Methods). To capture the spatial heterogeneity patterns, we measure the statistics of how often two FPs spatially co-occur, when compared with a background distribution. We use PMI to measure this statistic for each patient (STAR Methods) (Spagnolo et al., 2016). The resulting PMI values are either negative, zero, or positive indicating if an FP pair spatially co-occurs below, the same, or above average compared with a random background distribution (Figure 4A, see STAR Methods for details).

For each FP pair, we aggregate the PMI values to form NED-8yrs and REC-3yrs distributions (Figure S3D). We use a permutation test (Stanberry, 2013) to determine which FP pairs have a significant difference in spatial co-occurrence, compared with a background distribution, between the NED-8yrs and REC-3yrs patients (Table S3; STAR Methods). In addition, we perform the same permutation test analysis using the clinical variables (stage, grade, sex, and age) and FP fractions per tissue sample. We found that the spatial configurations determined by PMI are statistically more significant in relation to time to recurrence than the FP fractions per tissue sample and are similarly ranked to the stage and grade clinical covariates (Table S3).

We report nine FP pairs whose spatial co-occurrence compared with a background distribution is significantly different between the two patient cohorts (Figure 4B). Of these significant FP pairs, eight (of nine) FP pairs show PMI distributions skewed higher in the REC-3yrs cohort, indicating that these FP pairs spatially co-occur more significantly in this cohort as opposed to the NED-8yrs cohort (Figure 4B). One FP pair (FP2:FP4) shows the opposite, indicating that this FP pair spatially co-occurs more significantly in the NED-8yrs cohort.

The nine FP pairs form two microdomains (Figure 4C). Microdomain 1 (uD_1) is comprised of an epithelial-stromal network with interactions between two tumor FPs (FP1, FP7), two CSC FPs (FP5, FP6), two CAF FPs (FP9, FP12), and an immune cell FP (FP13) (Figures 4C and S3A–S3C). The association of stromal cells, particularly CAFs, with poor survival in CRC has been consistently demonstrated (Calon et al., 2015; Guinney et al., 2015; Isella et al., 2015). Microdomain 2 (uD_2) consists of a pairwise relationship between a tumor FP (FP2) and a hybrid tumor-macrophage FP (FP4) (Figures 4C and S3A–S3C). The discovery of uD_2 is particularly interesting because we also found the most non-specialized cells between these two FPs and a change in this population between the NED-8yr and REC-3yr cohorts (Figure S2C).



Tumor Cell Types	FP1	FP2	FP7
Cancer Stem Cell Types	FP3	FP5	FP6
Fusion Cell Type	FP4		
Macrophage Cell Type	FP8		
Cancer Associated Fibroblast Cell Types	FP9	FP12	
Immune Cell Types	FP10	FP11	FP13

Figure 3. LEAPH reveals cellular phenotypic heterogeneity and identifies specialized, transitional, and multi-transitional cell states in a CRC hyperplexed 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 FP (size key). Each leaf node represents a data-driven FP determined by the stopping criteria. The set of all leaf nodes, 13 FPs in total, form the components of the *final* mixture of factor analyzers model for the CRC patient cohort.

(B) **Left**: fraction of cells in each of the resulting data-driven FPs. **Right**: lower triangular matrix depicting the number of cells with shared ownership probabilities between each pair of FPs (see STAR Methods). We observe that the non-specialized cells are most likely in a state of transition between FP4 (hybrid tumor-macrophage cell type) and FP5 (cancer stem cell type). See also Figure S2.

(C) Example of a neighborhood of cells classified as specialized and non-specialized cell states. We point to multiple cells within the neighborhood with varying ownership probabilities. This neighborhood depicts a specialized, transitional, and multi-transitional cell states (see main text for more details).

(D) Based on biomarker-positive (+) signatures that are unique to each FP (subpanel d), we further group the 13 FPs into broad categories of tumor cells, CSCs, macrophages, CAFs, immune cells, and hybrid tumor-macrophage cells for ease of biological interpretation.

(E) Each FP has a unique biomarker-positive (+) signature which quantifies the fraction of biomarker-positive (+) cells in each FP. This signature enables us to characterize the functional properties of each FP based on known cell-type-specific markers (subpanel c).

(F) Illustrative tissue samples from NED-8yrs and REC-3yrs cohorts. Cell masks are colored to indicate the FP with has the highest ownership probability for the reference cell.

