Comp790-166: Computational Biology

Lecture 10

February 14, 2023

Today

- MELD for differential abundance (DA) analysis
- volume-dependent DA testing

Project Proposal Template

How are your projects going? I put a template for your proposals online, which can be found here, `https://github.com/natalies-teaching/CompBio2023/blob/main/Project_Proposal/Project_Proposal.tex`.

Discussion about Project Proposals

- **Abstract:** Sell your idea in 3-5 sentences. This is really good practice for figuring out what the story is with a project. Convince us why we should care.
- **Formal Problem Statement:** This should be a 1 to 2 sentence summary of what your problem is. Easier said than done.....
- **Contributions:** Think about a list of contributions and then put this into formal writing.
- **Intended Experiments:** Realistically you can aim for 1 to 2 experiments (more if you want!)
- **Implementation:** What is the product that you will give to the scientific community? (e.g. well-documented open source software)

Do you remember questions

- Which eigenvalues correspond to high and low frequency components?
- What does a low pass filter do?

Treatment as a Signal on a Graph

After creating a graph of cells, an indicator of treatment or control can be viewed as the signal on the graph. Interpretations of ‘signal‘ in relation to graph structure should help to inform treatment associated relative likelihood.

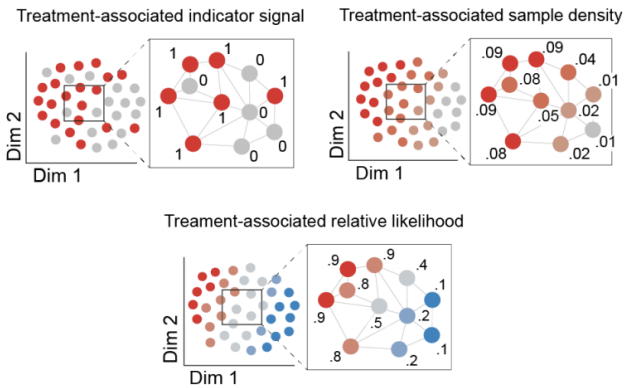

Figure: Burkhardt *et al.*, Nature Biotechnology. 2021.

General Overview of the Steps of MELD

- Build a graph between cells based on gene or protein expression measurements
- **Graph Signals:** Experimental label (a binary indicator) is used to label each cell according to experimental condition
- Using GSP techniques, MELD filters biological and technical noise to look at how much the experimental signal of a cell matches the true experimental label. This quantifies how prototypical each cell is in its condition.
- Relate back to cell-types and features that differ between experimental conditions

RES vs EES

EES represents the enhanced experimental signal, in comparison to RES, which was the raw, binary signal.

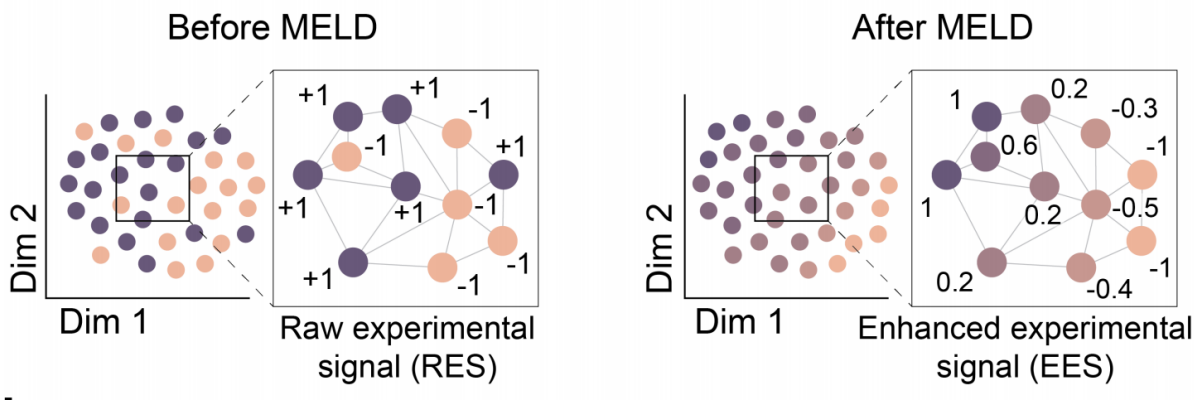

Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021

Sources of Noise

- Cells with similar feature measurements are said to be in the same state (biologically)
- **High Frequency Noise** : High frequency noise is when the labels of neighboring cells are rapidly fluctuating.
- **Graph Fourier Transform** is used to study the frequency of a signal over an irregular domain, like a graph.

Incorporating these ideas into meld

- Define a latent variable **z** that gives a score for how prototypical a cell is for a specific experimental or clinical condition.
- **z** will be computed using low-pass graph filters.
- Defining more specific variables
  - **x** is the vector of original labels (RES) for each cell
  - **z** is the vector of enhanced experimental signals (EES) for each cell.

Visualizing **x** and **z**

The left is RES (**x**) and the right is EES (**z**). **z** is what is being optimized.

