Comp790-166: Computational Biology

Lecture 12

February 18, 2022

Good Morning Question

We spent the last two days introducing the basic of GSP and filtering. Here are a couple of questions as a warm-up.

1. Explain the concepts of connectivity (e.g. edges) existing between cells that comprise a graph. Similarly, what is an example signal that we could have on these nodes (cells)?
2. Do low-pass filters prioritize the low-frequency components (e.g. small eigenvalues) or the high-frequency components (e.g. larger eigenvalues)?

Today

- Almost done with the GSP and filtering discussion.
- Identifying condition-specific prototypical cells with MELD
- Start differential abundance discussion with Cydar

Intermission for Announcements

- Homework 1 is due by 11:59pm eastern time on February 23
- Reminder : Monday office hours after class until noon.
- Project proposal template is now online `https://github.com/natalies-teaching/Comp790-166-CompBio-Spring2022/blob/main/Project_Proposal/Project_Proposal.pdf`
- Sign up a time to present your project on March 7 or March 9, `https://docs.google.com/spreadsheets/d/1fX52jKWDWbJO1iB6D7FHv1DNQ53LPs0S8yNCUnUbSwQ/edit?usp=sharing`

Reminder 1 : Local Variation Leads to Total Variation (Signal Smoothness)

The total variation of a signal on a graph is defined as follows and is also known as the Laplacian Quadratic Form

$$TV(\mathbf{f}) = \sum_{i,j} A_{ij}(\mathbf{f}(i) - \mathbf{f}(j))^2 \tag{1}$$

$$= \mathbf{f}^T \mathcal{L} \mathbf{f} \tag{2}$$

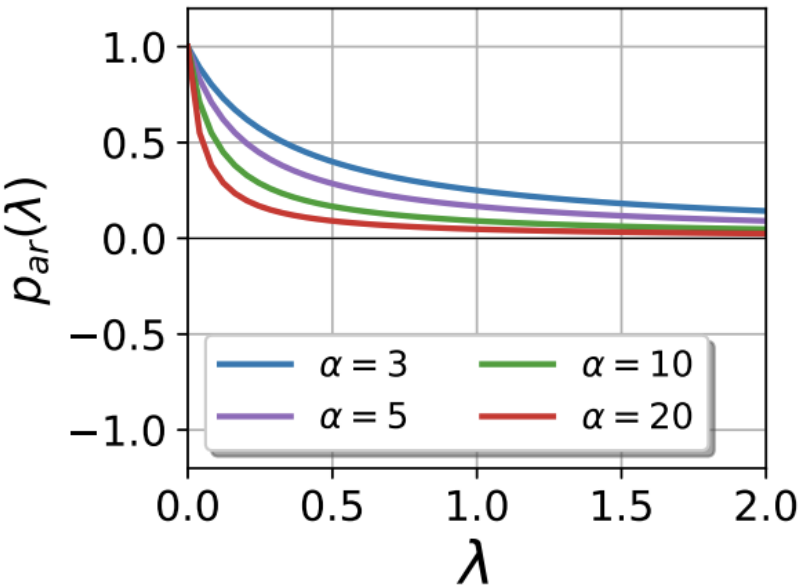
- Note here I have been assuming that we have an unweighted graph, but you could certainly substitute A_{ij} with a weighted version, W_{ij}