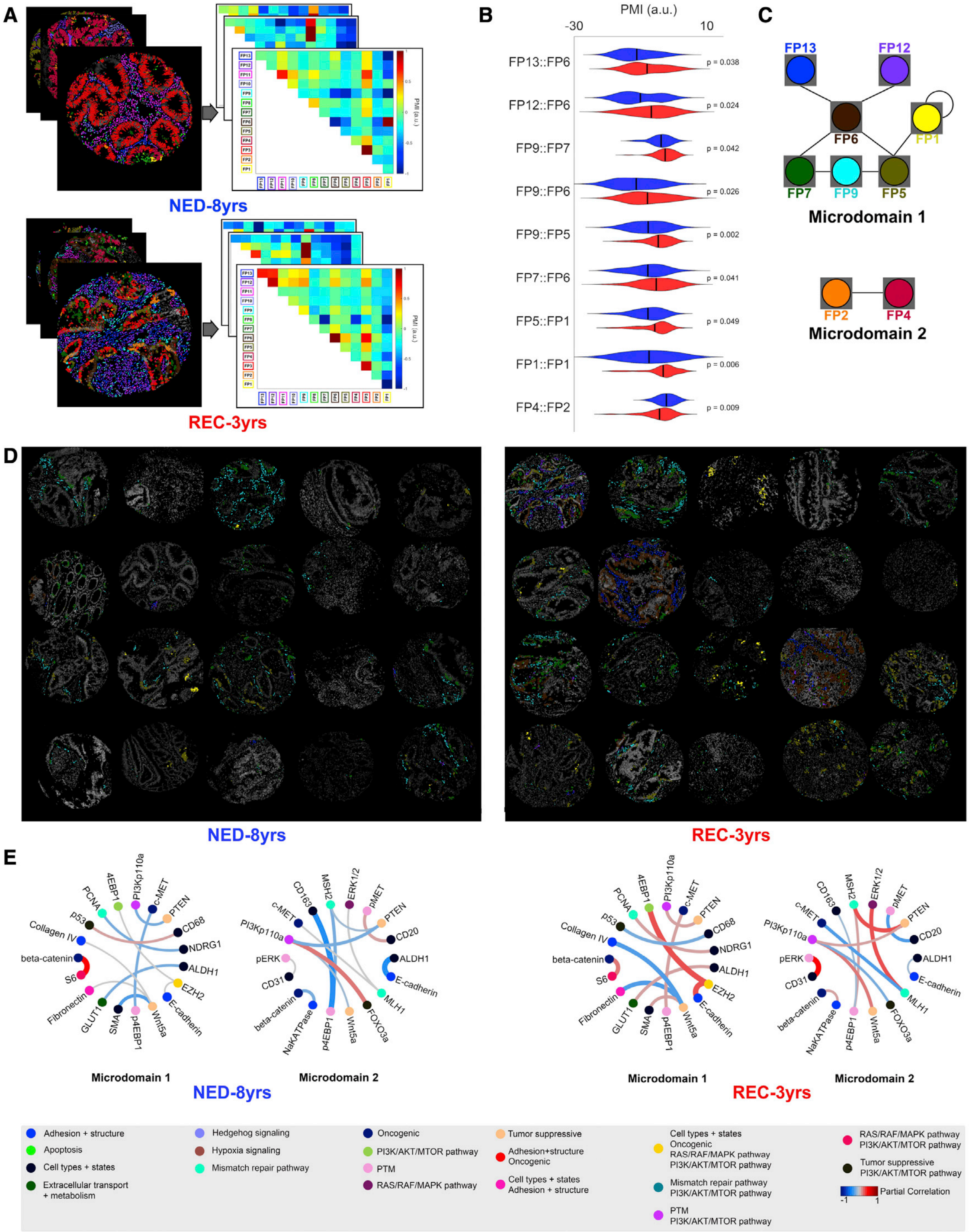


Figure 4. Discovery of microdomains driving a recurrence phenotype with PMI and spatial network biology

(A) PMI maps are computed for each tissue sample to quantify the spatial co-occurrence of each FP pair in relation to a random background distribution (see STAR Methods). The PMI values are normalized to the range −1 to 1 for visualization.

(B) The PMI values for each FP pair are grouped by the outcome data: NED-8yrs and REC-3yrs. Comparing the PMI distributions with a permutation test identified nine significant FP pairs ($p < 0.05$), including one FP pair to be highly significant ($p < 0.005$) (see STAR Methods, Table S3, and Figure S3D). Each significant FP pair (except FP4:FP2) has a distribution skewed higher for the REC-3yr group implicating a greater level of spatial co-occurrence compared with a random background distribution (vice versa for FP4:FP2).

(C) Two microdomains emerge from the FP pairwise significance analysis.

(legend continued on next page)

Unlike uD_1, uD_2 consists of an FP pair which shows a PMI distribution skewed higher in the NED-8yrs cohort (as opposed to the REC-3yr cohort), indicating that this FP pair has a stronger propensity to spatially co-occur in the NED-8yr cohort. Based on this observation, we hypothesize that uD_1 has tumor-promoting properties and uD_2 has tumor-suppressing properties.

We further investigated spatial biomarker network biology associated with recurrence. For each microdomain and patient cohort, we computed the partial correlation between each biomarker pair, controlling for the remaining confounding biomarkers (Uttam et al., 2020). We performed a differential connectivity analysis with a permutation test to determine the biomarker-pairs with a significant change in the partial correlation values between the NED-8yrs and REC-3yrs cohorts (see STAR Methods for details). This analysis results in a microdomain-specific partial correlation networks for each patient cohort (Figures 4E; Table S4). Importantly, most of these comparisons show a *change in sign* in the partial correlation values, suggesting that a distinct difference in network dysregulation in addition to co-occurrence of cell type per se is necessary for driving the recurrence phenotype. This result supports our working hypothesis that, within the evolving tumor microenvironment, the molecular signaling networks within each microdomain undergo a regulatory switch to confer a recurrence phenotype supported by cancer stem cell maintenance and immunosuppression (Augustin et al., 2016; Bienz and Clevers, 2000; de Jaeghere et al., 2019; Dienstmann et al., 2017; Galluzzi et al., 2019; Grasso et al., 2018; Kraman et al., 2010; Naito et al., 1998; Pai et al., 2017; Spranger and Gajewski, 2015, 2018; Turley et al., 2015; Wong et al., 2019; Yaeger et al., 2018). This is further supported by the reduction in number of transition cells across FP2 and FP4 (uD_2) between the NED-8yrs and REC-3yrs cohorts (Figure S2C). Thus, uD_1, and uD_2 microdomains represent the spatial manifestation of emergent recurrence-specific networks.

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 STAR Methods). For comparison, we use the maximum ownership probability from the results derived from the entire dataset (Figure 3) as the ground truth FP identity. We measure the accuracy as the percentage of cells with a matching FP identity to the ground truth.

At the first level, only two biomarkers (PCK26, E-cadherin) are needed to reproduce the cellular phenotypic assignments with an accuracy above 97% (Figure S4A). Increasing the number of biomarkers further increases the accuracy to almost 99% (Figure S4A). 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 1 (Figure S4A). The epithelial subtyping dissection reaches 89% accuracy with four biomarkers (4EBP1, pS6, pMAPKAPK2, FPNA) (Figure S4A). An increase in biomarkers does not make an overall substantial difference in the reproduction accuracy. The stromal subtyping dissection reaches above 93% accuracy with four 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 eight or more biomarkers (Figure S4A).

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 FPs (Figure 5).

DISCUSSION

We describe LEAPH, an unsupervised, spatially informed, recursive probabilistic clustering method, to describe FPs along a continuum of cell types and cell states. The traditional approaches to cell phenotyping focus on each cell type alone. In comparison, the functional cell phenotyping approach presented here focuses on discovering data-driven cell phenotypes and associating the resulting phenotypes with clinical outcomes. Inferring functional cell properties associated with outcomes enables the generation of testable hypotheses and thus moves well beyond typical cell phenotyping.

The computational time for applying LEAPH is correlated to the number of levels the hierarchy derives (Figure S4B). LEAPH is built as a generalizable model amenable to other datum sources, probabilistic clustering algorithms, and spatial regularization objective functions/optimizations. The probabilistic clustering step of LEAPH is amenable to other probabilistic mixture models such as but not limited to Gaussian mixture models (Reynolds et al., 2000) and mixtures of probabilistic principal-component analysis (Tipping and Bishop, 1999). 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. Currently, most of the computational effort is absorbed in

(D) We visualize microdomain 1 in the NED-8yrs and REC-3yrs cohort with the same set of tissue samples as in Figure 3F, but only coloring the cells that contribute to microdomain 1 (see STAR Methods).

(E) Within each microdomain, we identify a recurrence-associated biomarker network (see STAR Methods; Table S4). The biomarkers are grouped on the basis of their presumed cellular functions/processes and the color and thickness of each edge is coordinated to the partial correlation value between the biomarker pair (see legend).

