Comp790-166: Computational Biology

Lecture 20

April 15, 2021

Announcements

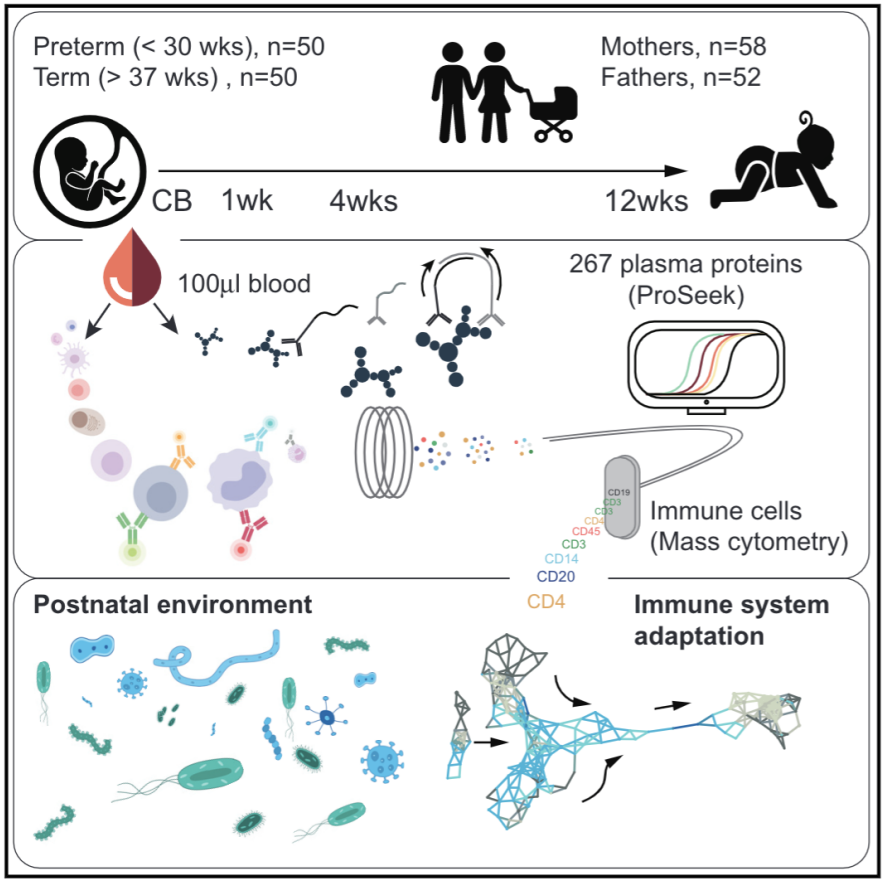
- Homework 2 is online. Due April 23. `https://github.com/stanleyn/Comp790-166-Comp-Bio/tree/main/Homework2`
- Project signup sheet will appear on the website today.
- Comments on projects? How are they going?

Homework Comment

Defining features based on hop neighborhoods of each node. These features defined by me are simplistic. Feel free to modify!

CompMed Seminar Today

Petter Brodin/2pm. Some of the coolest single cell work out there.



Today

- Spatial regularization in single-cell analysis
- LEAPH
- SpiceMix

CyTOF + Spatial Resolution

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

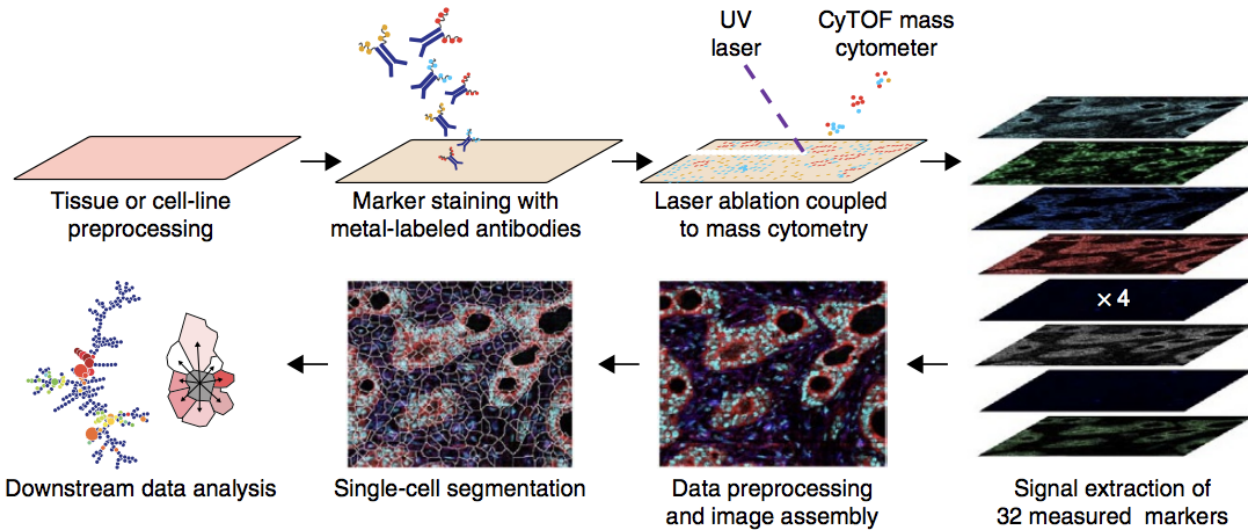


Figure: from Giesen *et al.* Nature Methods. 2016

Why Do We Care?