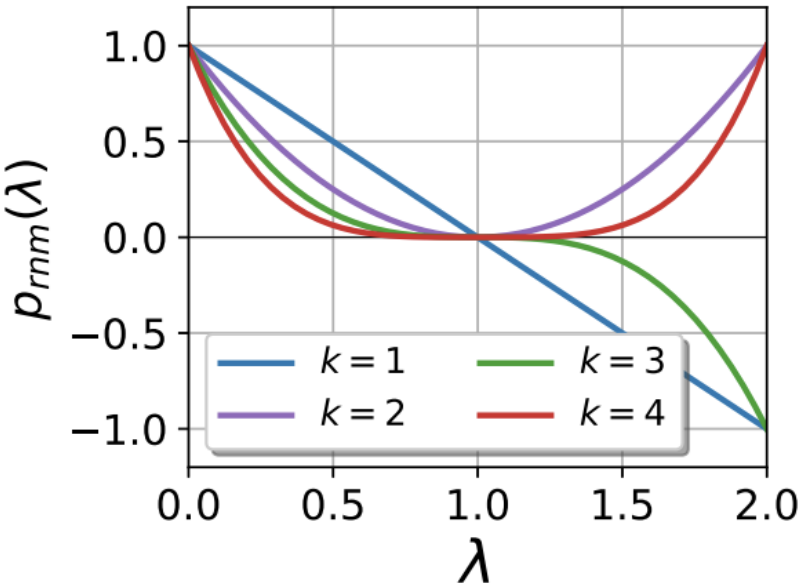
Applying a Filter in General

In general, you can filter an original signal, $\mathbf{f}$ in general as,

$$\underbrace{\boldsymbol{\Psi}(\mathbf{I} + \alpha\boldsymbol{\Lambda})^{-1}\boldsymbol{\Psi}^T}_{\text{Filtered Graph Laplacian}} \mathbf{f}. \tag{3}$$

Example Filters





(a) $p_{\text{ar}}(\lambda) = (1 + \alpha\lambda)^{-1}$ (b) $p_{\text{rnm}}(\lambda) = (1 - \lambda)^{k}$

Figure: from `https://openaccess.thecvf.com/content_CVPR_2019/papers/Li_Label_Efficient_Semi-Supervised_Learning_via_Graph_Filtering_CVPR_2019_paper.pdf`

Example in PyGSP

- Access PyGSP here, `https://pygsp.readthedocs.io/en/stable/tutorials/intro.html`

$$g(x) = \frac{1}{1+\tau x},$$

let's define and plot that low-pass filter:

```
>>> tau = 1
>>> def g(x):
...     return 1. / (1. + tau * x)
>>> g = filters.Filter(G, g)
>>>
>>> fig, ax = plt.subplots()
>>> g.plot(plot_eigenvalues=True, ax=ax)
>>> _ = ax.set_title('Filter frequency response')
```

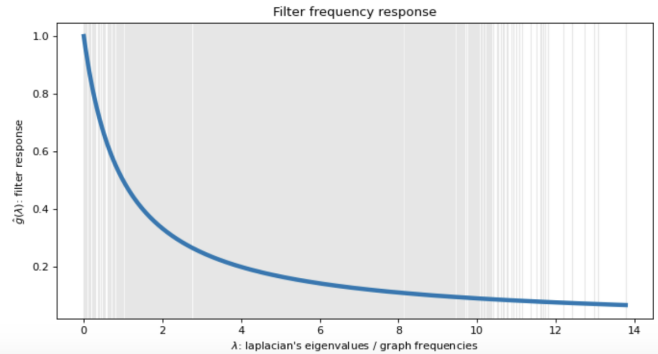
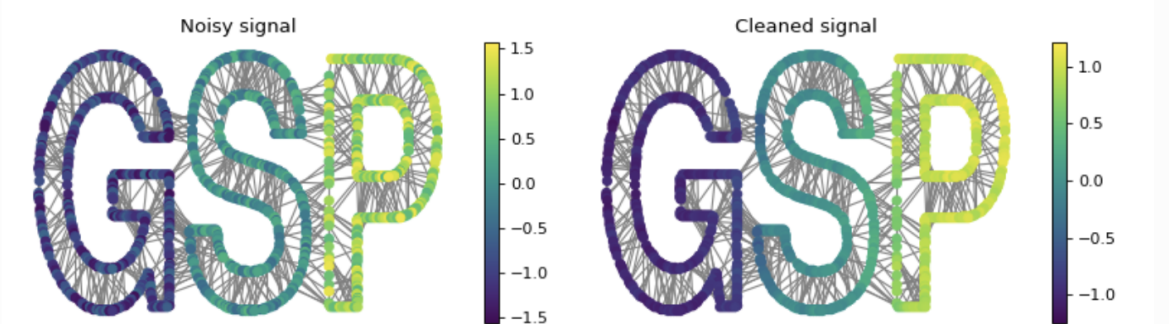


Figure: A simple filter for eigenvalues of **L**

Low-Pass Filtering a Noisy Signal

```
>>> s2 = g.filter(s)
>>>
>>> fig, axes = plt.subplots(1, 2, figsize=(10, 3))
>>> G.plot_signal(s, vertex_size=30, ax=axes[0])
>>> _ = axes[0].set_title('Noisy signal')
>>> axes[0].set_axis_off()
>>> G.plot_signal(s2, vertex_size=30, ax=axes[1])
>>> _ = axes[1].set_title('Cleaned signal')
>>> axes[1].set_axis_off()
>>> fig.tight_layout()
```



Linking Single Cell Data to External Information

MELD (an application of GSP!)