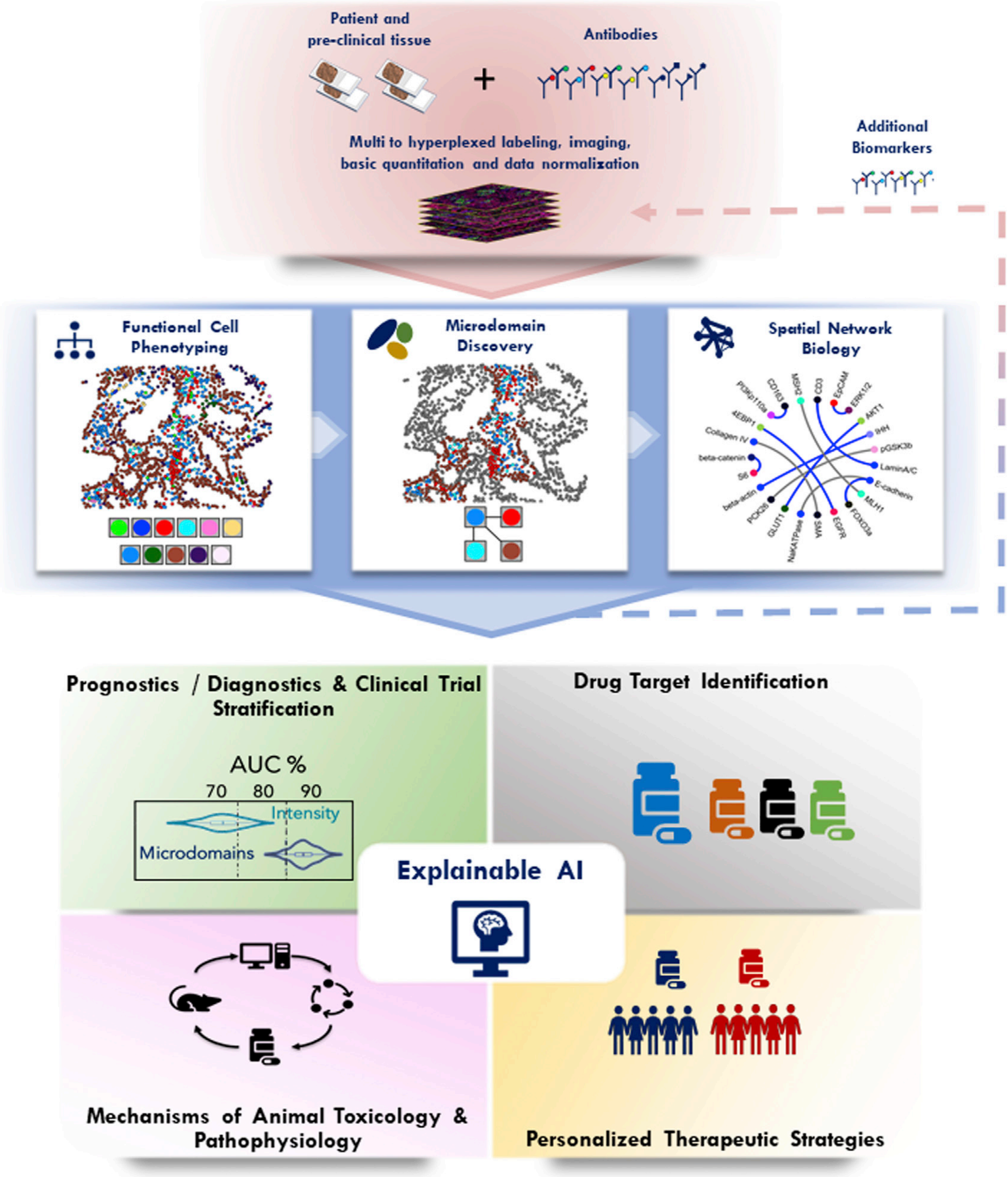


Figure 5. Analytical platform that can be applied to any multi- to hyperplexed image datasets

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 (Giesen et al., 2014; Goltsev et al., 2018; Lin et al, 2016, 2018; Vickovic et al., 2019). 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 spectrometry [Giesen et al., 2014]). Basic image processing and basic image quantitation is then followed by data normalization in preparation for applying the analytical platform. The hyperplexed imaging process can be initiated with a limited set of biomarkers from which LEAPH builds a data-driven, computationally unbiased phenotypic hierarchy of cell types, cell states capturing a phenotypic continuum, and their spatial configurations. LEAPH in combination with PMI discovers outcome-specific microdomains composed of spatially configured functional 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 (Giesen et al., 2014; Goltsev et al., 2018; Lin et al, 2016, 2018; Vickovic et al., 2019).

spatial regularization (Figure S4B) and therefore further optimization of this step will decrease the total time of LEAPH.

The input criteria for LEAPH are single-cell biomarker data vectors with corresponding spatial information per sample. Other sources for spatial multiparameter cellular and subcellular imaging data include transmitted light (H&E and IHC), fluorescence, immunofluorescence, live cell biomarkers, mass spectrometry, electron microscopy, etc. Also, we have demonstrated a non-spatial version of LEAPH in Figure S4C that can be applied to non-spatial data-generation methods such as scRNA-seq. As future work, we will extend LEAPH to analyze spatial transcriptomics data, as this technology is rapidly evolving with the goal of capturing single-cell spatial transcriptomic-wide information (Burgess, 2019; *Nature Methods*, 2021).

To suggest its potential for broad utility, we also applied LEAPH to a breast cancer IMC dataset (Jackson et al., 2020). LEAPH derives 35 FPs with functional properties of tumor cells, CSCs, epithelial mixed cells, tumor associated stromal cells, T cells, CAFs, endothelial cells, immune mixed cells, macrophages, and stromal mixed cells (Figure S5). Interestingly, there is a large population of non-specialized cells with shared ownership probability between FP31 and FP34, a tumor cell and CSC population that would be worth investigating in the future. We found a strong correlation between the LEAPH FPs, and the 71 cell clusters previously published and derived from PhenoGraph (Figure S5E) (Jackson et al., 2020; Levine et al., 2015). Although 71 cell clusters were dissected with PhenoGraph, the previous study chose to further group these clusters to form 27 metaclusters for downstream analysis and biological interpretation (Jackson et al., 2020). In Figures S5B and S5C, we present the interpreted biological functionalities of the phenotypes that LEAPH generates.

Beyond LEAPH, we have also laid out a framework utilizing the LEAPH-derived FPs 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 FPs and subsequently from microdomains based on the recurrence-specific FP pairs. We hypothesize that capturing the higher-order spatial relationships between groups of phenotypes might provide additional information on the spatial configurations of the microdomains. Furthermore, there are many graphical models (Barber and Drton, 2015; Foygel and Drton, 2010; Meinshausen and Bühlmann, 2006) that could also be used in the microdomain-specific network biology analyses.

To further investigate our findings, we can exploit the non-destructive property of cyclical imaging platforms (Giesen et al., 2014; Goltsev et al., 2018; Lin et al, 2016, 2018; Vickovic et al., 2019), which allow iterative analysis of additional biomarkers on the same tissue sample to test mechanistic hypotheses, and identify novel biomarkers and optimal therapeutic strategies (see below and Figure 5). Given the impact our computational framework can have, in the future, it will be imperative to apply our methods on a larger cohort of CRC data to test the universality of our discovered recurrence-associated microdomains and microdomain-specific biomarker networks.