Understanding the spatial organization of cells (for example, tumor and immune cells) can provide a more mechanistic understanding of the underlying biology. This can further translate to more accurate prediction of prognostic outcomes.

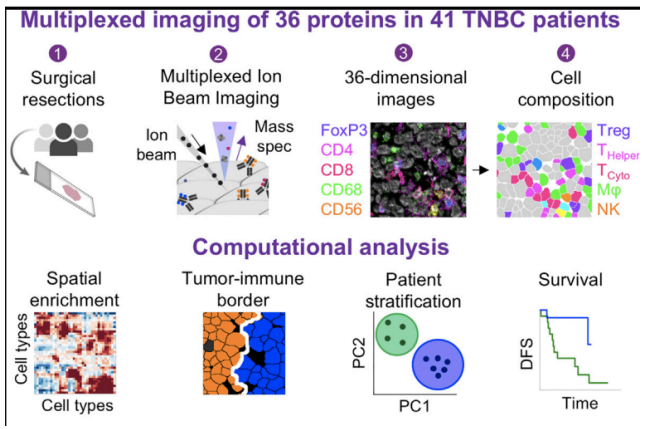


Figure: from Keren *et al.* Cell. 2018.

Recent Advances in Study The Relationship Between Immune Cells and Tumor

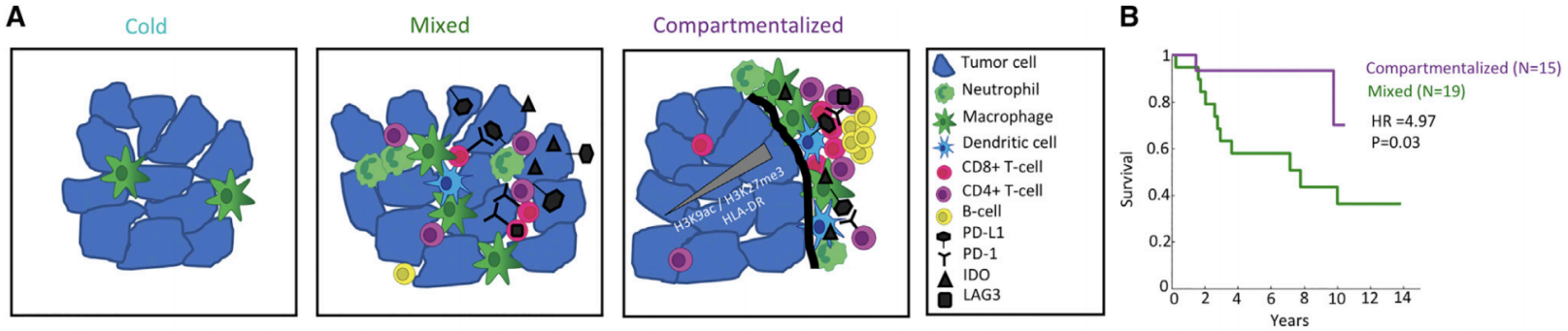


Figure: from Keren *et al.* Cell. 2018.

Studying Aging

Older mice were observed to have infiltrating T-cells in their neurogenic niches (the collection of neuronal progenitor cells)

