Comp790-166: Computational Biology

Lecture 15

March 6, 2023

Today

- Spatial Phenotyping with LEAPH
- *Lab Activity to Play with Trajectory Inference*

Project Related Announcements

- Please sign up with your group members and topics here, `https://docs.google.com/document/d/1x9mIJCZAkeogAhmGlpqJkXwuoAKlB_0gewV1LpXZDoU/edit?usp=sharing`
- Writeup due before spring break on March 8.
- Presentations will be the week after spring break. Stay tuned for your date and time.

Do You Remember Question

1. What was the main idea of the feature selection method in SLICER?
2. What kind of information can you get from spatial profiling modalities?

CyTOF + Spatial Resolution

An upgrade of regular CyTOF to image 32 proteins and their modifications at cellular resolution.

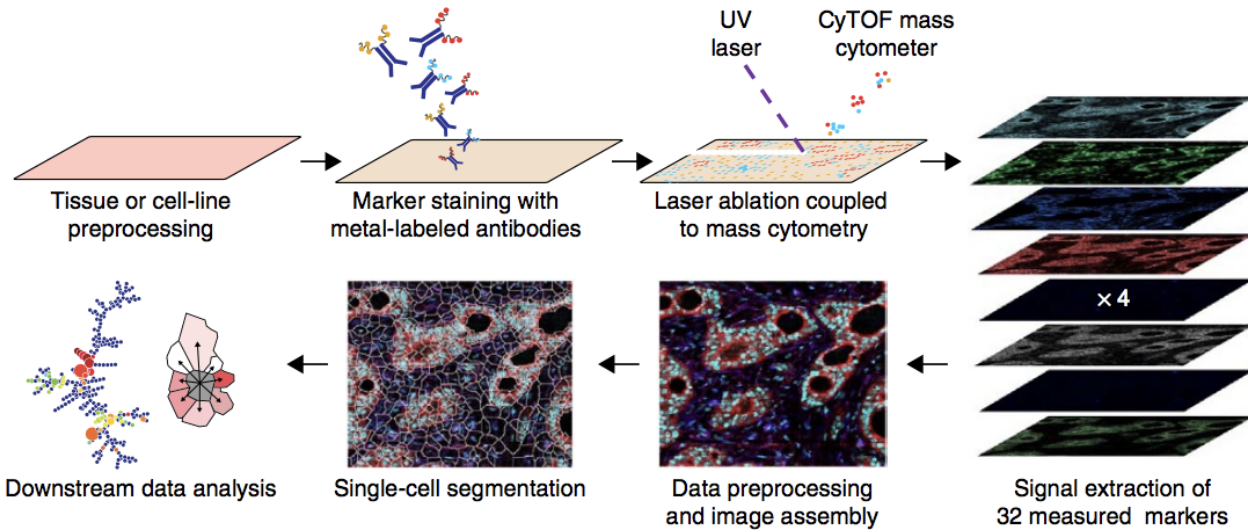

Figure: from Giesen *et al.* Nature Methods. 2016

Recent Advances in Study The Relationship Between Immune Cells and Tumor

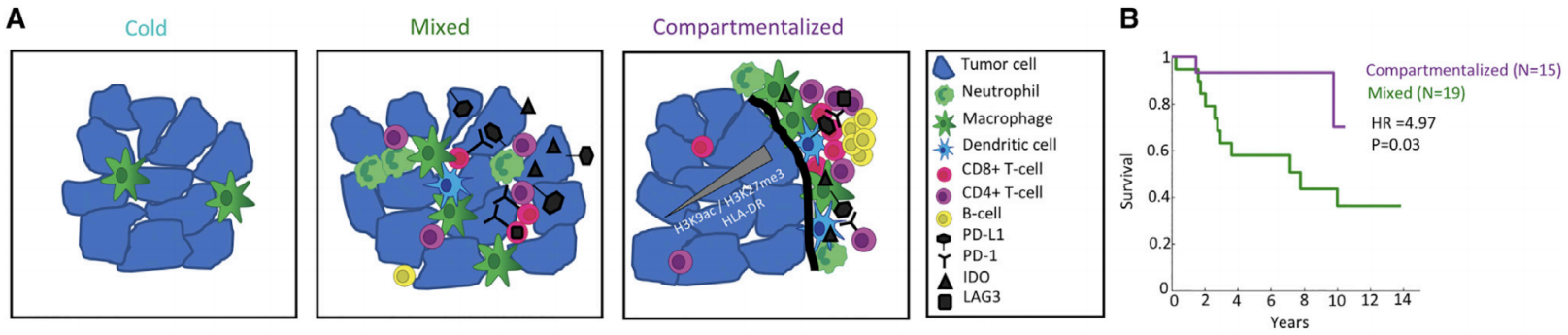

Figure: from Keren *et al.* Cell. 2018.