Here, we denoted the leaves of the tree generated by LEAPH as our FPs. 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 (Uttam et al., 2020). Indeed, the same analysis can be repeated by using the entire library of FPs derived from LEAPH. We expect the predictive analysis will be further enhanced with the additional phenotypes and the derived microdomains. In addition, given the impact our computational framework can have, in the future, it will be imperative to apply our methods on a larger cohort of CRC data to test the universality of our discovered recurrence-associated microdomains and microdomain-specific biomarker networks. On the basis of the statistical robustness of our approach and external experimental validation of our results found in the literature, we predict that we will find similar microdomains presented here in larger CRC datasets.

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 FPs 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 (Spagnolo et al., 2016; Uttam et al., 2020). LEAPH builds a statistical framework for the application of PMI (Spagnolo et al., 2016) 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 (Uttam et al., 2020) 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 a hypothesis for future experimental and computational studies to further investigate the signaling and crosstalk 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 (Tosun et al., 2020) with applications probing and modulating the 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).

Limitations of study

The 13 FPs reported here might not be comprehensive because of the choice made in selecting the panel of biomarkers to label and image. The number of FPs is also tied to the parameter choice (e.g., using a minimum cell fraction of 1%). Lowering the stopping criterion parameters for the recursive procedure would result in a more fine-grained decomposition of the cells.

This could show subtle differences in functionality and this option should be explored in future work. Our goal in this study is to characterize outcome-associated FPs and microdomains. To that end, the downstream analysis of relating the FPs to the outcome data can be done at any level of the recursive hierarchy, including if the hierarchy were to be expanded beyond the current 13 phenotypes.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Hyperplexed CRC data
  - Data pre-processing
  - LEAPH construction
  - Discovery of recurrence-associated microdomains
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100072.

ACKNOWLEDGMENTS

This work was supported in part by the grants NIH/NCI U01CA204836 (to S.C.C. and D.L.T.); NIH/NIBIB 5T32EB009403 and 1NIH/NCI 1F31CA254332 (to S.A.F.); NIH P30CA047904, and PA DHS 4100054875 (to D.L.T.); NIH/NHGRI U54HG008540, and UPMC Center for Commercial Applications of Healthcare Data 711077 (to S.C.C.). 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 CRC dataset generated on Cell DIVE from their previous study. We greatly appreciate the insightful comments and feedback from the reviewers in improving our manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, S.A.F., A.M.S., D.L.T., F.P., and S.C.C.; formal analysis, S.A.F., F.P., and S.C.C.; funding acquisition, S.A.F., D.L.T., and S.C.C.; investigation, S.A.F., A.M.S., F.P., and S.C.C.; methodology, S.A.F., S.U., F.P., and S.C.C.; project administration and resources, S.C.C.; software, S.A.F., F.P., and S.C.C.; supervision, D.L.T. and S.C.C.; validation, S.A.F., A.M.S., F.P., and S.C.C.; visualization, S.A.F., F.P., and S.C.C.; writing – original draft, S.A.F., A.M.S., D.L.T., F.P., and S.C.C.; writing – review & editing, S.A.F., A.M.S., S.U., D.L.T., F.P., and S.C.C.

DECLARATION OF 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.

Received: April 22, 2021
Revised: June 14, 2021
Accepted: August 9, 2021
Published: September 15, 2021

REFERENCES

Almendro, V., Marusyk, A., and Polyak, K. (2013). Cellular heterogeneity and molecular evolution in cancer. Annu. Rev. Pathol. Mech. Dis. https://doi.org/10.1146/annurev-pathol-020712-163923.

Aoki, T., Chong, L.C., Takata, K., Milne, K., Hav, M., Colombo, A., Chavez, E.A., Nissen, M., Wang, X., Miyata-Takata, T., et al. (2020). Single-cell transcriptome analysis reveals disease-defining T-cell subsets in the tumor microenvironment of classic Hodgkin lymphoma. Cancer Discov. *10*. https://doi.org/10.1158/2159-8290.CD-19-0680.

Augustin, I., Dewi, D.L., Hundshammer, J., Rempel, E., Brunk, F., and Boutros, M. (2016). Immune cell recruitment in teratomas is impaired by increased Wnt secretion. Stem Cell Res. *17*. https://doi.org/10.1016/j.scr.2016.10.010.

Method of the year 2020: spatially resolved transcriptomics. Nat. Methods. https://doi.org/10.1038/s41592-020-01042-x.

Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., et al. (2018). Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell *174*. https://doi.org/10.1016/j.cell.2018.05.060.

Balkwill, F.R., Capasso, M., and Hagemann, T. (2012). The tumor microenvironment at a glance. J. Cell Sci. *125*. https://doi.org/10.1242/jcs.116392.

Barber, R.F., and Drton, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. Electron. J. Stat. *9*. https://doi.org/10.1214/15-EJS1012.

Bartoschek, M., Oskolkov, N., Bocci, M., Lövrot, J., Larsson, C., Sommarin, M., Madsen, C.D., Lindgren, D., Pekar, G., Karlsson, G., et al. (2018). Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing. Nat. Commun. *9*. https://doi.org/10.1038/s41467-018-07582-3.

Bienz, M., and Clevers, H. (2000). Linking colorectal cancer to Wnt signaling. Cell. https://doi.org/10.1016/S0092-8674(00)00122-7.

Binnewies, M., Roberts, E.W., Kersten, K., Chan, V., Fearon, D.F., Merad, M., Coussens, L.M., Gabrilovich, D.I., Ostrand-Rosenberg, S., Hedrick, C.C., et al. (2018). Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat. Med. *24*. https://doi.org/10.1038/s41591-018-0014-x.

Bonaventura, P., Shekarian, T., Alcazer, V., Valladeau-Guilemond, J., Valsesia-Wittmann, S., Amigorena, S., Caux, C., and Depil, S. (2019). Cold tumors: a therapeutic challenge for immunotherapy. Front. Immunol. https://doi.org/10.3389/fimmu.2019.00168.

Burgess, D.J. (2019). Spatial transcriptomics coming of age. Nat. Rev. Genet. https://doi.org/10.1038/s41576-019-0129-z.

Bussard, K.M., Mutkus, L., Stumpf, K., Gomez-Manzano, C., and Marini, F.C. (2016). Tumor-associated stromal cells as key contributors to the tumor microenvironment. Breast Cancer Res. https://doi.org/10.1186/s13058-016-0740-2.

Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., Sevillano, M., Palomo-Ponce, S., Tauriello, D.V.F., Byrom, D., et al. (2015). Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat. Genet. *47*. https://doi.org/10.1038/ng.3225.

Caswell, D.R., and Swanton, C. (2017). The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. BMC Med. https://doi.org/10.1186/s12916-017-0900-y.

Chen, Z., Soifer, I., Hilton, H., Keren, L., and Jojic, V. (2020). Modeling multiplexed images with spatial-LDA reveals novel tissue microenvironments. J. Comput. Biol. *27*. https://doi.org/10.1089/cmb.2019.0340.

