# Comp683: Computational Biology

Lecture 13

February 26, 2024

# Today

- Finish differential abundance analysis with Cydar
- Milo as a graph-based method for differential abundance.
- Contrastive PCA for dealing with background data.

Announcements

- Homework is due today by 11:59pm
- Information on project proposals : `https://github.com/natalies-teaching/Comp683_CompBio_2024/tree/main/Project_Proposal`

# Good Morning Question

1. Who can tell us about how to define the hyperspheres with Cydar? What is a potential benefit of using the hypersphere approach rather than just clustering cells into populations from the beginning?
2. What is the problem with testing multiple cell-populations between two groups?

# Hyperspheres Illustrated in the Cydar Algorithm

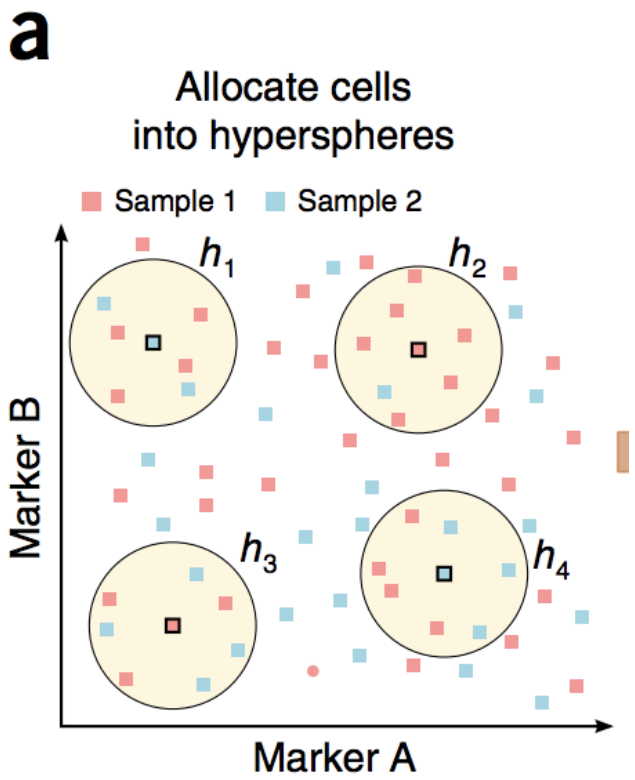



Figure: from Lun *et al.* Nature Methods. 2017.

# 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.

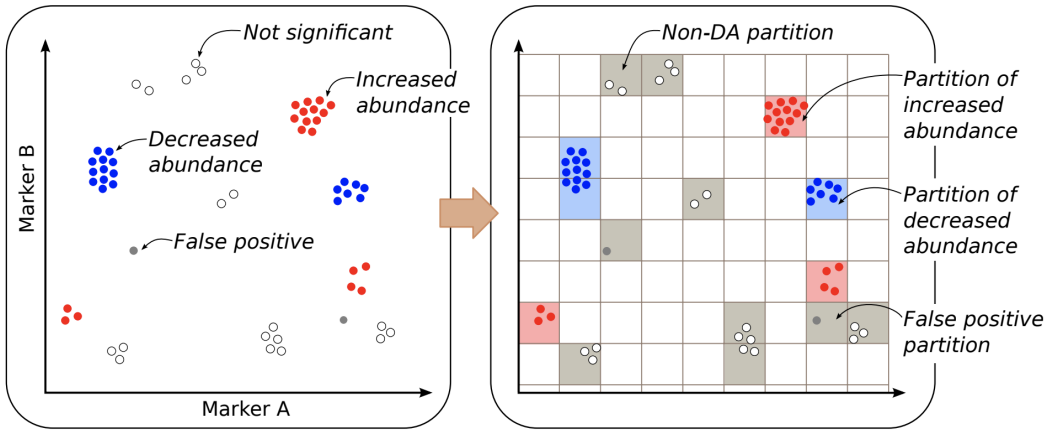



Figure: from Lun *et al.* Nature Methods. 2017.

# 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.

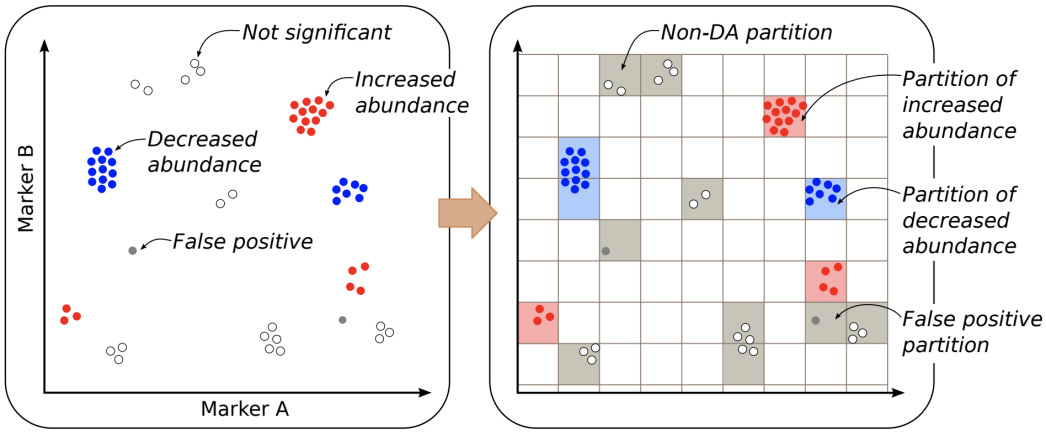



Figure: from Lun *et al.* Nature Methods. 2017.