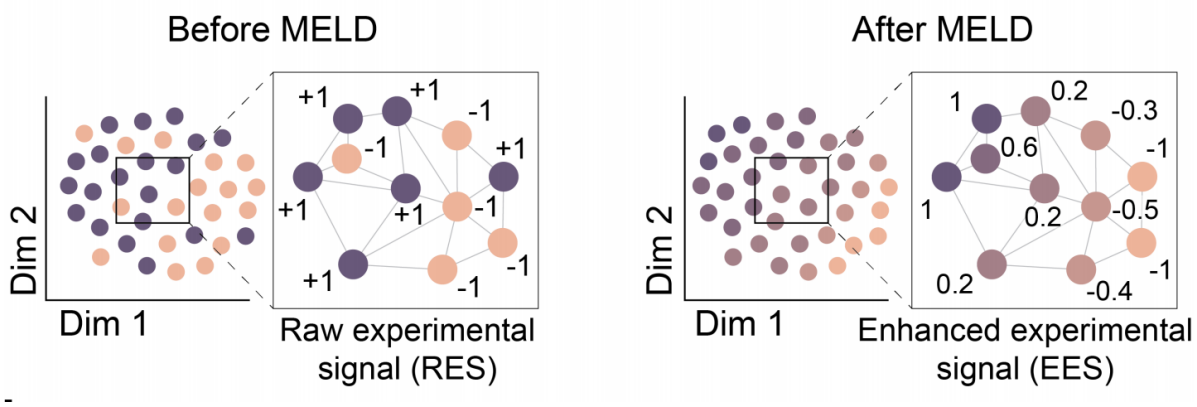

Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021

MELD Optimization Problem

To find an appropriate $\mathbf{z}$, an optimization problem can be defined as,

$$\mathbf{y} = \underset{\mathbf{z}}{\operatorname{argmin}} \underbrace{\|\mathbf{x} - \mathbf{z}\|_2^2}_{\mathbf{a}} + \underbrace{\beta \mathbf{z}^T \mathcal{L} \mathbf{z}}_{\mathbf{b}} \tag{1}$$

Unpacking

z is the EES or Enhanced Experimental Signal

$$\mathbf{y} = \underset{\mathbf{z}}{\operatorname{argmin}} \underbrace{\|\mathbf{x} - \mathbf{z}\|_2^2}_{\mathbf{a}} + \underbrace{\beta \mathbf{z}^T \mathcal{L} \mathbf{z}}_{\mathbf{b}} \tag{2}$$

- The Laplacian Regularization (term b) initially encourages smoothness for an input graph signal, **x**
- (**a**) Term a represents reconstruction between **x** and **z**
- (**b**) Term b represents Laplacian regularization or a measure of smoothness on the graph. Recall this looks a lot like total variation.

$$\beta \mathbf{z}^T \mathcal{L} \mathbf{z} = \beta \sum_{i,j} A_{ij} (\mathbf{z}(i) - \mathbf{z}(j))^2 \tag{3}$$

Introducing the MELD Filter

They adjust the filer a bit as follows. The following allows also for a flexible notion of figure order, ρ,

$$\begin{aligned} \mathbf{y} &= \underset{z}{\text{argmin}} \|\mathbf{x} - \mathbf{z}\|_2^2 + \mathbf{z}^T \mathcal{L}_* \mathbf{z} \\ &\text{where } \mathcal{L}_* = [\beta \mathcal{L} - \alpha \mathbf{I}]^\rho \end{aligned} \tag{4}$$

Takeaway

They show that their Laplacian Regularization is a filter with the following frequency response,

$$h_{\mathrm{MELD}}(\lambda) = \frac{1}{1 + (\beta\lambda - \alpha)^{\rho}} \tag{5}$$

This was a lot to unpack. I recommend staring at the details (if you are interested) in
`https://www.biorxiv.org/content/10.1101/532846v1.full.pdf`

Filter Variety

Here are some experiments showing what parameters on the MELD filter will do to the frequency response, $h(\lambda)$.