DeGregori, J. (2017). Connecting cancer to its causes requires incorporation of effects on tissue microenvironments (Cancer Research). https://doi.org/10.1158/0008-5472.CAN-17-1207.

Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., and Tabernero, J. (2017). Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat. Rev. Cancer. https://doi.org/10.1038/nrc.2016.126.

Epskamp, S., and Fried, E.I. (2018). A tutorial on regularized partial correlation networks. Psychol. Methods *23*. https://doi.org/10.1037/met0000167.

Foygel, R., and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010.

Francis, K., and Palsson, B.O. (1997). Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. Proc. Natl. Acad. Sci. U S A *94*. https://doi.org/10.1073/pnas.94.23.12258.

Galluzzi, L., Spranger, S., Fuchs, E., and López-Soto, A. (2019). WNT signaling in cancer immunosurveillance. Trends Cell Biol. https://doi.org/10.1016/j.tcb.2018.08.005.

Gast, C.E., Silk, A.D., Zarour, L., Riegler, L., Burkhart, J.G., Gustafson, K.T., Parappilly, M.S., Roh-Johnson, M., Goodman, J.R., Olson, B., et al. (2018). Cell fusion potentiates tumor heterogeneity and reveals circulating hybrid cells that correlate with stage and survival. Sci. Adv. *4*. https://doi.org/10.1126/sciadv.aat7828.

Gerdes, M.J., Sevinsky, C.J., Sood, A., Adak, S., Bello, M.O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R.J., et al. (2013). Highly multiplexed single-cell analysis of formalinfixed, paraffin-embedded cancer tissue. Proc. Natl. Acad. Sci. U S A *110*. https://doi.org/10.1073/pnas.1300136110.

Gerdes, M.J., Sood, A., Sevinsky, C., Pris, A.D., Zavodszky, M.I., and Ginty, F. (2014). Emerging understanding of multiscale tumor heterogeneity. Front. Oncol. https://doi.org/10.3389/fonc.2014.00366.

Ghahramani, Z., and Hinton, G.E. (1997). The EM algorithm for mixtures of factor analyzers. Compute.

Giesen, C., Wang, H.A.O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P.J., Grolimund, D., Buhmann, J.M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat. Methods *11*. https://doi.org/10.1038/nmeth.2869.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell *174*. https://doi.org/10.1016/j.cell.2018.07.010.

Graf, J.F., and Zavodszky, M.I. (2017). Characterizing the heterogeneity of tumor tissues from spatially resolved molecular measures. PLoS One *12*. https://doi.org/10.1371/journal.pone.0188878.

Grasso, C.S., Giannakis, M., Wells, D.K., Hamada, T., Mu, X.J., Quist, M., Nowak, J.A., Nishihara, R., Qian, Z.R., Inamura, K., et al. (2018). Genetic mechanisms of immune evasion in colorectal cancer. Cancer Discov. *8*. https://doi.org/10.1158/2159-8290.CD-17-1327.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*. https://doi.org/10.1038/nm.3967.

Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: clustering and optimization in large graphs. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2783258.2783313.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell. https://doi.org/10.1016/j.cell.2011.02.013.

Isella, C., Terrasi, A., Bellomo, S.E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., et al. (2015). Stromal contribution to the colorectal cancer transcriptome. Nat. Genet. *47*. https://doi.org/10.1038/ng.3224.

Jackson, H.W., Fischer, J.R., Zanotelli, V.R.T., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., and Bodenmiller, B. (2020). The single-cell pathology landscape of breast cancer. Nature *578*. https://doi.org/10.1038/s41586-019-1876-x.

de Jaeghere, E.A., Denys, H.G., and de Wever, O. (2019). Fibroblasts fuel immune escape in the tumor microenvironment. Trends Cancer. https://doi.org/10.1016/j.trecan.2019.09.009.

Junttila, M.R., and de Sauvage, F.J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. Nature. https://doi.org/10.1038/nature12626.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.ru, and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

Kraman, M., Bambrough, P.J., Arnold, J.N., Roberts, E.W., Magiera, L., Jones, J.O., Gopinathan, A., Tuveson, D.A., and Fearon, D.T. (2010). Suppression of antitumor immunity by stromal cells expressing fibroblast activation protein-α. Science *330*. https://doi.org/10.1126/science.1195300.

Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwé, H., Pircher, A., van den Eynde, K., et al. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med. *24*. https://doi.org/10.1038/s41591-018-0096-5.

Leinster, T., and Cobbold, C.A. (2012). Measuring diversity: the importance of species similarity. Ecology *93*. https://doi.org/10.1890/10-2402.1.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.A.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell *162*. https://doi.org/10.1016/j.cell.2015.05.047.

Lin, J.R., Fallahi-Sichani, M., Chen, J.Y., and Sorger, P.K. (2016). Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. Curr. Protoc. Chem. Biol. *8*. https://doi.org/10.1002/cpch.14.

Lin, J.R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P.M., Santagata, S., and Sorger, P.K. (2018). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. eLife *7*. https://doi.org/10.7554/eLife.31657.

Maynard, A., McCoach, C.E., Rotow, J.K., Harris, L., Haderk, F., Kerr, D.L., Yu, E.A., Schenk, E.L., Tan, W., Zee, A., et al. (2020). Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. Cell *182*. https://doi.org/10.1016/j.cell.2020.07.017.

McKinley, E.T., Sui, Y., Al-Kofahi, Y., Millis, B.A., Tyska, M.J., Roland, J.T., Santamaria-Pang, A., Ohland, C.L., Jobin, C., Franklin, J.L., et al. (2017). Optimized multiplex immunofluorescence single-cell analysis reveals tuft cell heterogeneity. JCI insight *2*. https://doi.org/10.1172/jci.insight.93487.

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. Ann. Stat. *34*. https://doi.org/10.1214/009053606000000281.

Menietti, E., Stoltzfus, C., Olin, B., Speziale, D., Kunz, L., Poeshinger, T., and Perro, M. (2020). Understanding the complexity of the tumor microenvironment by 3-dimensional multiplexed imaging and artificial intelligence. https://doi.org/10.1158/1538-7445.AM2020-370.

Nachmanson, D., Officer, A., Mori, H., Gordon, J., Evans, M.F., Steward, J., Yao, H., Hasteh, F., Stein, G.S., Jepsen, K., et al. (2021). The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ. https://doi.org/10.1101/2021.05.11.443641.

Naito, Y., Saito, K., Shiiba, K., Ohuchi, A., Saigenji, K., Nagura, H., and Ohtani, H. (1998). CD8+ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. Cancer Res. *58*, 3491–3494.

Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell *178*. https://doi.org/10.1016/j.cell.2019.06.024.

Nelson, D.A., Manhardt, C., Kamath, V., Sui, Y., Santamaria-Pang, A., Can, A., Bello, M., Corwin, A., Dinn, S.R., Lazare, M., et al. (2013). Quantitative single cell analysis of cell population dynamics during submandibular salivary gland development and differentiation. Biol. Open *2*. https://doi.org/10.1242/bio.20134309.