# 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, $l$, define $w_l$ as the weight, which is inversely related to the density of hypersphere $l$. 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 $\alpha$ 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.

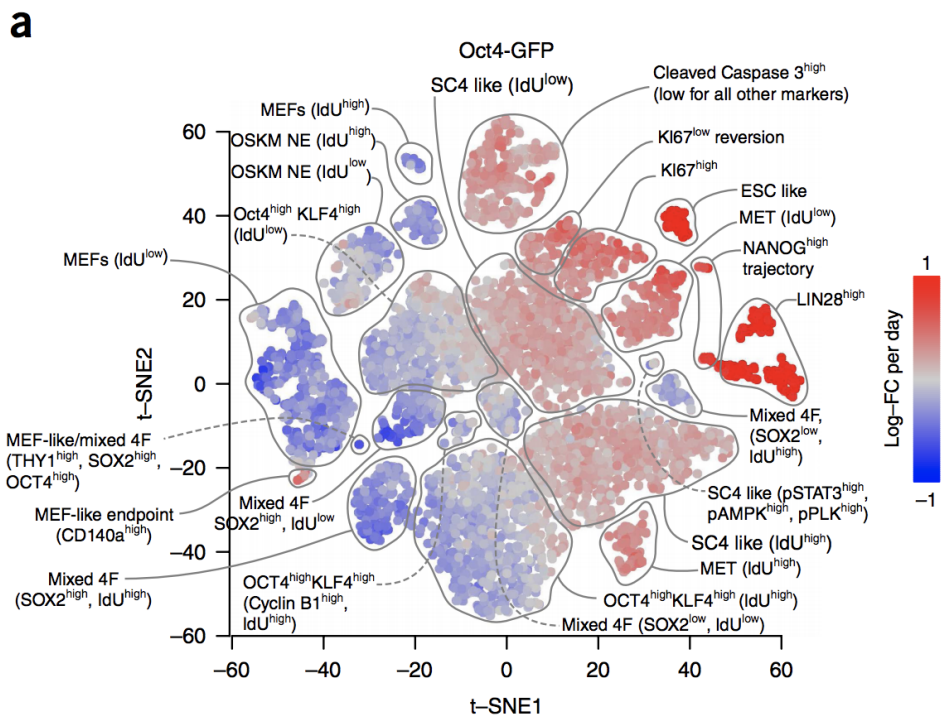



Figure: from Lun *et al.* Nature Methods. 2017. Each plotted point represents the median position of differentially abundant hyperspheres at an FDR of 5%

# 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.

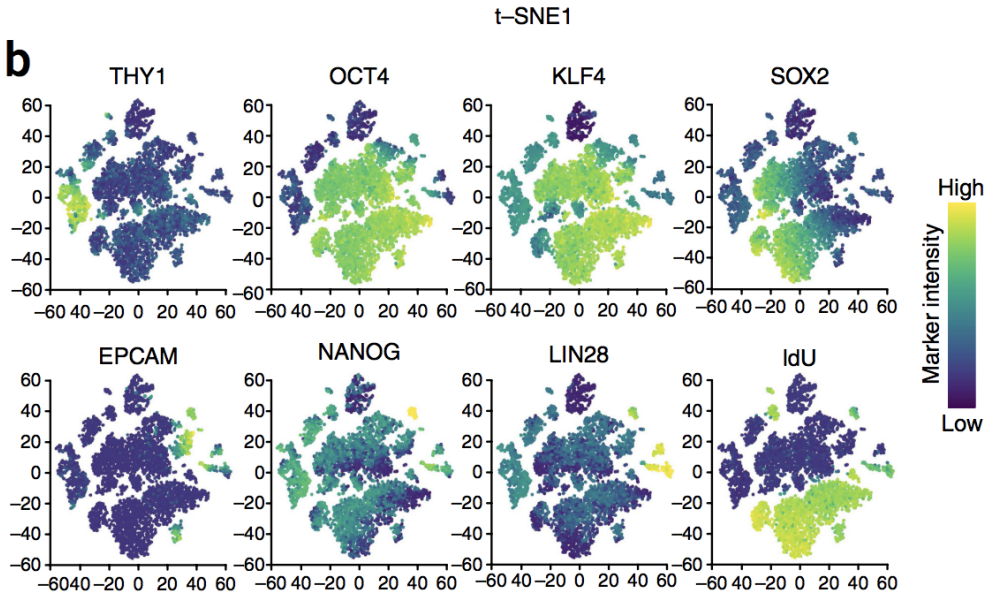



Figure: from Lun *et al.* Nature Methods. 2017. Color cells by the expression of individual markers.