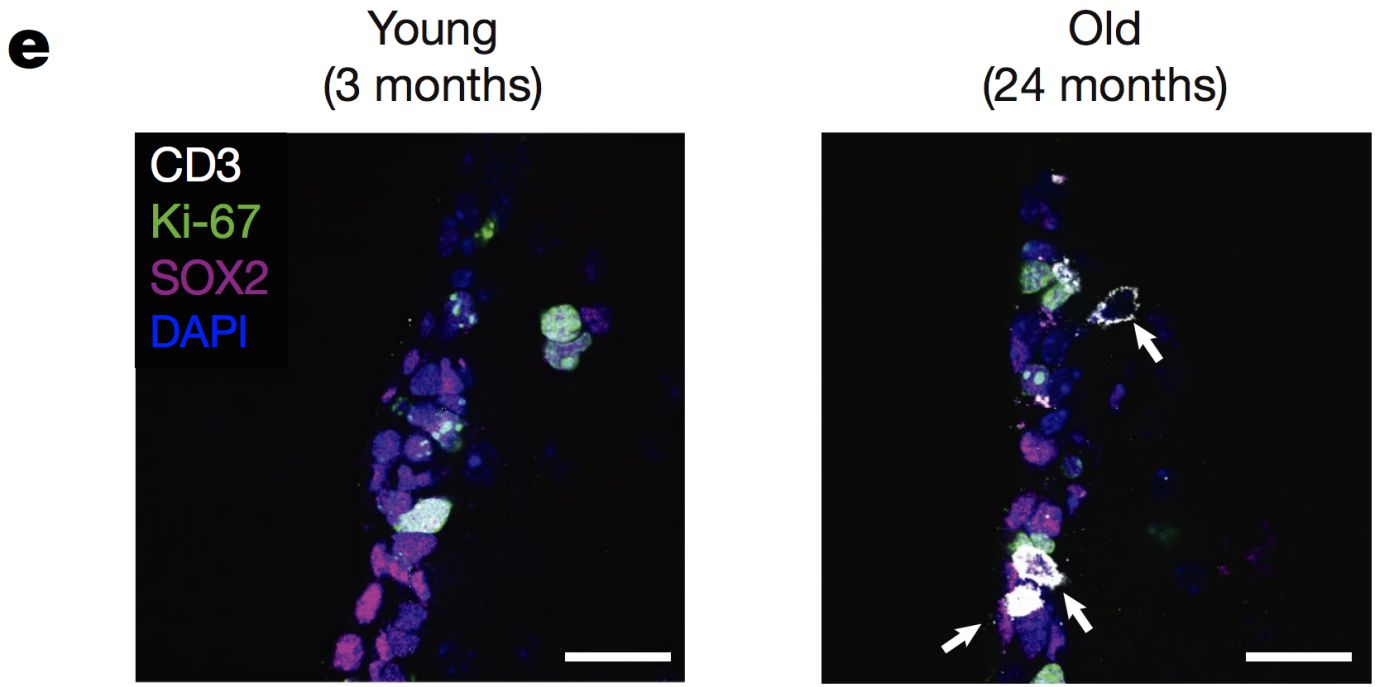
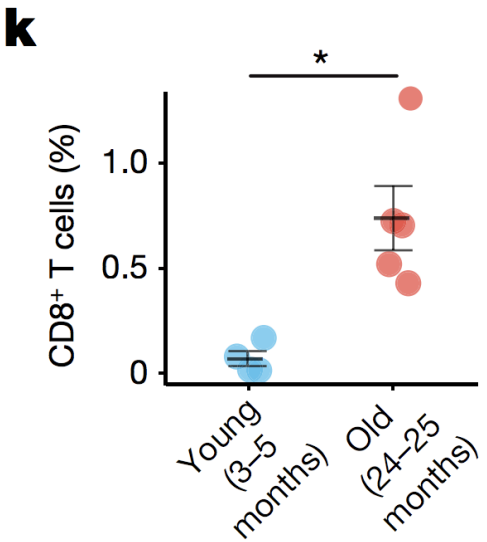


Figure: from Dulken *et al.* Nature 2019

Counting CD8+ T-cells

You can even compare the proportion of CD4+ T-cells there are in neurogenic niches between young and old mice. It's a pretty striking difference.



General Steps in Analyzing These Data

- Segmentation of cells
- Phenotype cells
- Identify microenvironments or characteristic co-occurences of particular cell-types within a region.

Example-Cell Phenotype Map

Cells are clustered and phenotyped according to protein expression.

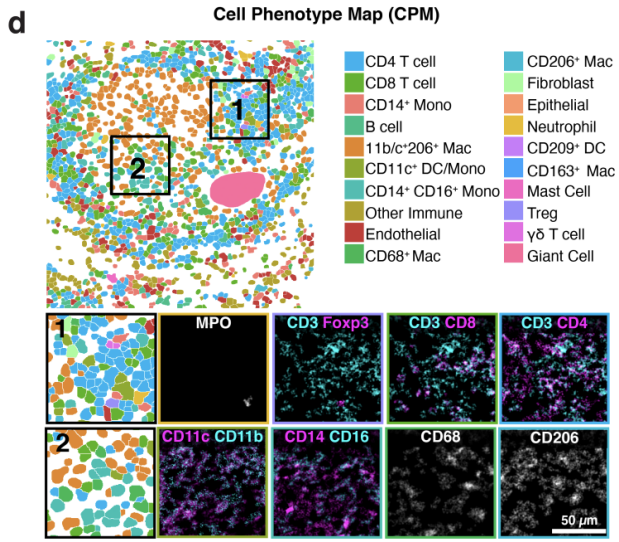


Figure: from `https://www.biorxiv.org/content/10.1101/2020.06.08.140426v1.full.pdf`

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.