Nirmal, A.J., Maliga, Z., Vallius, T., Quattrochi, B., Chen, A.A., Jacobson, C.A., Pelletier, R.J., Yapp, C., Arias-Camison, R., Chen, Y.-A., et al. (2021). The spatial landscape of progression and immunoediting in primary melanoma at single cell resolution. https://doi.org/10.1101/2021.05.23.445310.

Pai, S.G., Carneiro, B.A., Mota, J.M., Costa, R., Leite, C.A., Barroso-Sousa, R., Kaplan, J.B., Chae, Y.K., and Giles, F.J. (2017). Wnt/beta-catenin pathway: modulating anticancer immune response. J. Hematol. Oncol. https://doi.org/10.1186/s13045-017-0471-6.

Pinto, M.L., Rios, E., Durães, C., Ribeiro, R., Machado, J.C., Mantovani, A., Barbosa, M.A., Carneiro, F., and Oliveira, M.J. (2019). The two faces of tumor-associated macrophages and their clinical significance in colorectal cancer. Front. Immunol. *10*. https://doi.org/10.3389/fimmu.2019.01875.

Pozzi, F., di Matteo, T., and Aste, T. (2012). Exponential smoothing weighted correlations. Eur. Phys. J. B *85*. https://doi.org/10.1140/epjb/e2012-20697-x.

Qian, X., Harris, K.D., Hauling, T., Nicoloutsopoulos, D., Muñoz-Manchado, A.B., Skene, N., Hjerling-Leffler, J., and Nilsson, M. (2020). Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nat. Methods *17*. https://doi.org/10.1038/s41592-019-0631-4.

Reynolds, D.A., Quatieri, T.F., and Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Process. *10*. https://doi.org/10.1006/dspr.1999.0361.

Roerink, S.F., Sasaki, N., Lee-Six, H., Young, M.D., Alexandrov, L.B., Behjati, S., Mitchell, T.J., Grossmann, S., Lightfoot, H., Egan, D.A., et al. (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. Nature *556*. https://doi.org/10.1038/s41586-018-0024-3.

Santamaria-Pang, A., Padmanabhan, R.K., Sood, A., Gerdes, M.J., Sevinsky, C., Li, Q., Laplante, N., and Ginty, F. (2017). Robust single cell quantification of immune cell subtypes in histological samples. In 2017 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2017. https://doi.org/10.1109/BHI.2017.7897220.

Sautès-Fridman, C., Petitprez, F., Calderaro, J., and Fridman, W.H. (2019). Tertiary lymphoid structures in the era of cancer immunotherapy. Nat. Rev. Cancer. https://doi.org/10.1038/s41568-019-0144-6.

Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D.R., et al. (2020). Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive. Front. Cell *182*. https://doi.org/10.1016/j.cell.2020.07.005.

Shia, J., Schultz, N., Kuk, D., Vakiani, E., Middha, S., Segal, N.H., Hechtman, J.F., Berger, M.F., Stadler, Z.K., Weiser, M.R., et al. (2017). Morphological characterization of colorectal cancers in The Cancer Genome Atlas reveals distinct morphology-molecular associations: clinical and biological implications. Mod. Pathol. *30*. https://doi.org/10.1038/modpathol.2016.198.

Smith, E.A., and Hodges, H.C. (2019). The spatial and genomic hierarchy of tumor ecosystems revealed by single-cell technologies. Trends Cancer. https://doi.org/10.1016/j.trecan.2019.05.009.

Spagnolo, D.M., Gyanchandani, R., Al-Kofahi, Y., Stern, A.M., Lezon, T.R., Gough, A., Meyer, D.E., Ginty, F., Sarachan, B., Fine, J., et al. (2016). Pointwise mutual information quantifies intratumor heterogeneity in tissue sections labeled with multiple fluorescent biomarkers. J. Pathol. Inform. *7*. https://doi.org/10.4103/2153-3539.194839.

Spranger, S., and Gajewski, T.F. (2015). A new paradigm for tumor immune escape: β-catenin-driven immune exclusion. J. ImmunoTherapy Cancer. https://doi.org/10.1186/s40425-015-0089-6.

Spranger, S., and Gajewski, T.F. (2018). Impact of oncogenic pathways on evasion of antitumour immune responses. Nat. Rev. Cancer. https://doi.org/10.1038/nrc.2017.117.

Stanberry, L. (2013). Permutation test. In Encyclopedia of Systems Biology. https://doi.org/10.1007/978-1-4419-9863-7_1186.

Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol. *39*. https://doi.org/10.1038/s41587-020-0739-1.

Suda, K., Kim, J., Murakami, I., Rozeboom, L., Shimoji, M., Shimizu, S., Rivard, C.J., Mitsudomi, T., Tan, A.C., and Hirsch, F.R. (2018). Innate genetic evolution of lung cancers and spatial heterogeneity: analysis of treatment-naïve lesions. J. Thorac. Oncol. *13*. https://doi.org/10.1016/j.jtho.2018.05.039.

Tape, C.J. (2017). The heterocellular emergence of colorectal cancer. Trends Cancer. https://doi.org/10.1016/j.trecan.2016.12.004.

Tauriello, D.V.F., and Batlle, E. (2016). Targeting the microenvironment in advanced colorectal cancer. Trends Cancer. https://doi.org/10.1016/j.trecan.2016.08.001.

Tauriello, D.V.F., Calon, A., Lonardo, E., and Batlle, E. (2017). Determinants of metastatic competency in colorectal cancer. Mol. Oncol. https://doi.org/10.1002/1878-0261.12018.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The immune landscape of cancer. Immunity *48*. https://doi.org/10.1016/j.immuni.2018.03.023.

Tipping, M.E., and Bishop, C.M. (1999). Mixtures of probabilistic principal component analyzers. Neural Comput. *11*. https://doi.org/10.1162/089976699300016728.

Tosun, A.B., Pullara, F., Becich, M.J., Taylor, D.L., Fine, J.L., and Chennubhotla, S.C. (2020). Explainable AI (xAI) for anatomic pathology. Adv. Anat. Pathol. https://doi.org/10.1097/PAP.0000000000000264.

Turley, S.J., Cremasco, V., and Astarita, J.L. (2015). Immunological hallmarks of stromal cells in the tumour microenvironment. Nat. Rev. Immunol. https://doi.org/10.1038/nri3902.

Uttam, S., Stern, A.M., Sevinsky, C.J., Furman, S., Pullara, F., Spagnolo, D., Nguyen, L., Gough, A., Ginty, F., Lansing Taylor, D., and Chakra Chennubhotla, S. (2020). Spatial domain analysis predicts risk of colorectal cancer recurrence and infers associated tumor microenvironment networks. Nat. Commun. *11*. https://doi.org/10.1038/s41467-020-17083-x.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods *16*. https://doi.org/10.1038/s41592-019-0548-y.