The idea of MELD is to model how experimental perturbation alters cell states.

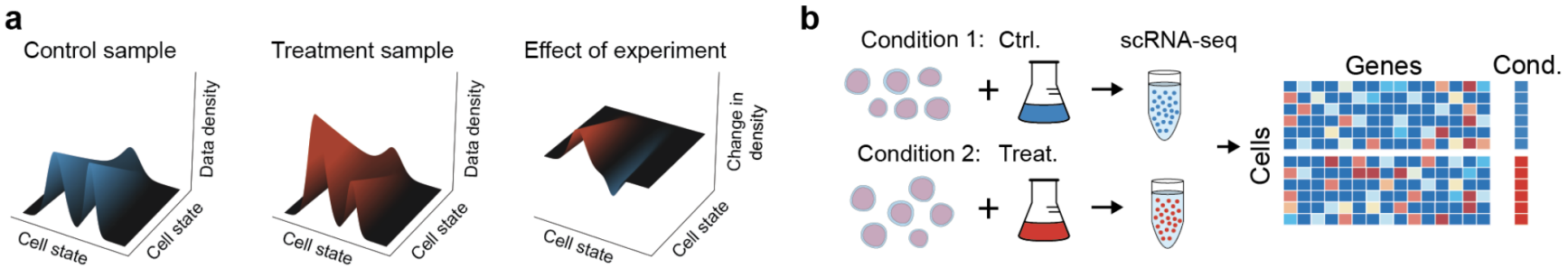


Figure: Burkhardt *et al.*, Nature Biotechnology. 2021. How more or less likely are we to observe a cell in a particular state in a control vs treatment sample?

A Question for You

If you were just given a bunch of cells and someone told you to find 200 cells that were associated with an experiment or a condition of interest, how would you choose those cells? What would cause you to trust that a particular cell was indeed representative of the experimental label?

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.

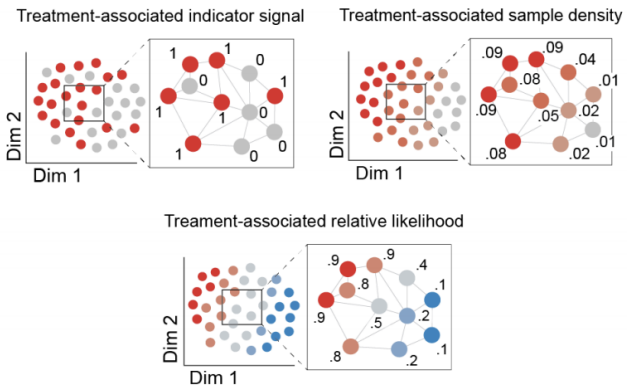


Figure: Burkhardt *et al.*, Nature Biotechnology. 2021.

What on Earth is a Cell-State

We have seen cell states already!

- Frequency $\rightarrow$ how many of a particular cell-type are there in a sample?
- Function $\rightarrow$ Which proteins are activated in a particular cell-type?

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.