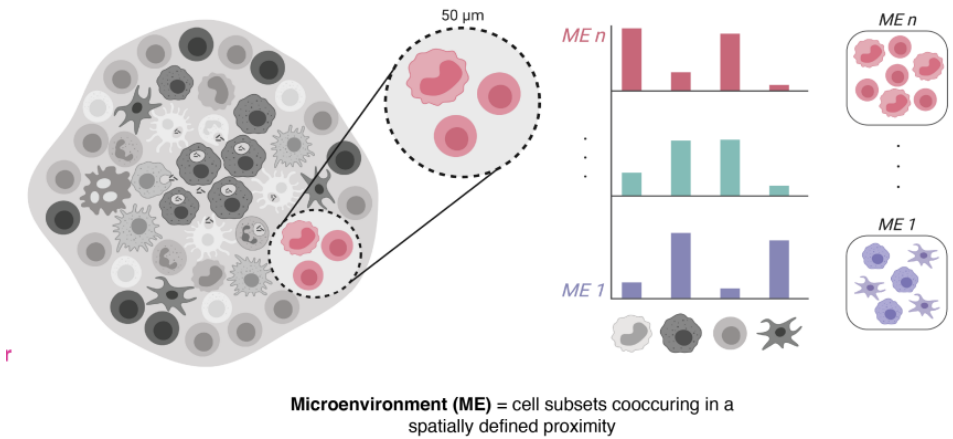


Figure: from `https://www.biorxiv.org/content/10.1101/2020.06.08.140426v1.full.pdf`

A New Problem: Identifying Microenvironments

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

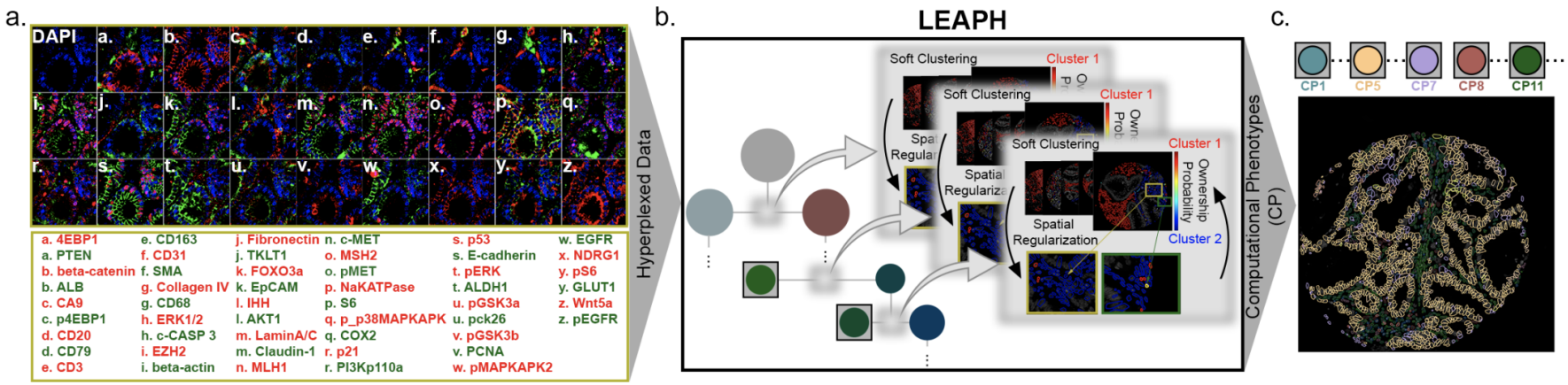


Figure: from Furman *et al.* BioArXiv. 2020.

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, $\mathbf{z} \in \mathbb{R}^{k \times 1}$
  - Noise term via, $\mathbf{v} \sim \mathcal{N}(0, \Psi)$
  - Mean vector, $\boldsymbol{\mu} \in \mathbb{R}^{p \times 1}$

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.

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,

$$\min_{\Omega} -\sum_{i=1}^{N}\sum_{j=1}^{M} \Omega_{ij} \log_2 (\Omega_{ij}) + \lambda \sum_{(j,k)} w_{jk} ||\Omega_j - \Omega_k||_2$$

Unpacking

$$\min_{\Omega} -\sum_{i=1}^{N}\sum_{j=1}^{M} \Omega_{ij} \log_2 (\Omega_{ij}) + \lambda \sum_{(j,k)} w_{jk} ||\Omega_j - \Omega_k||_2$$

- 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.