End-Goal of Identifying Particular Microenvironments

Ultimately, an objective is to identify ‘micro-environments‘ or spatially-localized subsets of cells with characteristic frequency patterns that are predictive of some outcome of interest.

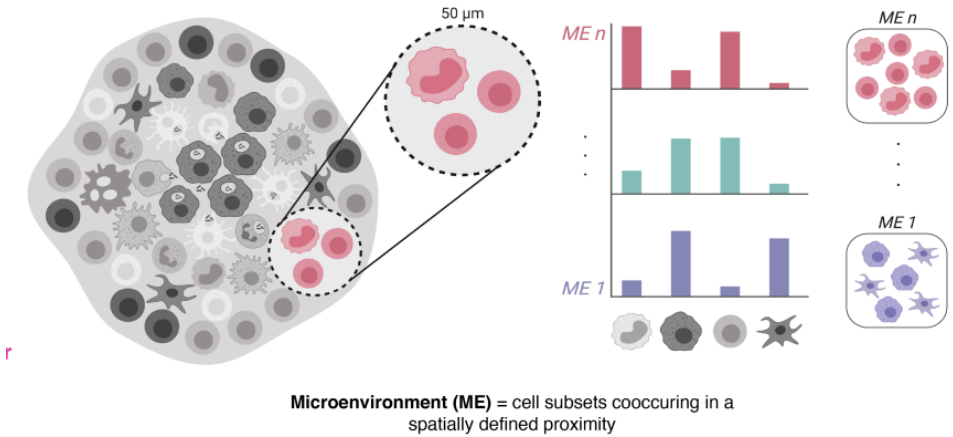

Figure: from `https://www.biorxiv.org/content/10.1101/2020.06.08.140426v1.full.pdf`

A New Problem: Identifying Microdomains

Welcome LEAPH. One of the first methods out there to identify phenotypically distinct microdomains of spatially configured cell phenotypes.

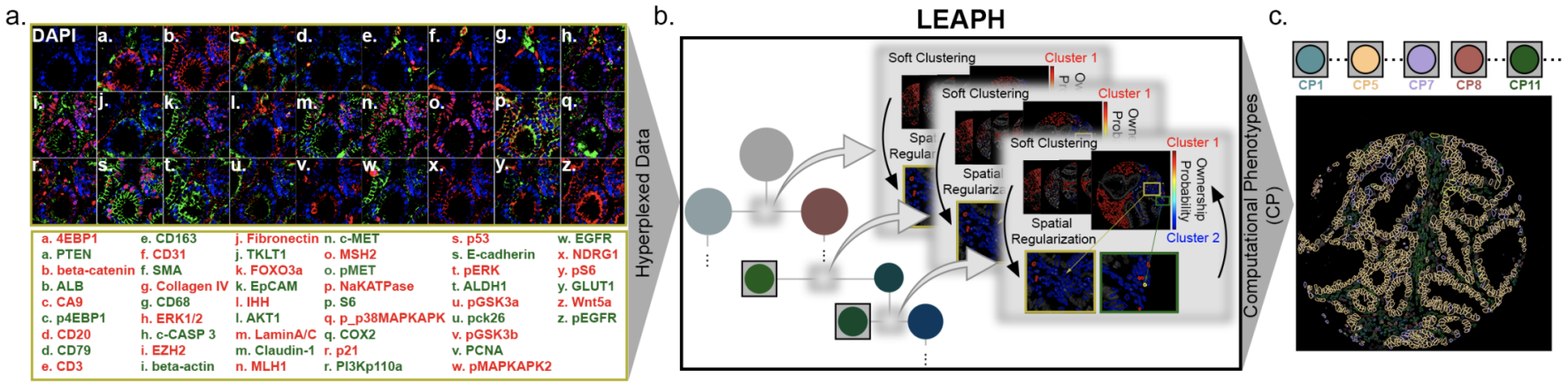

Figure: from Furman *et al.* Cell Reports Methods. 2021. Probabilistic clustering, incorporating spatial information can capture transitional states along a phenotypic continuum.