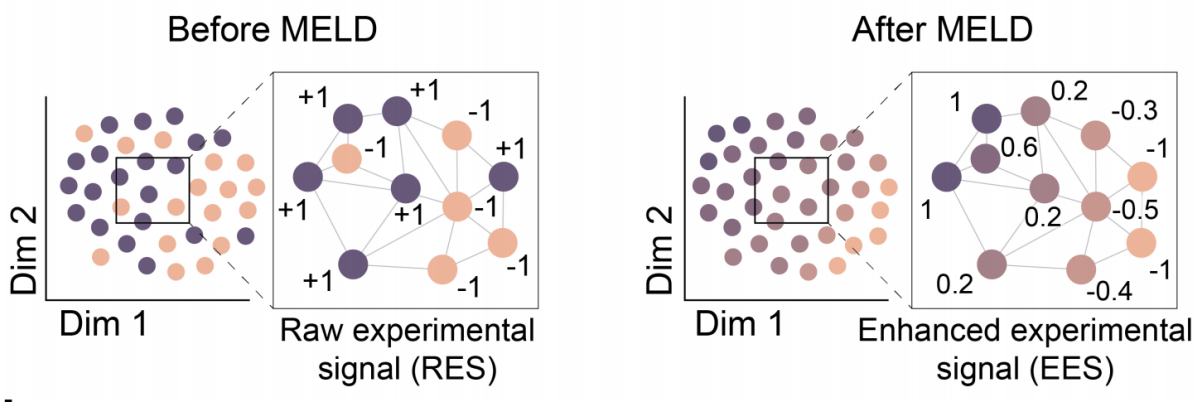


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021

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 $\mathbf{z}$ that gives a score for how prototypical a cell is for a specific experimental or clinical condition.
- $\mathbf{z}$ will be computed using low-pass graph filters.
- Defining more specific variables
  - $\mathbf{x}$ is the vector of original labels (RES) for each cell
  - $\mathbf{z}$ 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.

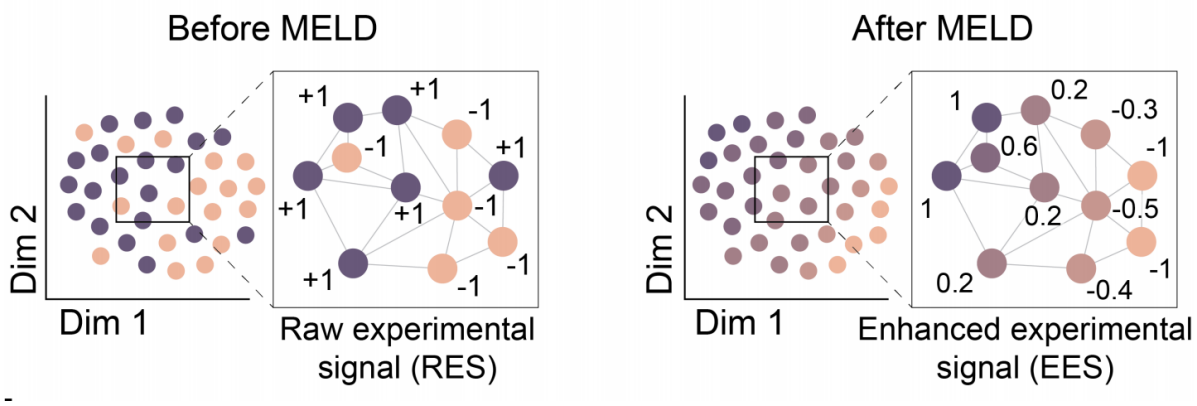


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021

MELD Optimization Problem

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

$$\mathbf{y} = \underset{\mathbf{z}}{\operatorname{argmin}} \underbrace{\|\mathbf{x} - \mathbf{z}\|_2^2}_{\mathbf{a}} + \underbrace{\beta \mathbf{z}^T \mathcal{L} \mathbf{z}}_{\mathbf{b}} \tag{4}$$

Unpacking

z is the EES or Enhanced Experimental Signal

$$\mathbf{y} = \underset{\mathbf{z}}{\operatorname{argmin}} \underbrace{\|\mathbf{x} - \mathbf{z}\|_2^2}_{\mathbf{a}} + \underbrace{\beta \mathbf{z}^T \mathcal{L} \mathbf{z}}_{\mathbf{b}} \tag{5}$$

- 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.

$$\beta \mathbf{z}^T \mathcal{L} \mathbf{z} = \beta \sum_{i,j} A_{ij} (\mathbf{z}(i) - \mathbf{z}(j))^2 \tag{6}$$

Introducing the MELD Filter

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

$$\begin{aligned} \mathbf{y} &= \underset{z}{\text{argmin}} \|\mathbf{x} - \mathbf{z}\|_2^2 + \mathbf{z}^T \mathcal{L}_* \mathbf{z} \\ &\text{where } \mathcal{L}_* = [\beta \mathcal{L} - \alpha \mathbf{I}]^\rho \end{aligned} \tag{7}$$

Takeaway

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

$$h_{\mathrm{MELD}}(\lambda) = \frac{1}{1 + (\beta\lambda - \alpha)^{\rho}} \tag{8}$$

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, $h(\lambda)$.