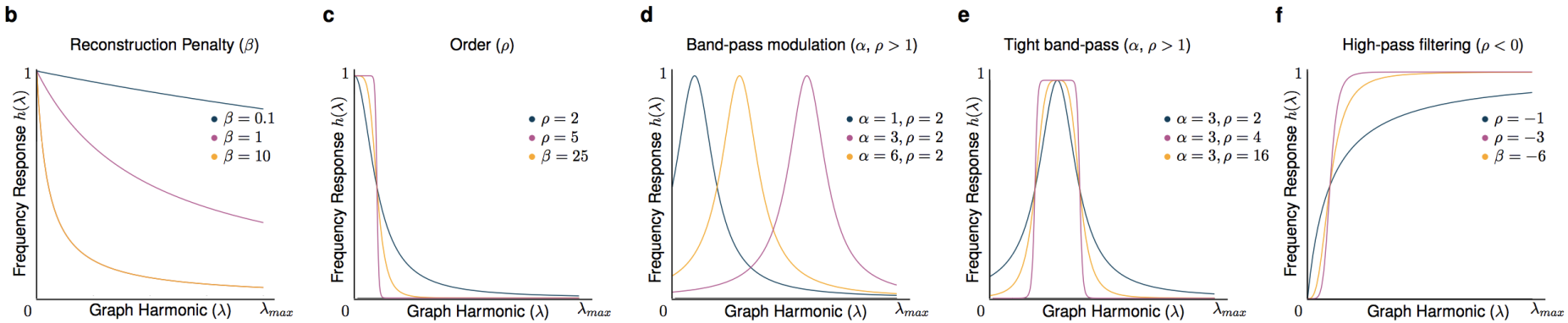

Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021. Negative values of ρ, for example, can produce a high-pass filter.

Meld Results

Computing the EES cleans up some of the noise and helps to better identify prototypical cells in each experimental condition.

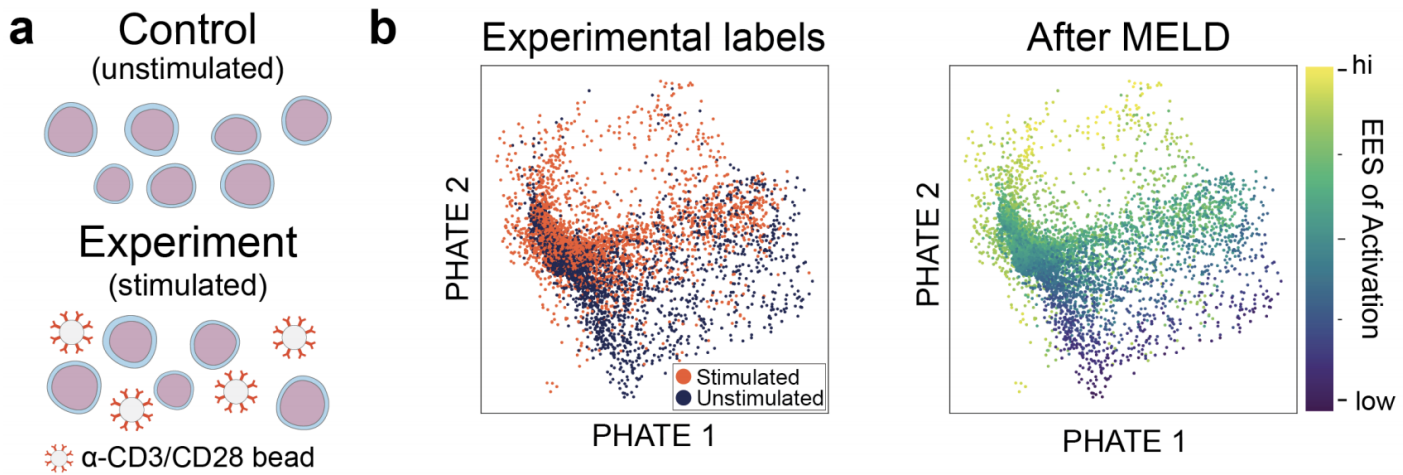

Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

Gene Expression Profiles Based on RES and EES

You can look at the gene expression profiles of cells with similar EES scores.