# 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.

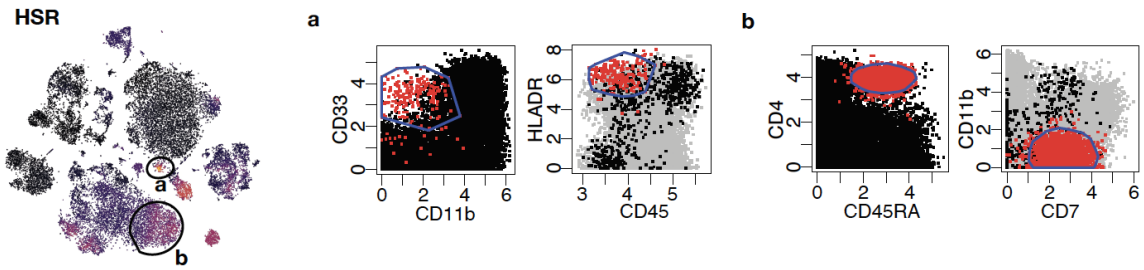



Figure: Nature Communications. 2020. We applied GateFinder to figure out what types of cells our prioritized cells were based on the combinations of markers that were expressed.

# 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.

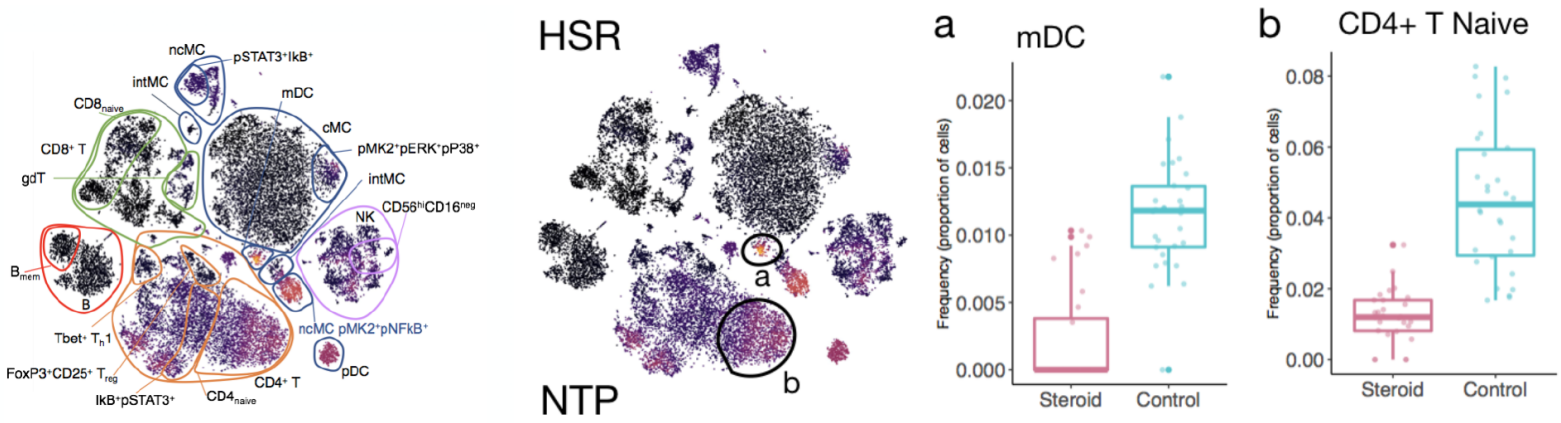
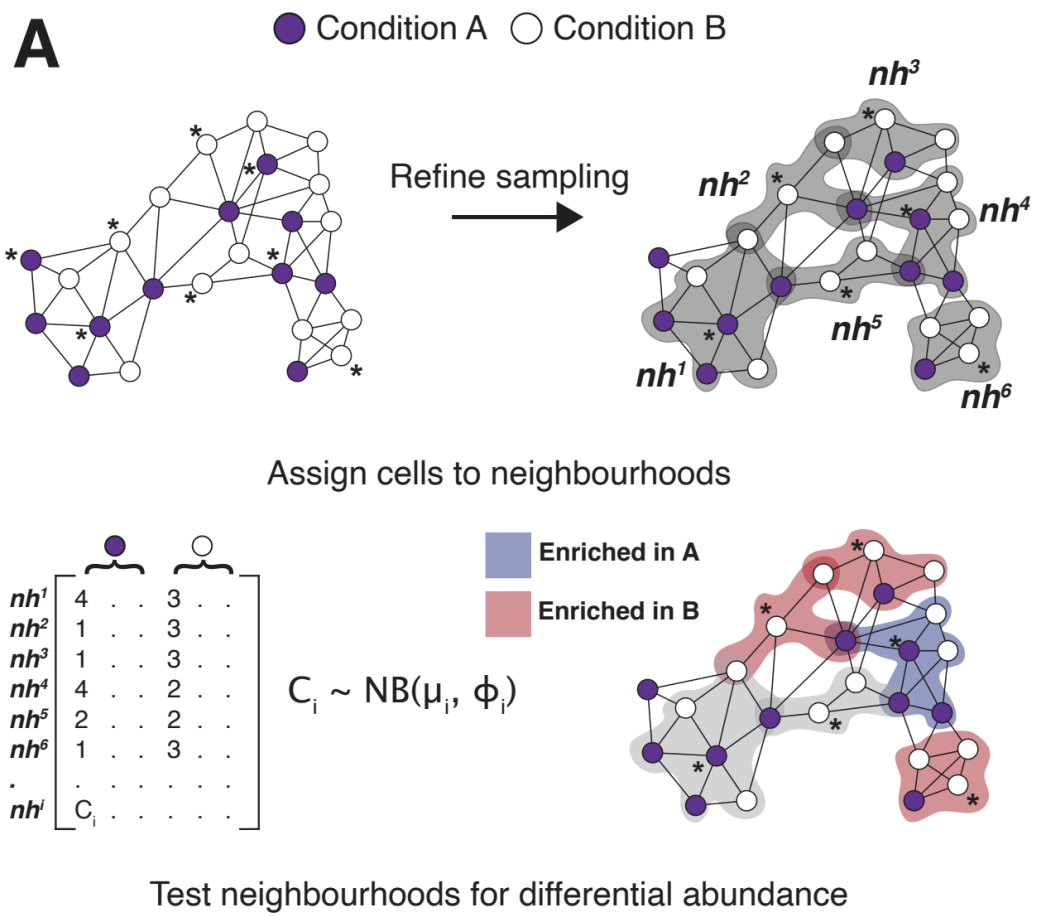