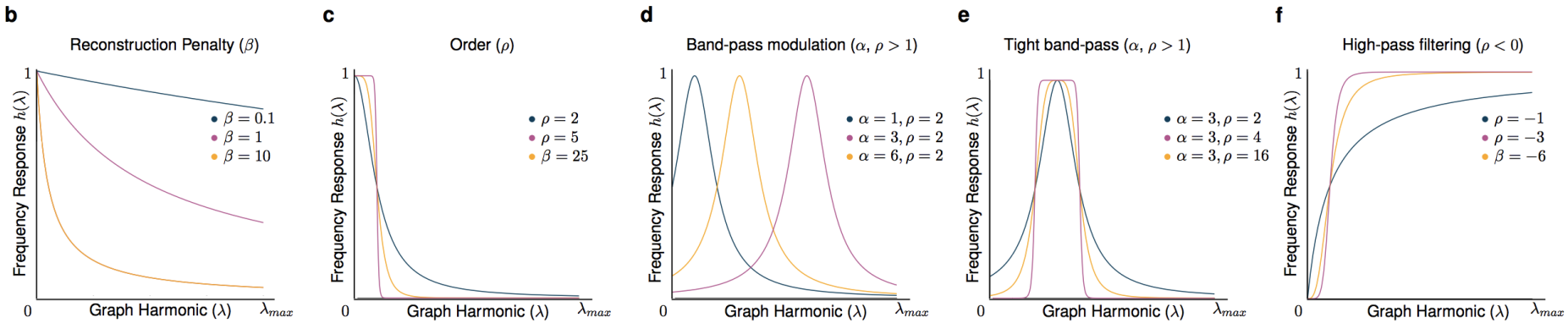


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021. Negative values of ρ, for example, can produce a high-pass filter.