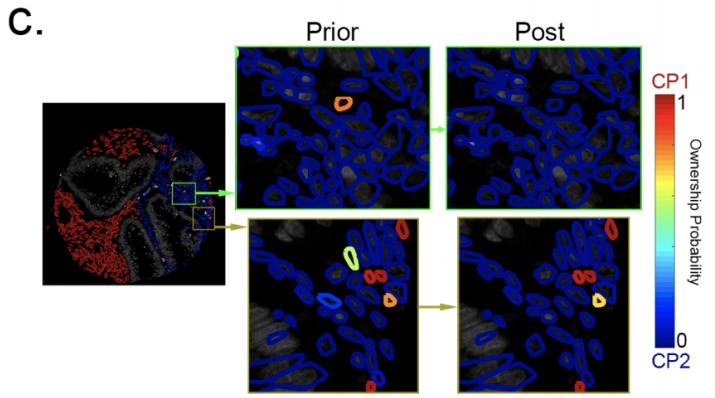


Figure: from Fig. 2 of `https://www.biorxiv.org/content/10.1101/2020.10.02.322529v3.full.pdf`

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 Breast 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, $d_{mn} < 1$.
- 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,

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

*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.

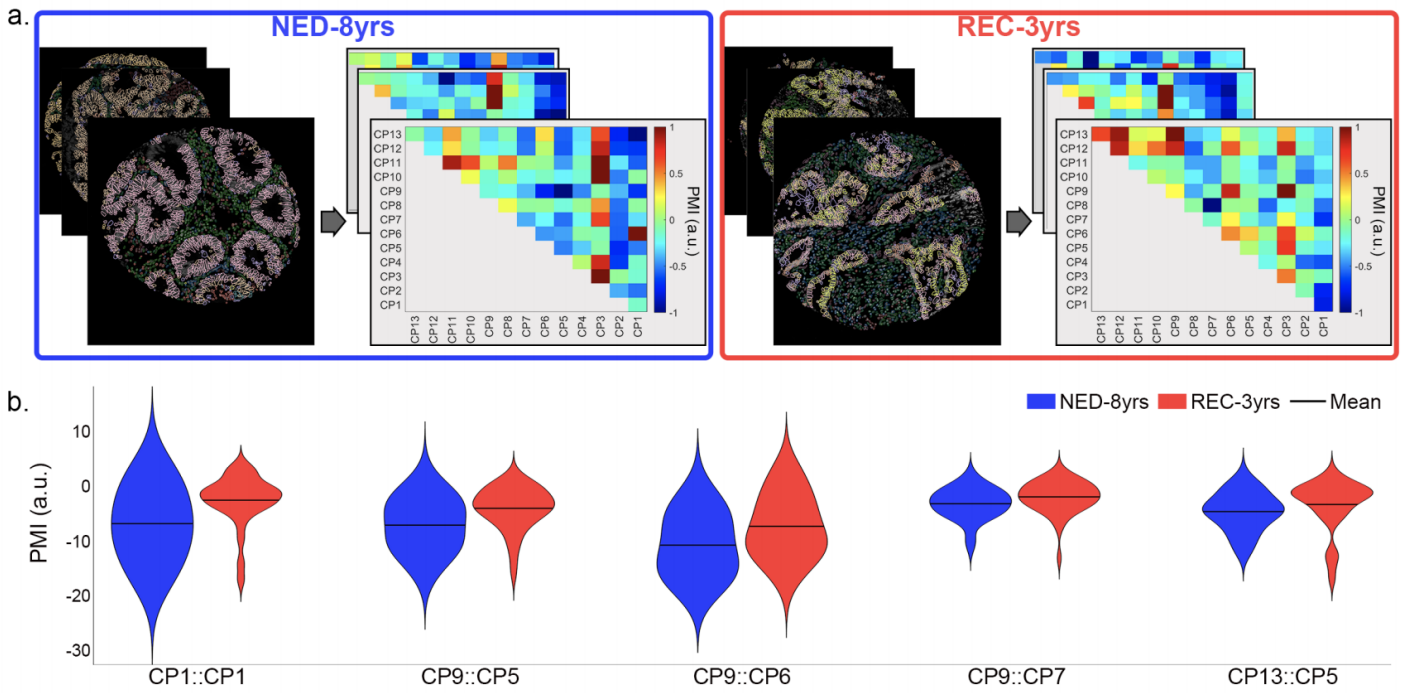


Figure: from Fig. 4 in `https://www.biorxiv.org/content/10.1101/2020.10.02.322529v3.full.pdf`

SeqFish+ and Cortex Cell-Class Composition

Each layer of the cortex has a different distribution of cells profiled through gene expression with SeqFish+ technnology.