Figure: from Nature Communications. 2020.

# Milo

Milo is effectively a graph-based version of Cydar.
`https://www.nature.com/articles/s41587-021-01033-z`.



# General Overview of Milo

- Build $k$-NN graph of cells
- Define a representative set of nodes to serve as the ‘center‘ of neighborhoods across the graph
- Define the neighborhood of a node, $j$, as the collection of cells that are connected to node $j$ by an edge.
- Count cells in each neighborhood. You end up with a matrix of samples $\times$ counts of cells across neighborhoods.
- Test for differential abundance in neighborhoods Spatial FDR again to control for the proportion of neighborhoods that are false-positive.

# Comparing to some other approaches we looked at

| Method | DA quantification | Clustering-free | Statistical testing | Flexible experimental design |
|---|---|---|---|---|
| Milo | Negative binomial GLM | ✓ | ✓ | ✓ |
| Cydar | Negative binomial GLM | ✓ | ✓ | ✓ |
| DA-seq | Logistic classifier prediction | ✓ | ✓ | ✗ |
| Louvain + GLM | Negative binomial GLM | ✗ | ✓ | ✓ |
| MELD | Kernel density estimation | ✓ | ✗ | ✗ |

# How Does Milo Do?

MCC (Matthews Correlation Coefficient) is a performance metric that measures performance from integrating multiple performance metrics.