LEAPH Overview

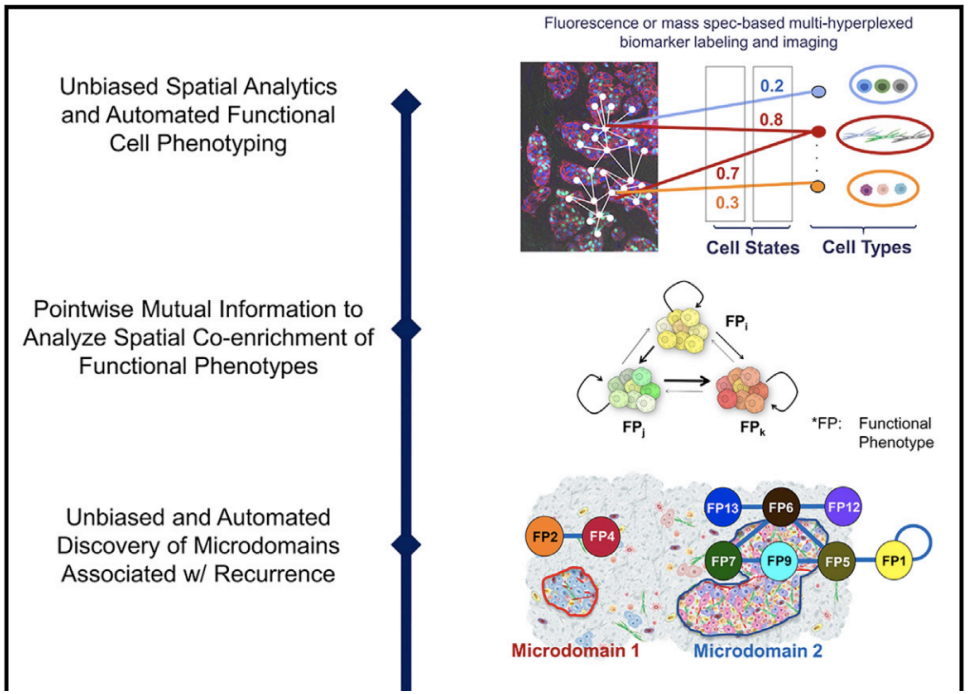

Figure: from Furman *et al.* Cell Reports Methods. 2021. Functional responses of cells are determined both by internal state and interactions with neighbors and stimulus from local environment.

Notation in LEAPH

- For cell i, let its protein expression be represented as $\mathbf{x}_i \in \mathbb{R}^p$.
- Mixture of factors setup, with k dimensions in the latent space, with $\mathbf{x}_i = \Lambda \mathbf{z} + \boldsymbol{\mu} + \mathbf{v}$
  - Loadings in $\Lambda \in \mathbb{R}^{p \times k}$
  - Latent variables, $\mathbf{z} \in \mathbb{R}^{k \times 1}$. Assume zs generated from a standard normal with 0 mean and unit variance.
  - Noise term via, $\mathbf{v} \sim \mathcal{N}(0, \Psi)$
  - Mean vector, $\boldsymbol{\mu} \in \mathbb{R}^{p \times 1}$
  - Under this formulation, each $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Lambda\Lambda^T + \Psi)$

Probabilistic Clustering Setup

The mixture of factor analyzers setup with the notation that we defined can be written as,

$$(\{\pi_j, \boldsymbol{\mu}_j, \lambda_j\}_{j=1}^{M}\}, \Psi)$$

.

Note that π_j is the component or mixing weight for component j.

Mixture Model

Each $p(\mathbf{x}_i)$ is computed as

$$p(\mathbf{x}_i) = \sum_{j=1}^{M} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \Lambda_j \Lambda_j^T + \Psi)$$

- π_j is the mixing weight for cluster j.

The parameters, λ and Ψ are updated in a close-form way using the EM algorithm. See here,
`https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=766f4465747394d304d162197e091f1ae8f7f577`.

Practicalities

- Overall, parameters being estimated are $\{\pi_j, \boldsymbol{\mu}_j, \Lambda_j\}_{j=1}^{M}, \Psi)$.
- They 2-dimensions for each latent space, so, $k = 2$.
- Ultimately, they get a prediction that each cell belongs to of the M components, and in particular for class j, $p(j \mid \mathbf{x}_i) = \frac{p(\mathbf{x}_i|j)p(j)}{\sum_{c=1}^{M} p(\mathbf{x}_i|c)p(c)}$
- Use the estimated probability between a cell i and a cluster c and create a matrix, $\Omega \in \mathbb{R}^{N \times M}$ where Ω_{ic} gives the probability that cell i belongs to cluster c.
- This gives a soft clustering interpretation for each cell.