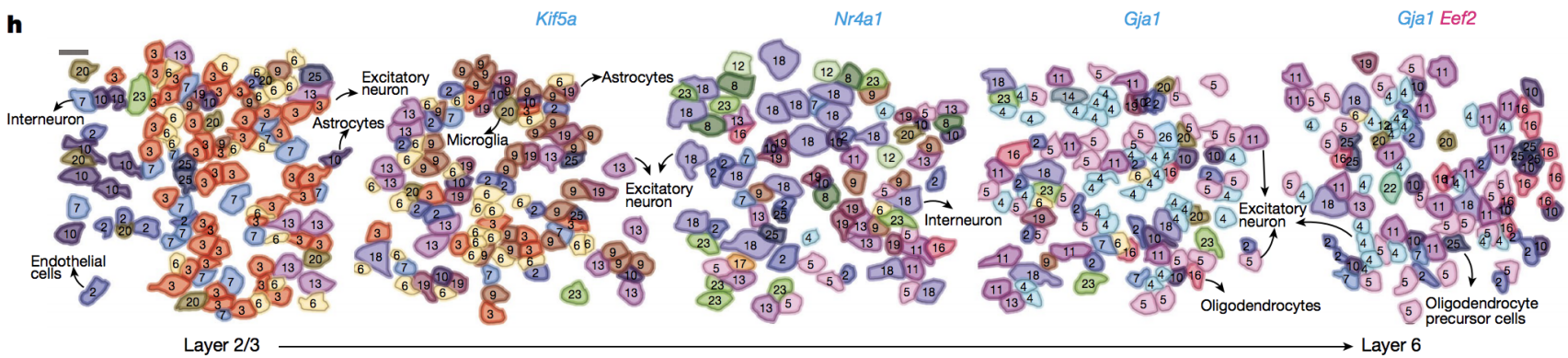


Figure: from Eng *et al.* Nature. 2019

Welcome SpiceMix

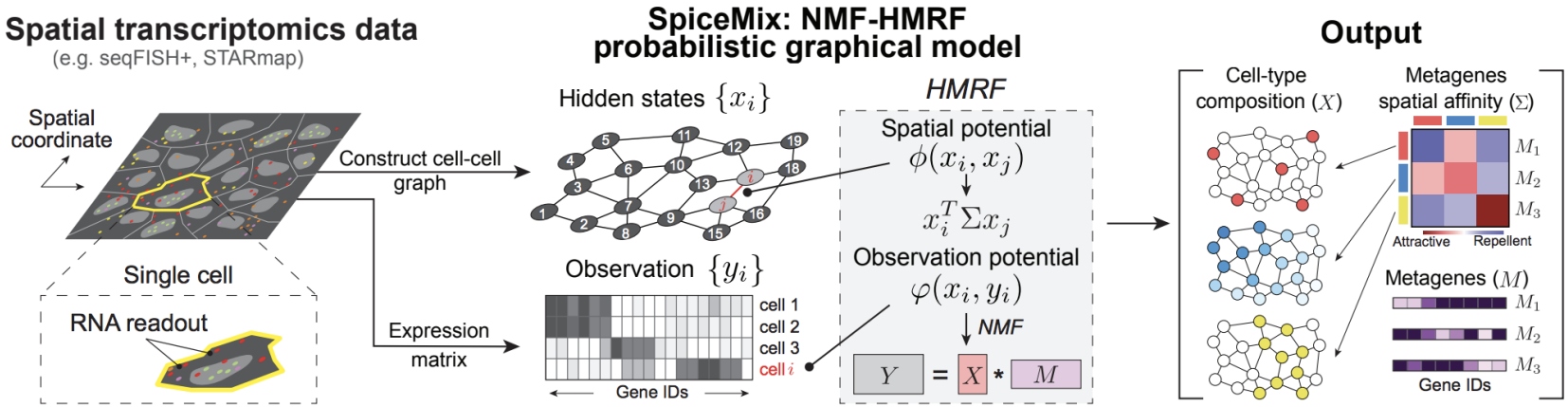


Figure: from Fig. 1 of `https://www.biorxiv.org/content/biorxiv/early/2020/11/30/2020.11.29.383067.full.pdf`

NMF + gene expression

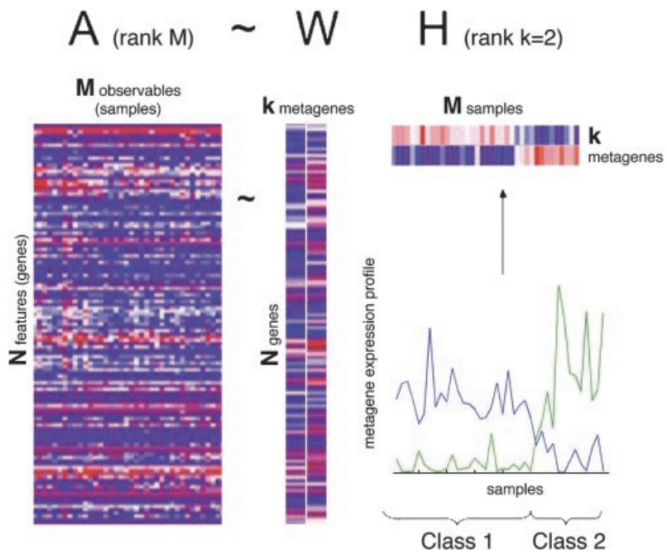


Figure: from Brunet *et al.* PNAS. 2004. The W matrix is recording the coefficient of each gene in each metagene.

Setup

Consider the expression of G genes in N cells.