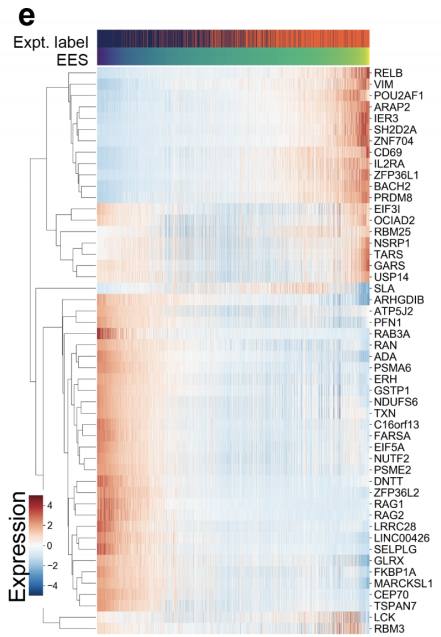

Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

Zooming in on High and Low Frequency Regions

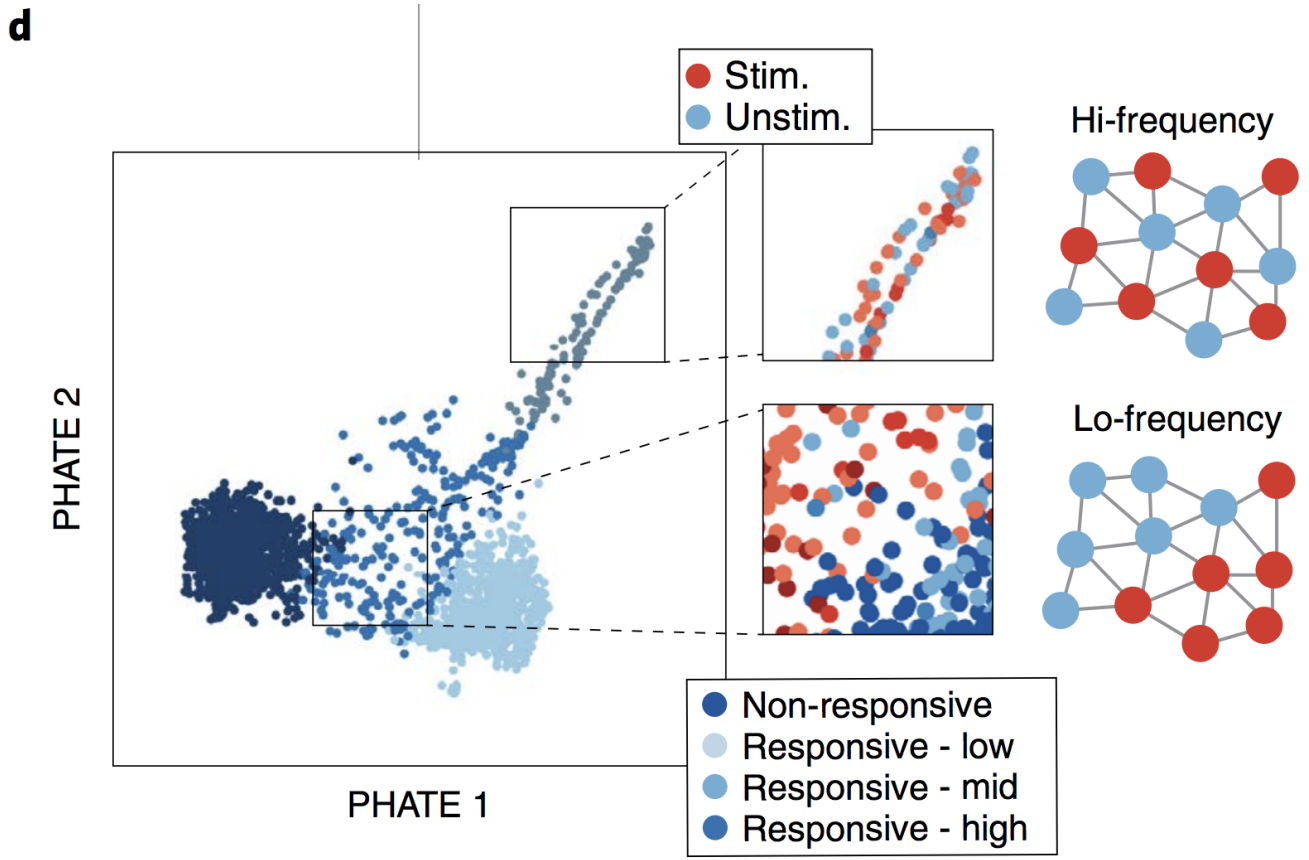

Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

A General Question: What's Different Between Clinical Groups

In this example, the abundances of particular cell-types are being compared between patients who have varying mortality likelihoods from COVID.

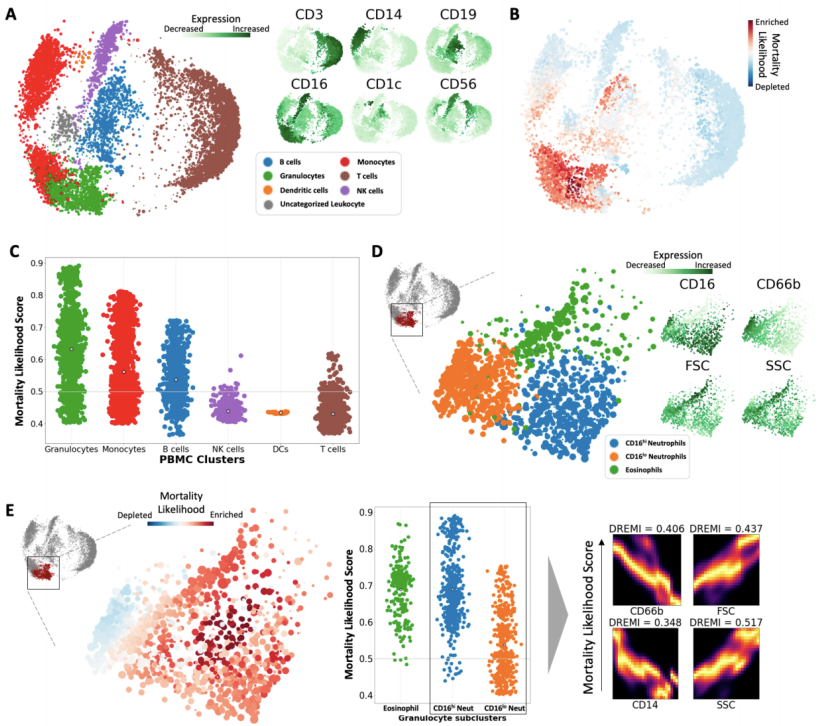

Figure: from Kuchroo *et al.* 2020. `https://www.biorxiv.org/content/10.1101/2020.11.15.383661v1.abstract`

Problem Formally Stated

Given two patient phenotypes, which cell-populations are statistically, significantly different between groups in terms of **frequency**,**function**, or 'state'?