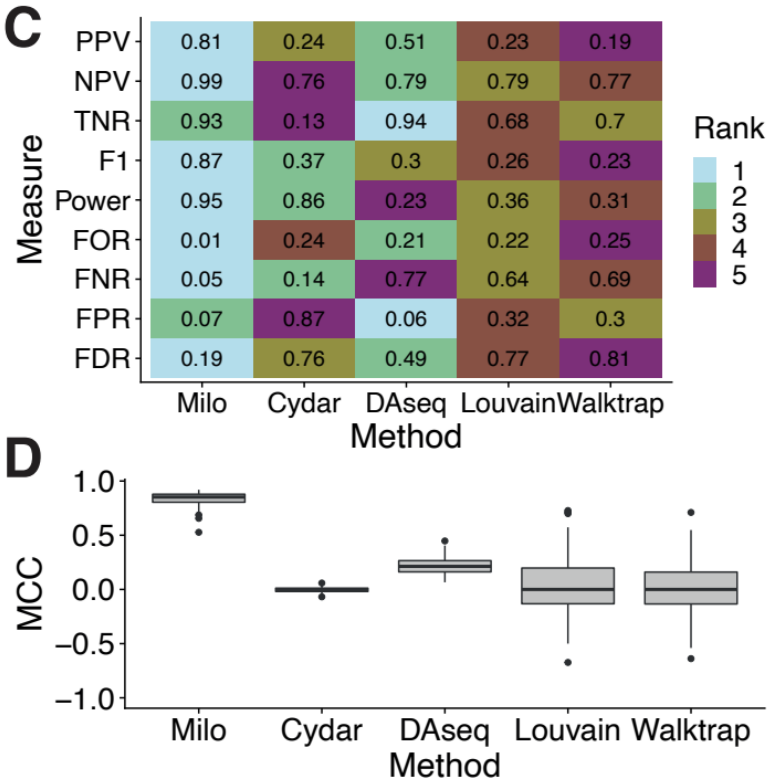



Figure: from Dann *et al.* 2020. BioArXiv

# Resulting Visualization of Milo

Milo visualizes the graph between the subsets of selected nodes that were used to form neighborhoods.

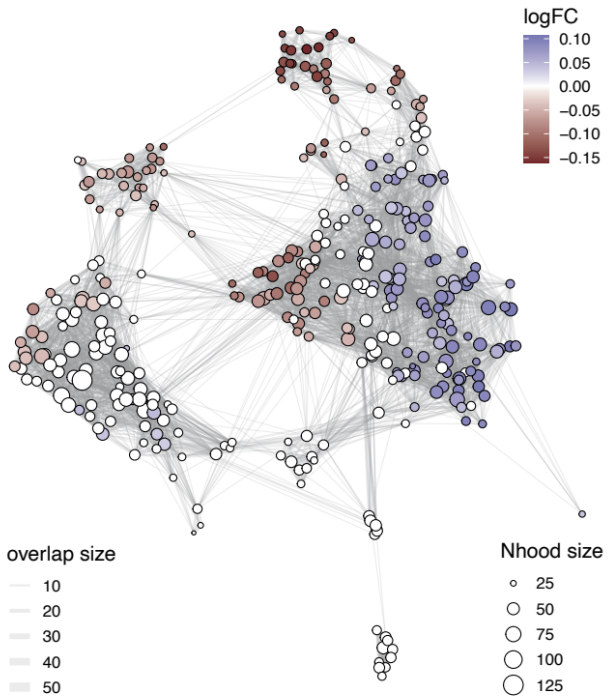



Figure: from Dann *et al.* 2020. BioArXiv. Here the data are single thymic epithelial cells sampled from mice from age 1 to 52 weeks.

# Connecting to Ground Truth Labels of the Cells

Each cell as an 'age' associated with it. We can see that cells belonging to neighborhoods that had an increased abundance of cells with age are colored blue (previous slide).

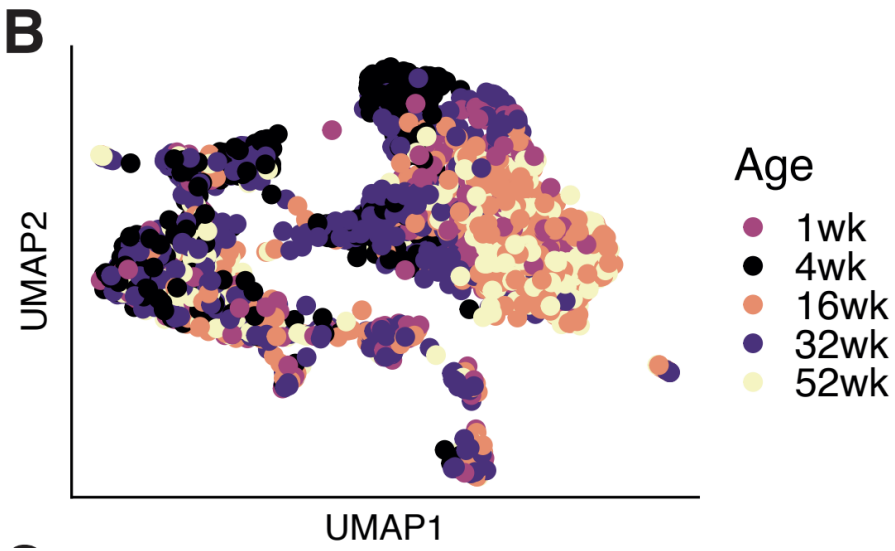



Figure: from Dann *et al.* 2020. BioArXiv. Cells are colored by the age of the mouse that they came from.