Spatial Regularization Intuition

- Based on prior biological knowledge, there are known properties that for example, epithelial/tumor cells should be surrounded by or spatially proximal to other epithelial/tumor cells.
- There should also be some allowance for tumor-infiltrating cells, such as lymphocytes and other stromal cells.

A new Ω is optimized that encodes spatial information as follows,

$$\min_{\Omega} -\sum_{i=1}^{N}\sum_{j=1}^{M} \Omega_{ij} \log_2 (\Omega_{ij}) + \lambda \sum_{(m,n)} w_{mn} ||\Omega_m - \Omega_n||_2$$

Unpacking

$$\min_{\Omega} -\sum_{i=1}^{N}\sum_{j=1}^{M} \Omega_{ij} \log_2 (\Omega_{ij}) + \lambda \sum_{(m,n)} w_{mn} ||\Omega_m - \Omega_n||_2$$

- w_{jk} is a weight, calculate as the reciprocal of distance between cells j and k in the image
- The first term is basically an entropy term of ownership confidence
- The second term is promoting spatial coherence.
- λ controls the tradeoff between spatial coherence and membership confidence.

Effect of Spatial Regularization

In particular in the first example, a cell with a highly predicted assignment towards CP1 transitioned towards a phenotype of CP2 after spatial regularization.

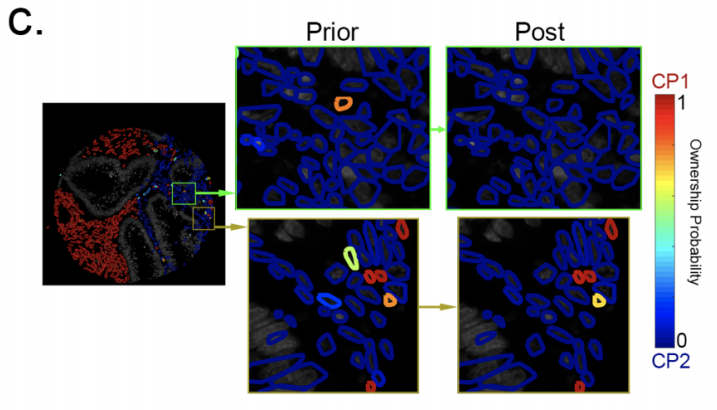

Figure: from Fig. 2 of `https://www.biorxiv.org/content/10.1101/2020.10.02.322529v3.full.pdf`

Determining Specialized Cells

- Based on the Ω, assign each cell to one of the M phenotypes based on the j that gives the maximum probability.
- For a particular patient, p, create a feature vector $\mathbf{f}_p$ which gives the proportion of its cells assigned to each of the cell phenotypes.
- At times, the authors refer to specialized cell-types (membership probability $> 95\%$) in contrast to transitional and rare cells.

Recap and Transition

- The clustering part is straight-forward : Assume each cell is from one of M 2-dimensional latent factors
- Calculate a probability that each cell was from each of these latent factors
- Add penalties that enforce spatial coherence and certainty of assignment
- **Next step:** Identify microdomains with a collection of cells that are predictive of some penotype of interest.

Predicting Time to Recurrence in Colorectal Cancer

- Consider cohorts of patients with the following properties.
  - 45 patients in ‘NED-8‘ category that have no evidence of disease for over 8 years
  - 46 patients in ‘NED-3‘, where cancer came back within 3 years.

The goal is to translate the distributions of cell phenotypes that spatially co-occur to a signal that can be used for prediction.

Constructing a Cell Network For Each Patient

- Connectivity is determined by proximity in the image of the tissue
- For a pair of cells, m, and n, connect them with a weights, $w_{mn} = 1$ if their spatial distance, $d_{mn} < \tau$.
- Otherwise, $w_{mn} = 0$ and there are no edge between the cells

Identifying Spatial Co-Occurrence Between Cell Phenotype Pairs

Consider two phenotypes, f_i and f_j for a given set (e.g. a subset of patients, etc). The pairwise mutual information between these two phenotypes is defined as,

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

- $p(f_i^s)$ is the probability of a particular phenotype, i occurring in a network set, s.
- $p(f_i^t)$ is the background probability of phenotype i.