Meld Results

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

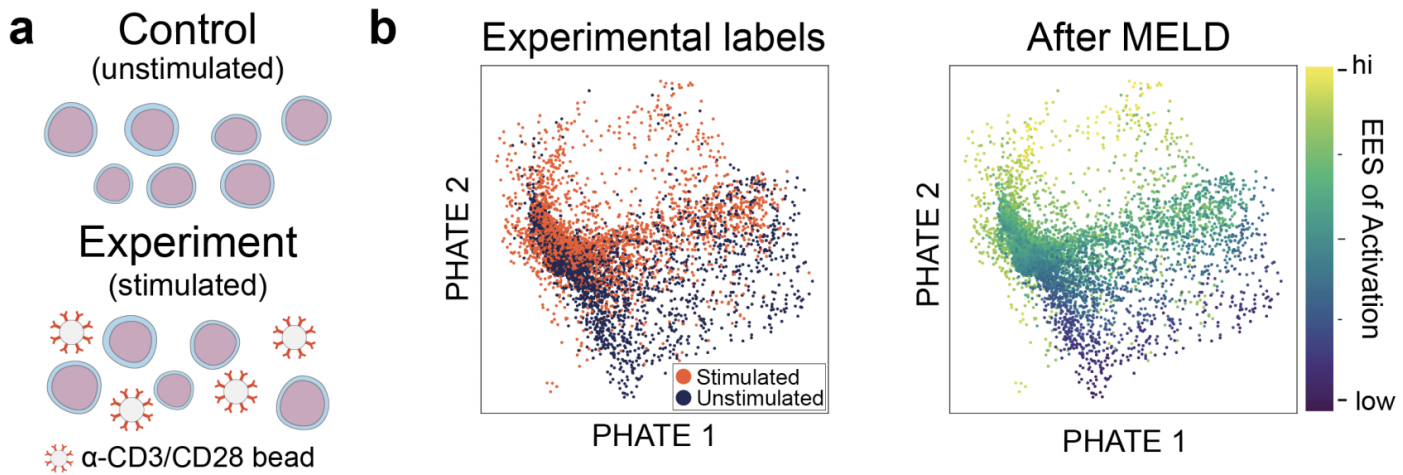


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

Gene Expression Profiles Based on RES and EES

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

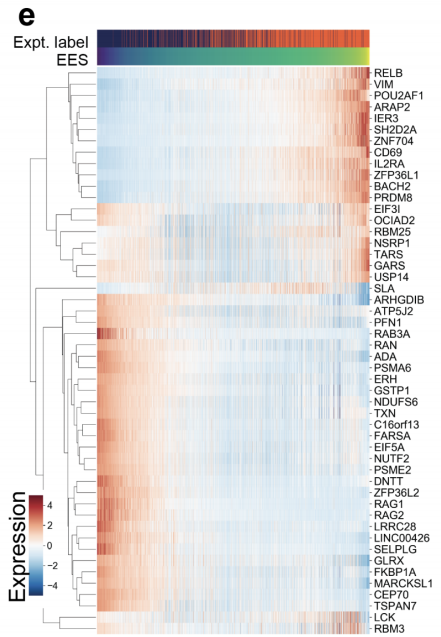


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

Zooming in on High and Low Frequency Regions

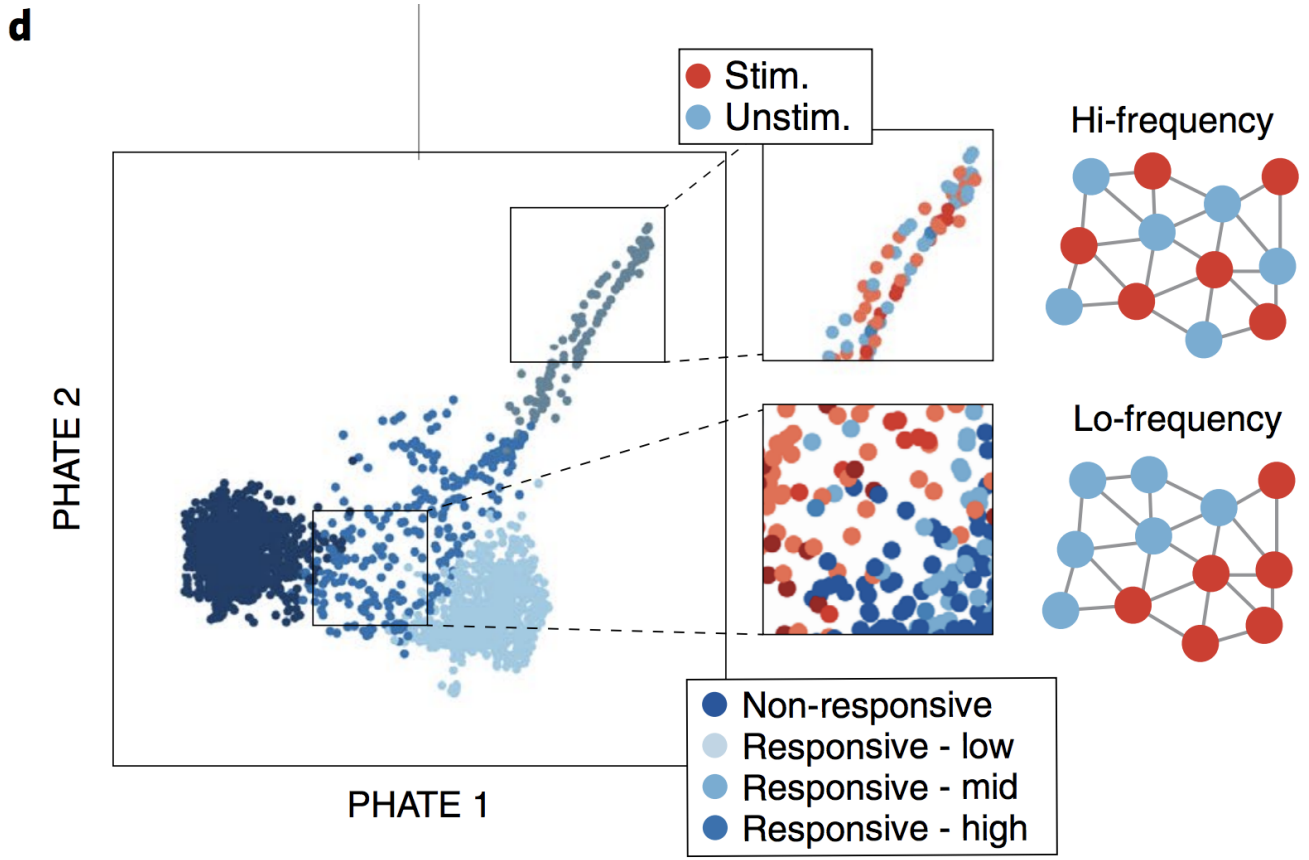


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

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.