# Example 2: Healthy vs Cirrhotic Mice

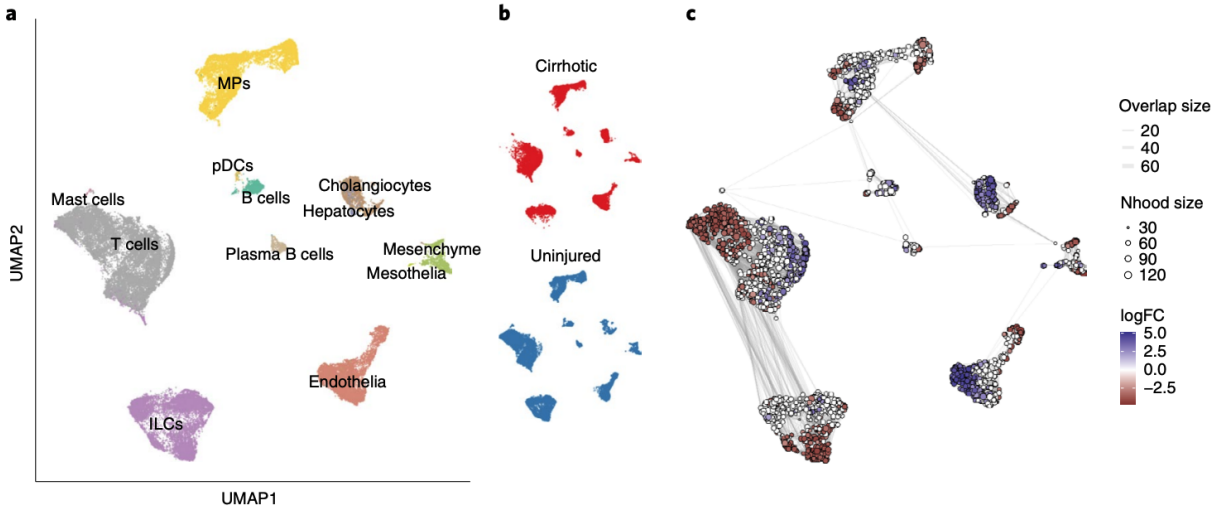



Figure: Look at neighborhoods prioritized by Milo and their log fold change in abundance.

# Zooming in on only the endothelial lineage

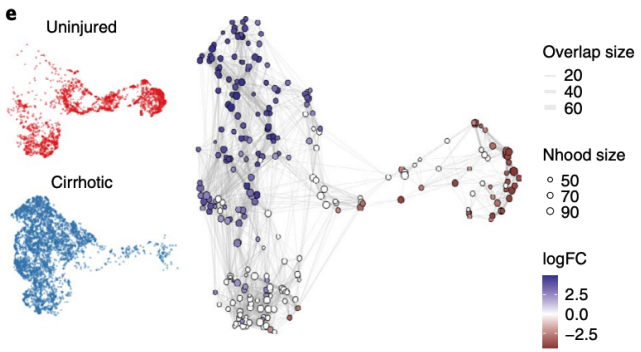



Figure: Same as previous slide, but only the endothelial lineage.

# Thoughts on Milo + Comparison with Meld

- **Neighborhoods Initialized Randomly.** It seems that we can really do better than choosing nodes at random to serve as the centers of the neighborhoods.
- This problem of choosing seeds on a graph is actually hard. How do you choose seeds that are sufficiently equidistant from other seeds. (For example, this would be easier on a grid)

# Example : Sampling on a Grid

Fig. 1. Images segmented using SLIC into superpixels of size 64, 256, and 1,024 pixels (approximately).

Figure: from `https://ieeexplore.ieee.org/abstract/document/6205760`

What if we thinking about a set of control samples as a *background* that we can compare samples from our experiments to?

# Switching Gears - Dealing with Background Populations and Data

- Consider high-dimensional gene expression measurements collected from people from all over the world.
- Suppose these patient samples also correspond to healthy and cancer patients.
- If the question is to find gene expression patterns associated with cancer subtypes, PCA on our samples may mostly reflect demographic variation between patients, rather than biological variation related to cancer subtypes.