- We are exploring this question in a statistical way, rather than through building a classifier or a model. Therefore, we need to look out for multiple testing problems!
- In contrast to Meld, we are no longer looking for prototypical cell examples associated with each condition. Instead, we are testing overlapping subsets of cells for significance.

Welcome Cydar (Nature Methods 2017).

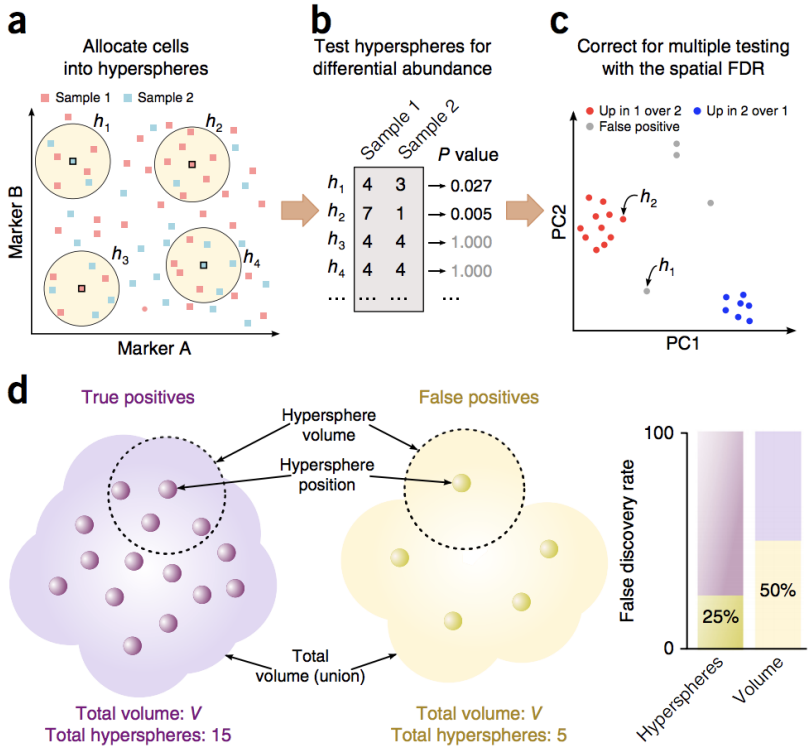

Figure: from Lun *et al.* Nature Methods. 2017.

Clustering vs Other Approaches

- The authors claim they can just cluster cells and then study differential abundance (DA) in each cluster
- The rightfully suggest that the quality of downstream interpretation can be affected by noisy cells, clustering parameters, etc.
- To deal with this, they develop a ‘hypersphere‘ based approach

Assigning Cells to Hyperspheres

Hyperspheres are designed to be centered around existing individual cells. In particular, around every $10th$ cell.

- Define the radius of each hypersphere, r, as $r = 0.5\sqrt{M}$. M is the number of markers or parameters measured for each cell
- Any cell is assigned to the hypersphere if their distance to the hypersphere center is within $r = 0.5\sqrt{M}$
- Note that this definition imposes overlap between hyperspheres, or having cells assigned to more than one hypersphere.

Hyperspheres Illustrated

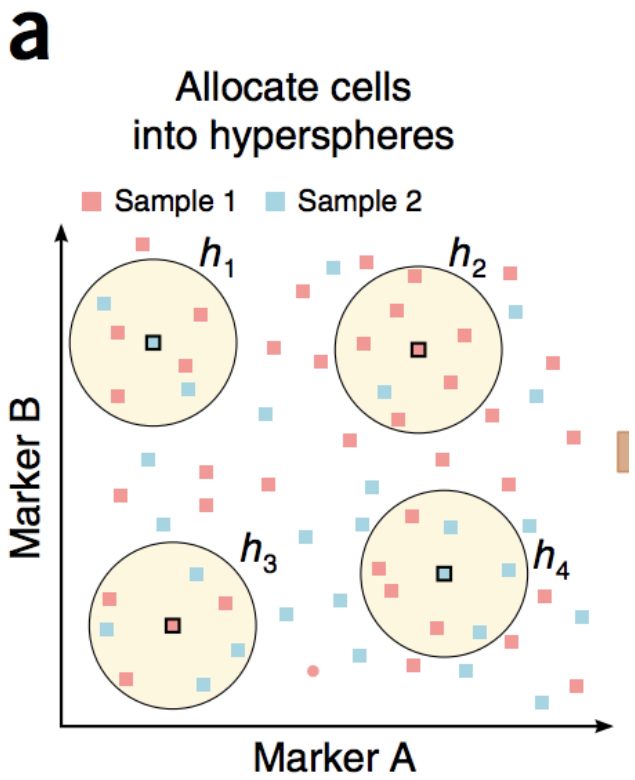

Figure: from Lun *et al.* Nature Methods. 2017.