Calculating Joint Phenotypic Probability for a Single Patient

Letting Ψ encode the set of edges for a particular patient, the joint probability of phenotypes i and j is given as,

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

*Here z is a normalization over all combinations of i and j according to the computational phenotypes.

Specifying a Background Distribution

The background probability for a phenotype, i is simply the mean assignment probability over all cells, or,

$$p(f_i^t) = \frac{1}{N} \sum_{c=1}^{N} \Omega_{ci}$$

Ultimately, for each cell phenotype pair, (f_i, f_j) compute the PMI for each sample and consider how this relates to the patient re-occurence outcomes.

Looking at Significant Microdomains Between Groups

There were a few cellular phenotypes that tended to co-occur between the two patient groups.

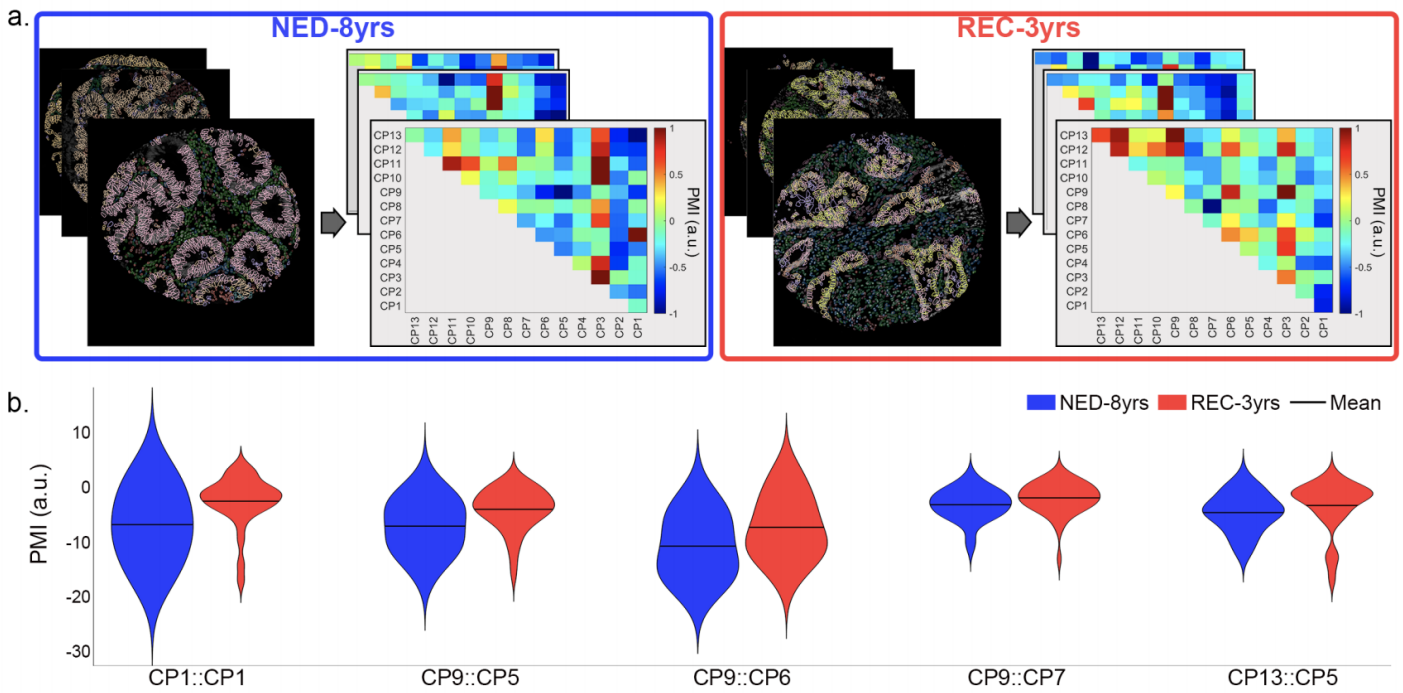

Figure: from Fig. 4 in `https://www.biorxiv.org/content/10.1101/2020.10.02.322529v3.full.pdf`

Now, switching gears to trajectory activity

Please team up and go here, `https://colab.research.google.com/drive/14LRc76ltDEnHi4v2jUWCS9Tad-RbXOLz?usp=sharing#scrollTo=zU82kaKhfKEN`.
Things to try:

- Put clusters as input to PAGA rather than ground-truth cell labels
- Try varying how the clustering is done and see how it affects the input and psuedo time
- Choose other genes to visualize with respect to the trajectory
- Change root cells to something like intuitive.