# Intro to Contrastive PCA

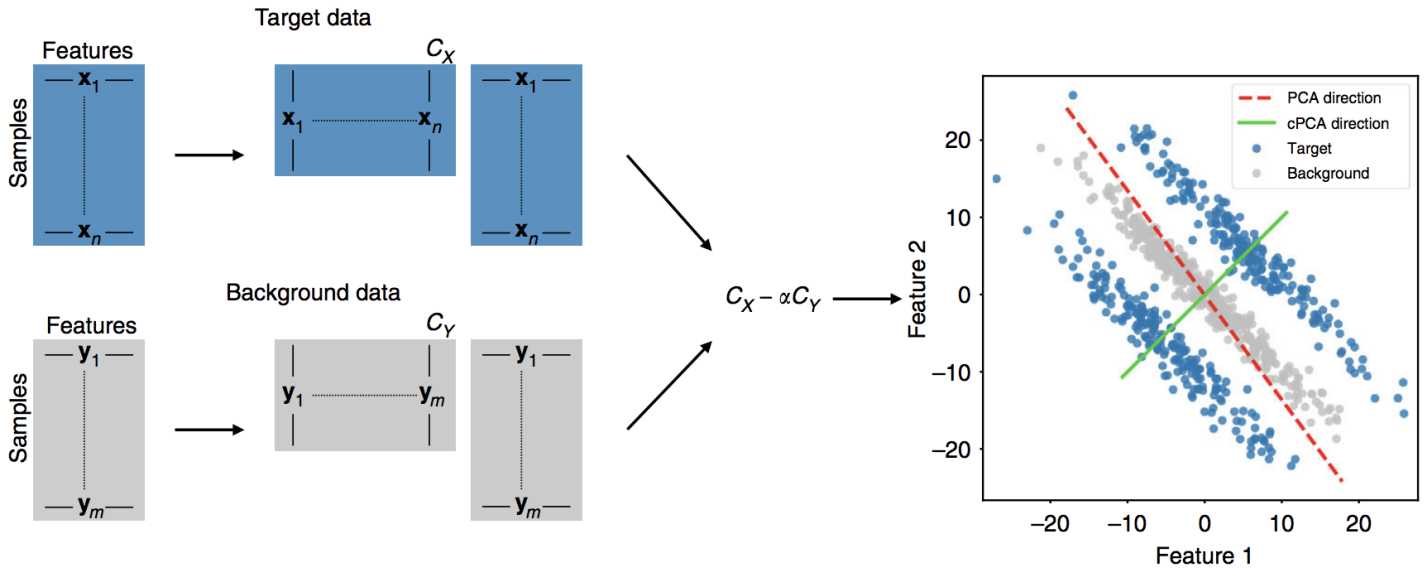



Figure: from Abid *et al.* Nature Communications. 2018. When projecting the data, the goal is to find the target direction that has the highest variance in the target data in comparison to the background data.

# Thinking About Background Data

- Given two groups of datapoints (e.g. patient measurements), you can imagine there is variance common to both datasets and variance characteristic of each one.
- For example, thinking about a control group and a disease group, both have population-level variation, but the disease group has particular disease subtypes.
- As another example, consider time series data when you want to decouple variation from a particular timepoint from variation across the entire time series.
- Choice of background dataset is important here and should ideally contain ‘structure‘ that we would like to remove form the target data.

# Synthetic MNIST Example

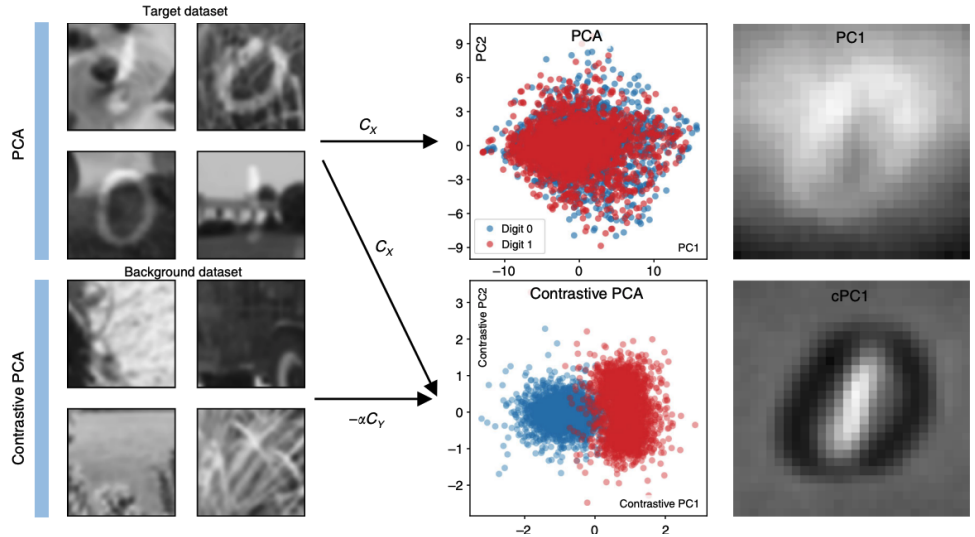



Figure: MNIST 1s and 0s are overlaid on scenes. A background dataset of grass is introduced, which allows better separation of 1s and 0s. In **c**, bright pixels shown the contribution to PCs or cPCs.