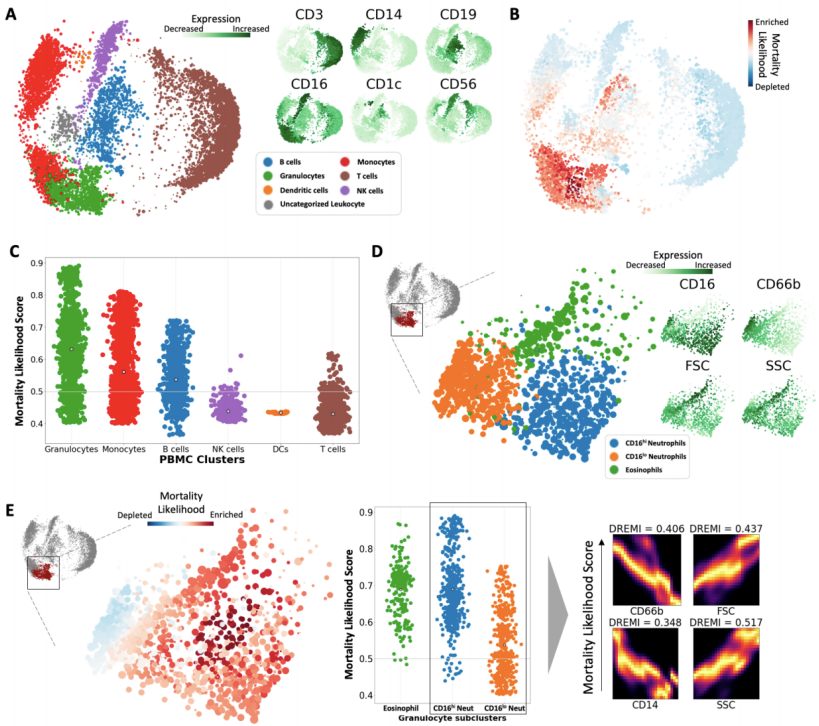


Figure: from Kuchroo *et al.* 2020. `https://www.biorxiv.org/content/10.1101/2020.11.15.383661v1.abstract`

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).

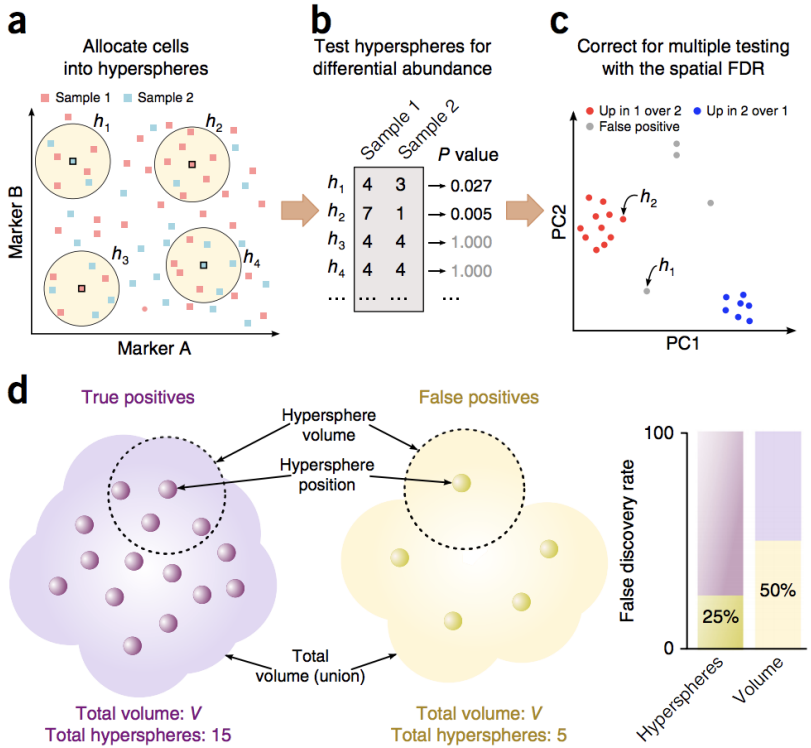


Figure: from Lun *et al.* Nature Methods. 2017.