- Let $Y \in \mathbb{R}^{G \times N}$ be the matrix of gene expression in cells.

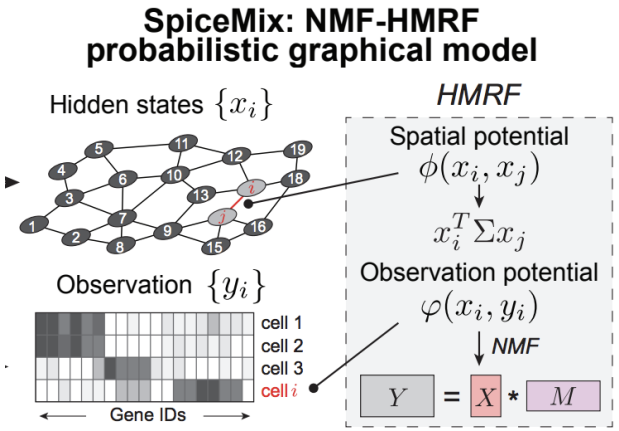
In an NMF formulation, let this matrix of genes and cells be the product of metagenes and weights as,

$$Y = MX + E$$

- M is a matrix of metagenes, with $M \in \mathbb{R}^{G \times K}$
- X is a matrix of weights, $X \in \mathbb{R}^{K \times N}$

Enhanced NMF

For a cell, i, the y_is are observed and the x_is are the hidden variables. The observed and hidden variables are related to each other through a potential function, ϕ. Another potential function φ will capture the spatial dependencies.



Incorporating Spatial Consistency

Given a graph between cells according to spatial coordinates, a hidden Markov random field (HMRF) formulation is used to model $P(Y, X \mid \Theta)$ as,

$$P(Y, X \mid \Theta) = \frac{1}{Z(\Theta)} \prod_{(i,j)\in\mathcal{E}} \varphi(x_i, x_j) \prod_{i\in\mathcal{V}} \phi(y_i, x_i)\, \pi(x_i)$$

What are these potentials?

$$\phi\left(y_i, x_i\right) = \exp\left(-U_y\left(y_i, x_i\right)\right), \quad \varphi\left(x_i, x_j\right) = \exp\left(-U_x\left(x_i, x_j\right)\right)$$

The energy functions are given by:

$$U_y\left(y_i, x_i\right) = \frac{\left(y_i - M x_i\right)^2}{2\sigma_y^2}, \quad U_x\left(x_i, x_j\right) = \frac{x_i^\top}{\|x_i\|_1} \Sigma_x^{-1} \frac{x_j}{\|x_j\|_1}$$

- U_y is the reconstruction error of the measured expression of cell i according to the estimated x_i
- U_x measures the inner product between the normalized factorization of neighboring vells i and j, weighted by a leaned pairwise correlation matrix, Σ_x^{-1}, which captures the spatial affinity of metagenes.
- σ_y^2 is the variation of expression, or noise, of the NMF model.

Approximating the Joint Probability of Hidden States

The joint probability of the hidden states are approximated by pesueolikelihood as follows as the conditional probability of the hidden states given their neighbors as,

$$P(X \mid \Theta) \approx \prod_{i \in \mathcal{V}} P\left(x_i \mid x_{\eta(i)}, \Theta\right)$$

Here $\Theta = \{\Delta, M\}$ is the set of model parameters and metagenes

The MAP Estimate of X

The MAP estimate of X is given by,

$$\hat{X} = \underset{X \in \mathbb{R}_+^{K \times N}}{\operatorname{argmax}} P(X \mid Y, \Theta) = \underset{X \in \mathbb{R}_+^{K \times N}}{\operatorname{argmax}} P(Y, X \mid \Theta) = \underset{X \in \mathbb{R}_+^{K \times N}}{\operatorname{argmax}} \{\log P(Y, X \mid \Theta)\}$$

$$= \underset{X \in \mathbb{R}_+^{K \times N}}{\operatorname{argmax}} \left\{ \sum_{i \in \mathcal{V}} \left[-U_y(y_i, x_i) + \log \pi(x_i) \right] - \sum_{(i,j) \in \mathcal{E}} U_x(x_i, x_j) \right\}$$

Updating Θ

Given an estimate for X, Θ can be updated as,

$$\hat{\Theta} = \underset{\Theta}{\text{argmax}} P(\Theta \mid Y, X) = \underset{\Theta}{\text{argmax}} P(Y, X \mid \Theta) P(\Theta) = \underset{\Theta}{\text{argmax}} \{\log P(Y, X \mid \Theta) + \log P(\Theta)\}$$

$$= \underset{\Theta}{\text{argmax}} \left\{ \sum_{i \in \mathcal{V}} \left[-U_y \left(y_i, x_i \right) + \log \pi \left(x_i \right) \right] - \sum_{(i,j) \in \mathcal{E}} U_x \left(x_i, x_j \right) - \log Z(\Theta) + \log P(\Theta) \right\}$$

$$\approx \underset{\Theta}{\text{argmax}} \left\{ \sum_{i \in \mathcal{V}} \left[-U_y \left(y_i, x_i \right) + \log \pi \left(x_i \right) - \log Z_i(\Theta) \right] - \sum_{(i,j) \in \mathcal{E}} U_x \left(x_i, x_j \right) + \log P(\Theta) \right\}.$$

Overall Algorithm

- To initialize NMF, the data are clustered with k-means, where k is the number of metagenes.