Vitale, I., Shema, E., Loi, S., and Galluzzi, L. (2021). Intratumoral heterogeneity in cancer progression and response to immunotherapy. Nat. Med. https://doi.org/10.1038/s41591-021-01233-9.

Wong, P.F., Wei, W., Gupta, S., Smithy, J.W., Zelterman, D., Kluger, H.M., and Rimm, D.L. (2019). Multiplex quantitative analysis of cancer-associated fibroblasts and immunotherapy outcome in metastatic melanoma. J. ImmunoTherapy Cancer *7*. https://doi.org/10.1186/s40425-019-0675-0.

Yaeger, R., Chatila, W.K., Lipsyc, M.D., Hechtman, J.F., Cercek, A., Sanchez-Vega, F., Jayakumaran, G., Middha, S., Zehir, A., Donoghue, M.T.A., et al. (2018). Clinical sequencing defines the genomic landscape of metastatic colorectal cancer. Cancer Cell *33*. https://doi.org/10.1016/j.ccell.2017.12.004.

Zanotelli, V.R., Leutenegger, M., Lun, X., Georgi, F., de Souza, N., and Bodenmiller, B. (2020). A quantitative analysis of the interplay of environment, neighborhood, and cell state in 3D spheroids. Mol. Syst. Biol. *16*. https://doi.org/10.15252/msb.20209798.

Zeng, X., Xia, Y., and Tong, H. (2018). Jackknife approach to the estimation of mutual information. Proc. Natl. Acad. Sci. U S A *115*. https://doi.org/10.1073/pnas.1715593115.

Zhang, A.W., McPherson, A., Milne, K., Kroeger, D.R., Hamilton, P.T., Miranda, A., Funnell, T., Little, N., de Souza, C.P.E., Laan, S., et al. (2018). Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. Cell *173*. https://doi.org/10.1016/j.cell.2018.03.073.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Cell DIVE colorectal cancer data	Gerdes et al., 2013	https://doi.org/10.1073/pnas.1300136110
IMC breast cancer data	Zenodo	https://doi.org/10.5281/zenodo.3518284
Software and algorithms		
LEAPH	This paper	https://github.com/ChakraLab/LEAPH

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, S. Chakra Chennubhotla (chakra@spintellx.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- For academic use, an executable version of this code is publicly available on GitHub as of the date of this publication. DOIs are listed in the key resources table. For non-academic use, please note that the computational and systems pathology intellectual property is owned by the University of Pittsburgh and is exclusively licensed to SpIntellx Inc., Pittsburgh, PA (info@spintellx.com).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Hyperplexed CRC 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) (Gerdes et al, 2013, 2014; Graf and Zavodszky, 2017; McKinley et al., 2017; Nelson et al., 2013; Santamaria-Pang et al., 2017; Uttam et al., 2020) 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 (Gerdes et al., 2013). 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 (Gerdes et al, 2013, 2014; Graf and Zavodszky, 2017; McKinley et al., 2017; Nelson et al., 2013; Santamaria-Pang et al., 2017; Uttam et al., 2020). 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 (Gerdes et al, 2013, 2014; Graf and Zavodszky, 2017; McKinley et al., 2017; Nelson et al., 2013; Santamaria-Pang et al., 2017; Uttam et al., 2020). The biomarkers chosen are protein markers for specific cell lineages, oncogenes, tumor suppressors, and post-translational protein modifications indicative of cellular activation states (Table S1). The data also includes corresponding clinical information including the histological tumor grade, cancer stage, gender, age, and follow up monitoring for 10 years (Table S2).

Data pre-processing

Cell quantitation

Cellular segmentation is done using a collection of structural biomarkers: NaKATPase (cell membrane, border), S6 (cytoplasm), and DAPI (nucleus) (Figure S1A). Protein expression and standard deviation were quantified by the median biomarker intensity value within each cell mask and transformed to the log2 scale (Gerdes et al., 2013). 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 dataset to deceased patients with recurrence within 5 years (post-surgical 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 S1B. The final dataset used is composed of 213 TMA spots (Table S2). 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. To demonstrate properties of patients at the extrema's, we further group patients with no evidence of disease beyond 8 years (NED-8yrs) and evidence of recurrence within 3 years (REC-3yrs) (Table S2).

Biomarker selection and distributions

We removed biomarkers showing batch effects 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 S1D. 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 construction

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 (Λ $(p \times k)$), 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 (Ghahramani and Hinton, 1997). The latent factors, $\vec{z}$, are generated from zero-mean, unit-variance Normal distribution $N(0,I)$, 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 $\Lambda\Lambda^T + \Psi$ (Ghahramani and Hinton, 1997).

Probabilistic 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. Using a probabilistic clustering approach, with a Mixture of Factor Analyzers (MFA), we model the cells as M components (clusters) with the parameters $(\{\pi_j, \vec{\mu}_j, \Lambda_j\}_{j=1}^{M}, \Psi)$ where π_j is the component weight: $p(\vec{x}) = \sum_{j=1}^{M} N(\vec{x} | \vec{\mu}_j, \Lambda_j\Lambda_j^T + \Psi)$. 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 S1C). The expectation-minimization (EM) algorithm is utilized to estimate the model parameters (Ghahramani and Hinton, 1997). 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 order 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 S2A). The MFA model results in probabilistic 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 probabilistic 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: $\min_{\Omega}\left(-\sum_{i=1}^{N}\sum_{j=1}^{M} \Omega_{cj} \log_2(\Omega_{cj}) + \lambda \sum_{(m,n)} w_{mn} ||\Omega_m - \Omega_n||_2 \right)$. The first term minimizes the entropy of the ownership probabilities promoting specialization in cells. The second term promotes spatial coherence where w_{mn} is the weight between cell m and cell n and is computed as the reciprocal of the distance between two cells: $w_{mn} = \frac{1}{dist(cell_m,\ cell_n)}$. 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 (Francis and Palsson, 1997).

The objective function is optimized using Alternating Directions Method of Multipliers (ADMM) (Hallac et al., 2015). 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: $\lambda = \frac{N_{opt} \times \text{maxEntropy}}{\sum_{\{m,n\}} w_{mn}}$ 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 homogeneous 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

We make two key observations about the hyperplexed data: 1) a varying spread of biomarker values likely due to the individual biomarker response sensitivity (Figure S1D) and 2) existence of dominant cellular phenotypes (e.g., epithelial and stromal cells) and nested cell subtypes (e.g., immune, cancer stem cells) derived from a cellular phenotypic continuum. To account for these observations and automate the process of cellular phenotype discovery, we propose a recursive probabilistic approach where each step attempts to dissect the most dominant clusters with $M = 2$ components.

At each recursive step, the probabilistic clustering step utilizes a low-dimensional latent space MFA. Subsequently, within each recursive step, 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 the stopping criteria based on the angle between the mean vector and factor loading space.

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: $\Delta_j = \left|\vec{\mu}_{1j} - \vec{\mu}_{2j}\right| * \max(\vec{\mu}_{1j},\ \vec{\mu}_{2j})$. 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.