Testing Differences Between Groups

- First, for each hypersphere, and for each sample, you are going to calculate the proportion of sample's cells assigned to the hypersphere. Again, cells will be counted twice because they can belong to multiple hyperspheres.
- Then you can apply some statistical test (like Wilcoxon Rank Sum Test) to test if the proportion of cells assigned to each hypersphere are the same
- The Null hypothesis is that the mean proportion of cells belonging to each hypersphere should be the same between groups (e.g. treatment and control)

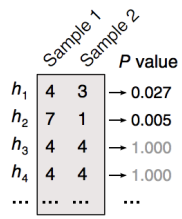

Figure: from Lun *et al.* Nature Methods. 2017.

Example Null Hypothesis for tReg Hypersphere

H_0 :The number of tRegs is the same between healthy and sick people
H_1 : The number of tRegs between healthy and sick people is not the same
The caveat here is were are doing many, many of such tests so significance can arise by chance!

Intuition Behind Spatial FDR Idea

The spatial FDR can be interpreted as the proportion of the total volume (rather than the sum of individual hypersphere volumes) that is occupied by false positively differentially abundant hyperspheres.

- Hypersphere density differs across the high-dimensional space. So, we will soon see that each hypersphere is weighted by the recipricol of its density of neighboring hyperspheres.

Spatial FDR

False discoveries are when the null hypothesis (that the abundance is the same between groups) is *falsely* rejected. Cydar computes a spatial FDR, which considers the proportion of the total volume of differentially abundant hyperspheres that are occupied by false-positive hyperspheres.

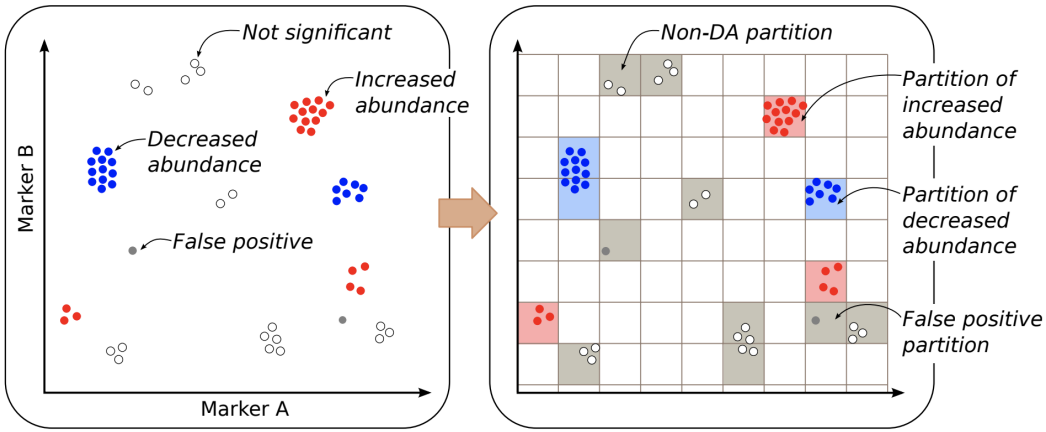

Figure: from Lun *et al.* Nature Methods. 2017.

Spatial FDR, Continued

Each circle is representing a hypersphere colored by increase in abundance (red), decrease (blue), no change (white), or false positive (gray). On the right shows a partition of the space, which is ultimately labeled depending on the hyperspheres comprising it. Partition can be labeled as false positive.

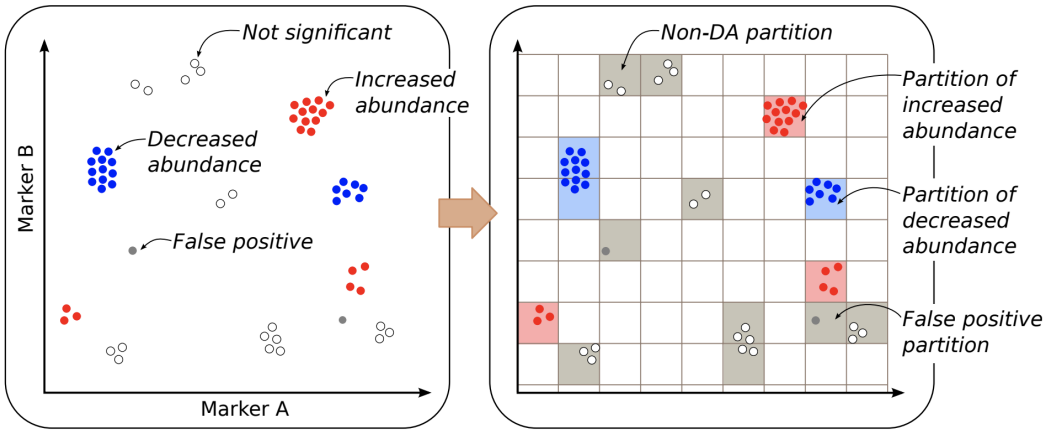

Figure: from Lun *et al.* Nature Methods. 2017.