Algorithm 1 NMF-HMRF model-fitting and hidden state estimation.

Derive an initial estimate M^0 using K-means clustering assuming no spatial relationships.
for $1 < t \leq T_0$ **do**
    Derive an estimate X^t by minimizing $R(X) = ||Y - M^{t-1}X||_2^2$.
    Derive an estimate M^t by minimizing $R(M) = ||Y - MX^t||_2^2$.
end for
Set $M^{(0)} = M^{T_0}$, $X^{(0)} = X^{T_0}$.
Derive an initial estimate $\sigma_y^{(0)} = \sqrt{R(X^{(0)})/(G \times N)}$.
Initialize $(\Sigma_x^{-1})^{(0)}$ to a zero matrix.
for $1 < t \leq T$ **do**
    Derive an estimate $X^{(t)}$ given $\Theta^{(t-1)}$ by maximizing $P(X|Y, \Theta = \Theta^{(t-1)})$.
    Derive an estimate $\Theta^{(t)}$ given $X^{(t)}$ by maximizing $P(\Theta|Y, X = X^{(t)})$.
end for

SpiceMix vs NMF

The latent representation was used to assign cells to cell-types through hierarchical clustering. SpiceMix notably identified the VIP and SSI subtypes of inhibitory neurons, while these were grouped into a single subtype under regular NMF.

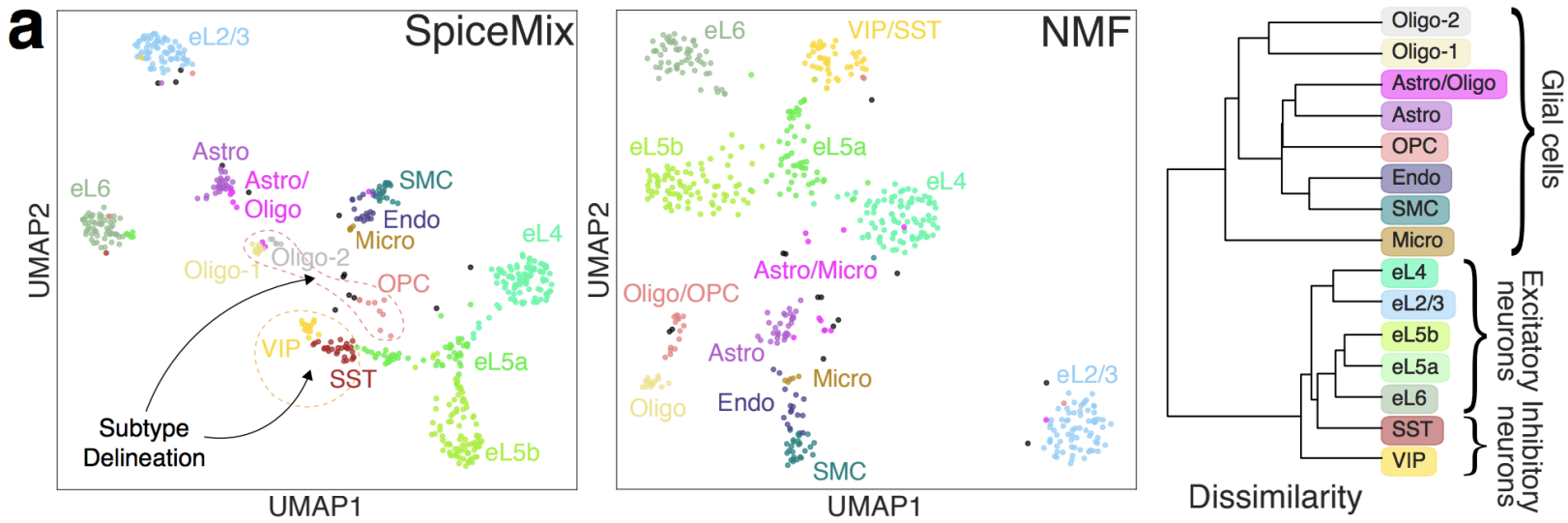


Figure: from Fig. 3 https://www.biorxiv.org/content/biorxiv/early/2020/11/30/2020.11.29.383067.full.pdf

Conclusion

- Incorporating spatial information is important for cell phenotyping is important.
- Where cells are located is important biologically.
- We are interested in co-occurence of particular cell-types
- bf There are hardly any solutions for these problems yet!