Analyzing non-specialized cell states

To investigate the likelihood of non-specialized cells holding ownership probabilities shared between an FP-pair, i and j, we compute the inner product between the ownership probability vectors, $<\Omega_i,\ \Omega_j>$. The inner product determines the total sum of shared ownership probability between the FP-pair, i and j. Generally, we perform this computation as $\Omega^T\Omega$ and visualize in Figure 3B (and Figure S5B).

FP-specific biomarker positive (+) signatures

We follow the approach of Schurch et al. in using cellEngine (www.cellengine.com) to validate LEAPH (Schürch et al., 2020). We use a threshold for each biomarker equal to its mean to determine biomarker-positive (+) cells. With this threshold in place for every biomarker, each cell is now represented as a vector, $\vec{b}\ (p \times 1)$, with values 1 or 0 if the biomarker expression is positive or negative, respectively. To summarize each FP with a biomarker-positive (+) signature, we compute a biomarker-positive (+) fraction for each FP, j, biomarker, k, pair defined as: $B_{jk} = \frac{\sum_{c=1}^{N} \Omega_{cj} \vec{b}_{ck}}{\sum_{c=1}^{N} \vec{b}_{ck}}$.

Biomarker selection virtual simulation

For comparison, we assume the LEAPH derived FP'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 S5). Assigning each cell to one FP 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 S4A). This virtual simulation demonstrates the reproducibility and capabilities to integrate LEAPH into an integrated iterative imaging and computational platform.

Discovery of recurrence-associated microdomains

Step 1: patient selection

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 the two groups, NED-8yrs (N = 45) and REC-3yrs (N = 46). See Table S2 for patient statistics.

Step 2: spatial network of neighboring cells

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: PMI for spatial co-occurrence of FPs

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 (Leinster and Cobbold, 2012).

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

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

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:

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

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:

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

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 FP-pair spatially co-occurs more, less, or the same as random. To construct the random distribution, we set the probability of each FP to the probability of each FP 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) (Zeng et al., 2018). 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: $b_{jack} = (n-1)(\overline{T}_n - T_n)$, where $\overline{T}_n = \frac{1}{n}\sum_i T_{(-i)}$. 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 FP-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 FP-pair for a patient, p, as the summation of ownership probabilities, $N^p_{(f_i,\, f_j)}$, over all cells in the network: $N^p_{(f_i,\, f_j)} = \sum_{c}^{N^p} \overrightarrow{\Omega}_{cf_i} + \overrightarrow{\Omega}_{cf_j}$. 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.

Step 4: identify significant FP-pairs

For each FP-pair, (f_i, f_j), we aggregate the PMI values to form two distributions (NED-8yrs, REC-3yrs) (Figure S3D). We perform a permutation test (10,000 permutations) using a test statistic, t, to evaluate the absolute difference between NED-8yrs and REC-3yrs PMI distribution means (Table S3) (Stanberry, 2013). FP-pairs are identified as being significant based on the resulting permutation test p value ($p_t < 0.05$) (Figure 4B). We found the significant FP-pairs form two microdomains: uD_1 — an interconnecting network between FP1, FP5, FP6, FP7, FP9, FP12, FP13, uD_2 — a pairwise interaction between FP2, FP4.

Visualizing cells contributing to each microdomain- Each microdomain consists of a set of FP-pairs (e.g., uD_2 = FP2:FP4). Each cell and its neighbor within 100 pixels (50 μm) contribute to the relevant uD with a probability that is determined by the joint product of their ownership probabilities. For example, for FP-pair FP2:FP4, we sum the product of probabilities of the reference cell, c, and its neighbor d belonging to FP2 and FP4 respectively, and vice versa to determine the overall contribution of the link connecting c and d to the microdomain with an FP-pair FP2:FP4.

Step 5: spatial network biology analysis

For each microdomain, we aim to identify a biomarker network representation for the NED-8yrs and REC-3yrs patient groups. Within a single microdomain ($\{f_i\}_{i \in uD}$) and each patient cohort, we gather all cells, c, where $\sum_{\{f_i\}_{i \in uD}} \Omega_{ci} > 0$ and compute the partial correlation between each biomarker-pair, controlling for the remaining confounding biomarkers (Uttam et al., 2020). First, using the biomarker data, X ($N \times p$), and weight vector, $\overrightarrow{w}$ $(p \ \times 1)$, where the weight for each cell, c, is the normalized sum of the ownership probabilities of each FP in the microdomain ($\overrightarrow{w}_c = \frac{\sum_{\{f_i\}_{i \in uD}} \Omega_{ci}}{\sum_{c=1}^{N} \sum_{\{f_i\}_{i \in uD}} \Omega_{ci}}$), we compute the weighted covariance matrix, C ($p \times p$) defined as: $C = \widehat{X}^T W \widehat{X}$ where $\widehat{X} = X - \overrightarrow{w}^T X$ and W denotes the weights as a diagonal matrix (Pozzi et al., 2012). Second, we invert the covariance matrix to obtain the precision matrix, $\Phi = C^{-1}$. Third, we compute the partial correlation coefficient between each biomarker pair, ρ_{ij}, through normalization: $\rho_{ij} = \ - \frac{\Phi_{ij}}{\sqrt{\Phi_{ii} \Phi_{jj}}}$. The partial correlation coefficients range from -1 to 1 representing the true correlation between each biomarker pair when all other confounding factors (other biomarkers) are removed (Epskamp and Fried, 2018). For the sake of simplicity, we chose to pool cells across the two patient outcome cohorts: total number of cells in NED-8yrs cohort = 89,963 and REC-3yrs cohort = 89,596 (Table S2). However, one could also construct patient-specific networks, and based on the number of cells, an additional step of matrix regularization may have to be incorporated.

To compare the partial correlation networks between the cohorts, NED-8yrs (ρ^{NED}) and REC-3yrs (ρ^{REC}), we use differential connectivity analysis with a permutation test. We define a symmetric differential connectivity matrix, Δ_0:$p \times p$, where each entry, (i,j), represents the differential connectivity between a biomarker pair: $\Delta_0(i,j) = \left| \rho_{ij}^{NED} - \rho_{ij}^{REC} \right|$. To test the significance of the differential, we use a permutation test (with B = 10,000 permutations). For each permutation iteration, the patients are randomly shuffled across the NED-8yrs and REC-3yrs cohort and sampled to mirror the same number of tissue samples as the original NED-8yrs and REC-3yrs cohorts. Each permutation, k, results in a differential connectivity matrix of the permuted data, Δ_k. The significance value (p value) for a specific biomarker-pair (i,j) is computed as $\frac{b+1}{B+1}$ where $b = \sum_{k=1}^{B} I(\Delta_0(i,j) \leq \Delta_k(i,j))$. We define the biomarker-pairs with a differential connectivity score in the top 99th percentile and p value below 0.05 as significant (Table S4).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical details and data analysis methods used for this study are cited in the appropriate sections in the STAR Methods.