Weighted Benjamini-Hochberg

- Assume that for n hyperspheres that there are n ordered p-values from the statistical test with $p_1 \leq p_2 \leq \ldots p_n$.
- Imagine a partition of your M dimensional space. For a particular hypersphere, its local density will represent how representative it is of that partition.
- For hypersphere, I, define w_I as the weight, which is inversely related to the density of hypersphere I. Local density is defined by the distance of each hypersphere to its 50th nearest neighbor.

Weighted BH, Coontinued

Assuming ordered pvalues $p_1 < p_2 < ... < p_n$, then the weighted BH method will reject any null hypothesis where the p-value is less than the following. Here α is some threshold at which you would like to control your FDR.

$$\max_i \left\{ p_{(i)} : p_{(i)} \leq \alpha \frac{\sum_{l=1}^{i} w_{(l)}}{\sum_{l=1}^{n} w_{(l)}} \right\}$$

Cydar Applied in Practice

They applied Cydar to an MEF dataset (mouse embryonic fibroblast). Samples were collected at 13 timepoints between day 0 and day 20. The goal was to detect subppulations that change over time.

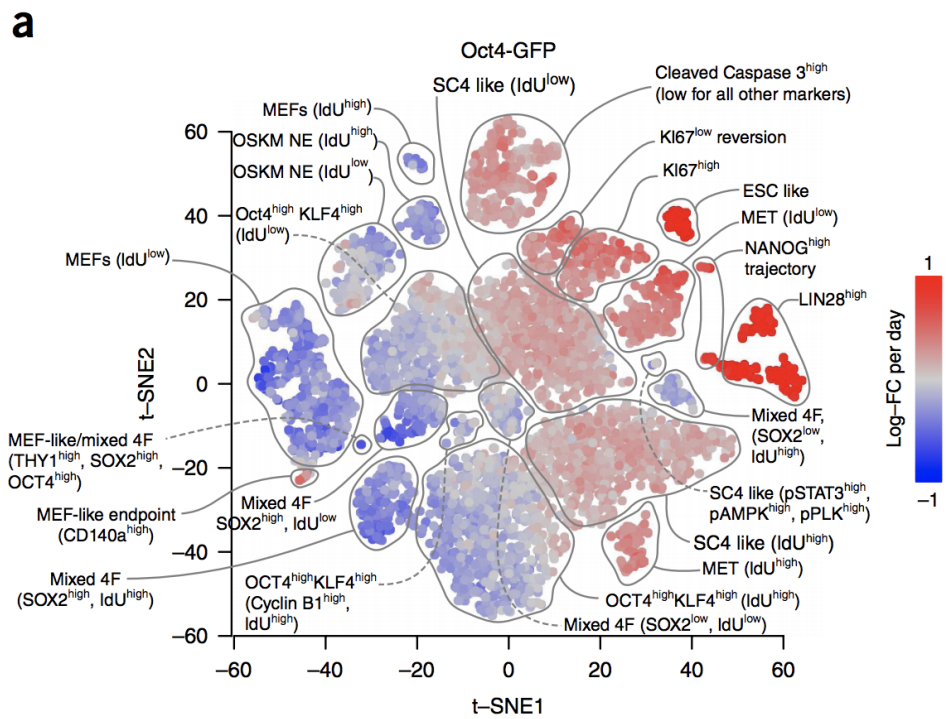

Figure: from Lun *et al.* Nature Methods. 2017. Each plotted point represents the median position of differentially abundant hyperspheres at an FDR of 5%

A Practical Point about Annotation

Getting a colored tSNE like this is just the beginning. You then need to do the following to describe your cell-populations, or to annotate them by hand.

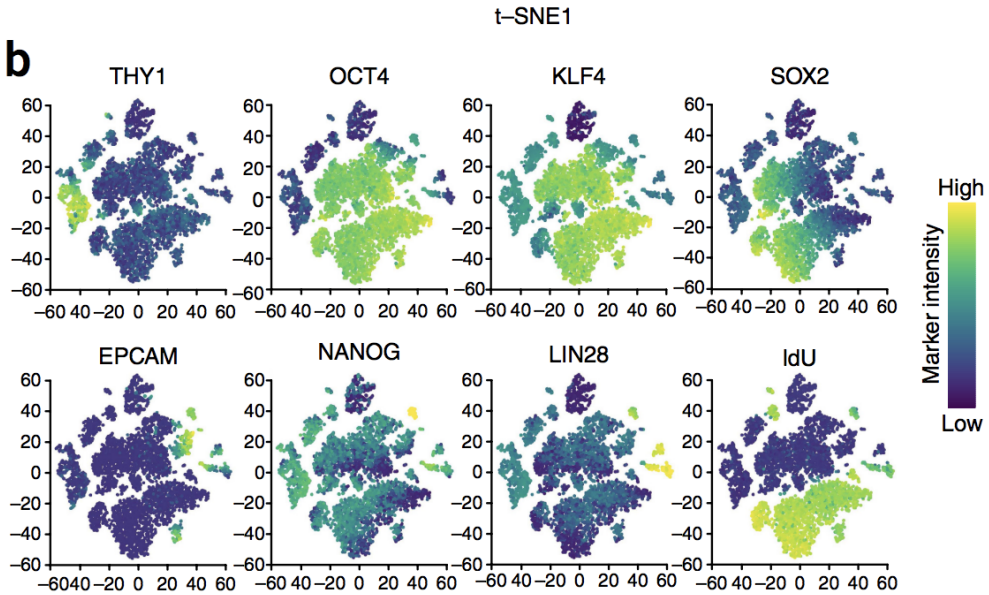

Figure: from Lun *et al.* Nature Methods. 2017. Color cells by the expression of individual markers.

There has been some work automating this.

Welcome GateFinder (Aghaeepour *et al.* Bioinformatics. 2018). The goal of GateFinder is to tell you the combinations of markers that characterize your cell-population of interest.

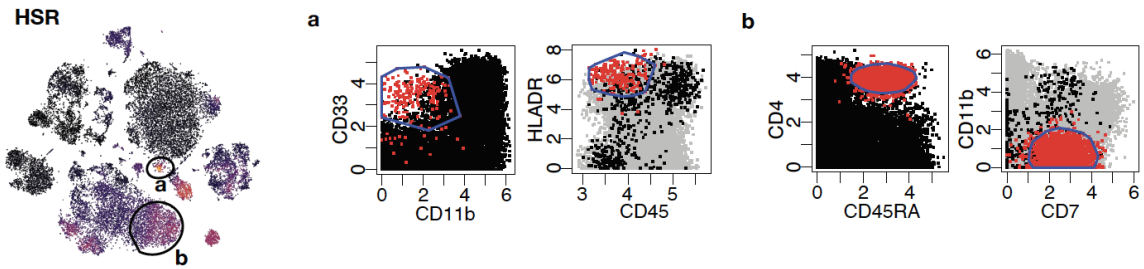

Figure: Nature Communications. 2020. We applied GateFinder to figure out what types of cells our prioritized cells were based on the combinations of markers that were expressed.

Thoughts...

- Why not just define clusters, test clusters, and do something simple like dividing the p-value threshold for significance by the number of tests? (aka Bonferonni)
- My guess it that that is completely driven by visualization. They want a way to visualize individual cells, not cluster centers.
- All of this hypersphere business seems a bit expensive and time consuming if you could just do k-means $\rightarrow$ test $\rightarrow$ correct

Example of Cluster-Based Testing

We have done something simpler. We calculated a score for each cell based on a linear combination of similarity (in marker space) to each cluster and that cluster's score.

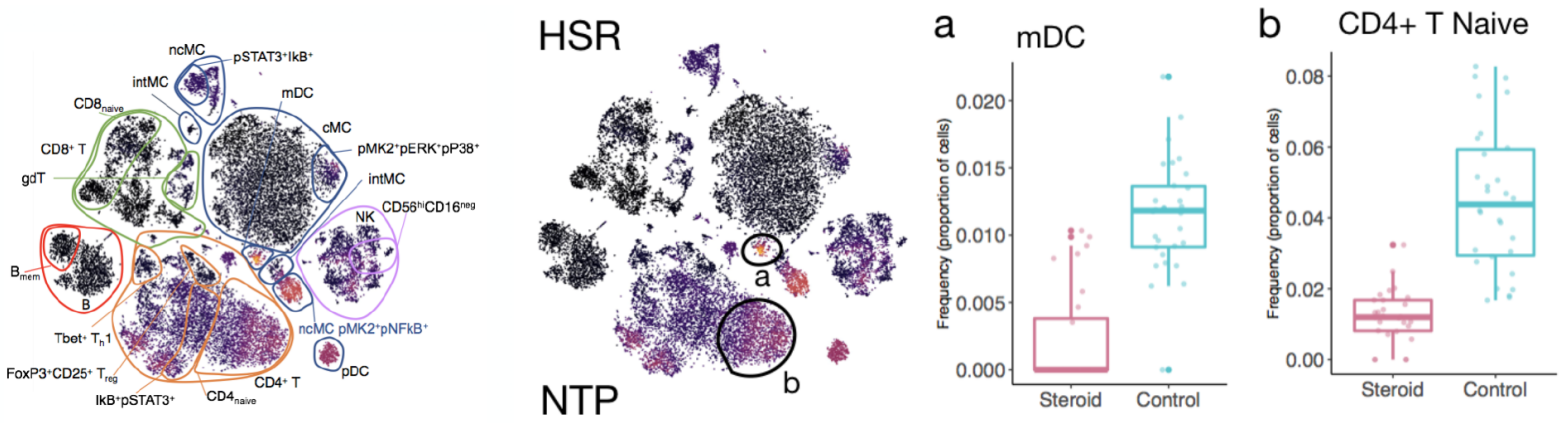

Figure: from Stanley *et al.* Nature Communications. 2020.