# Motivating Biological Examples

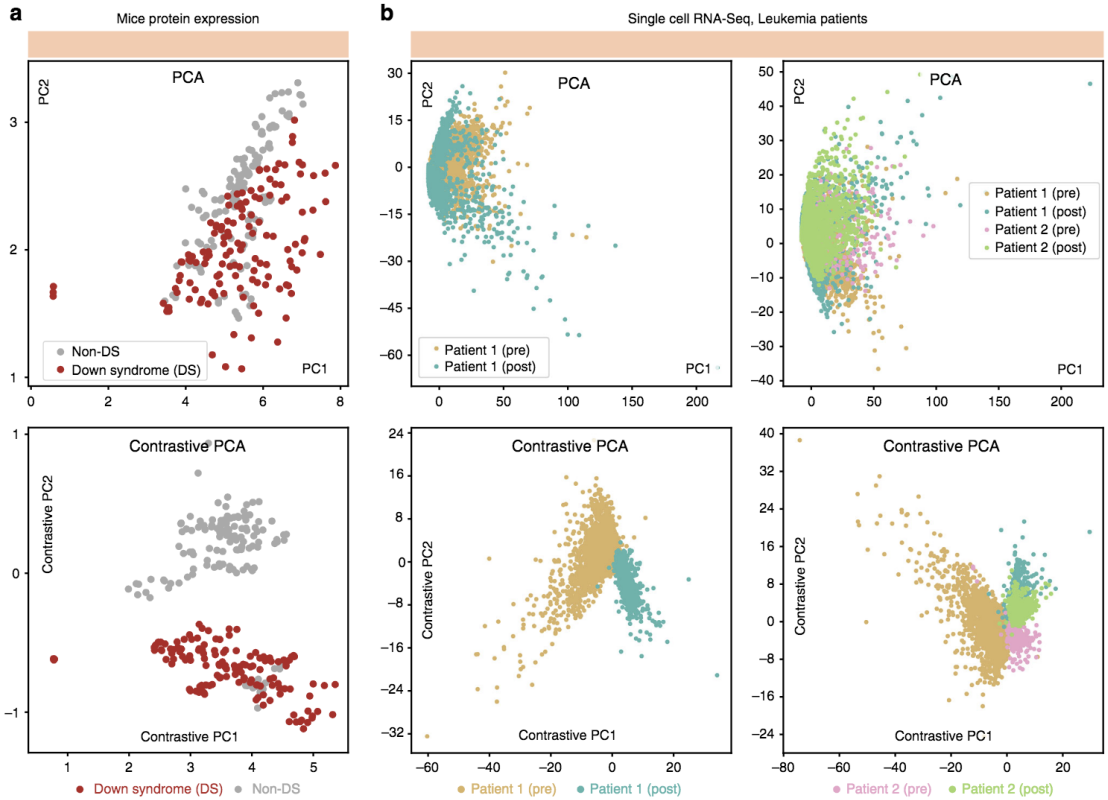



Figure: from Abid *et al.* Nature Communications. 2018. (Left) : Protein expression in Down Syndrome vs Non Down Syndrome Mice. Background data are control mice that have not been exposed to shock therapy. (Right) scRNA-seq of pre and post transplant, with healthy donor as background.

# cPCA Problem Setup

- Assuming we start with $d$-dimensional target data $\{\mathbf{x}_i \in \mathbb{R}^d\}$ background data $\{\mathbf{y}_i \in \mathbb{R}^d\}$

For some direction vector, $\mathbf{v} \in \mathbb{R}^d_{\text{unit}}$ the variance it accounts for in the target and background data can be expressed as,

$$\text{Target data variance} : \lambda_X(\mathbf{v}) \stackrel{\text{def}}{=} \mathbf{v}^T \mathbf{C}_X \mathbf{v}$$

$$\text{Background data variance} : \lambda_Y(\mathbf{v}) \stackrel{\text{def}}{=} \mathbf{v}^T \mathbf{C}_Y \mathbf{v}$$

# What is happening here and what does this remind you of?

Given a contrast parameter $\alpha \geq 0$ that quantifies the trade-off between having high target variance and low background variance, cPCA computes the contrastive direction $\mathbf{v}$ by optimizing

$$\mathbf{v}^* = \text{argmax}_{\mathbf{v} \in \mathbb{R}^d_{\text{unit}}} \ \lambda_X(\mathbf{v}) - \alpha \lambda_Y(\mathbf{v})$$

This problem can be rewritten as

$$\mathbf{v}^* = \text{argmax}_{\mathbf{v} \in \mathbb{R}^d_{unit}} \ \mathbf{v}^T \left(C_X - \alpha C_Y\right) \mathbf{v}$$

# cPCA is Quite Simple!

**Algorithm 1** cPCA for a Given $\alpha$

**Inputs:** target data $\{\mathbf{x}_i\}_{i=1}^n$; background data $\{\mathbf{y}_i\}_{i=1}^m$; contrast parameter $\alpha$; the number of components $k$.

Centering the data $\{\mathbf{x}_i\}_{i=1}^n$, $\{\mathbf{y}_i\}_{i=1}^m$.
Calculate the empirical covariance matrices:

$$C_X = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T, C_Y = \frac{1}{m}\sum_{i=1}^{m} \mathbf{y}_i\mathbf{y}_i^T.$$

Perform eigenvalue decomposition on

$$C = (C_X - \alpha C_Y).$$

Compute the the subspace $V \in \mathbb{R}^k$ spanned by the top $k$ eigenvectors of $C$.
**Return:** the subspace $V$.

Figure: Just do eigendecomposition on **C** and consider the eigenvectors corresponding to the top $k$ eigenvalues of $C$.

# Effect of Varying $\alpha$

For $\alpha = 0$, cPCA will create directions that maximize the target variance. For higher $\alpha$, directions with smaller background variance become more important.

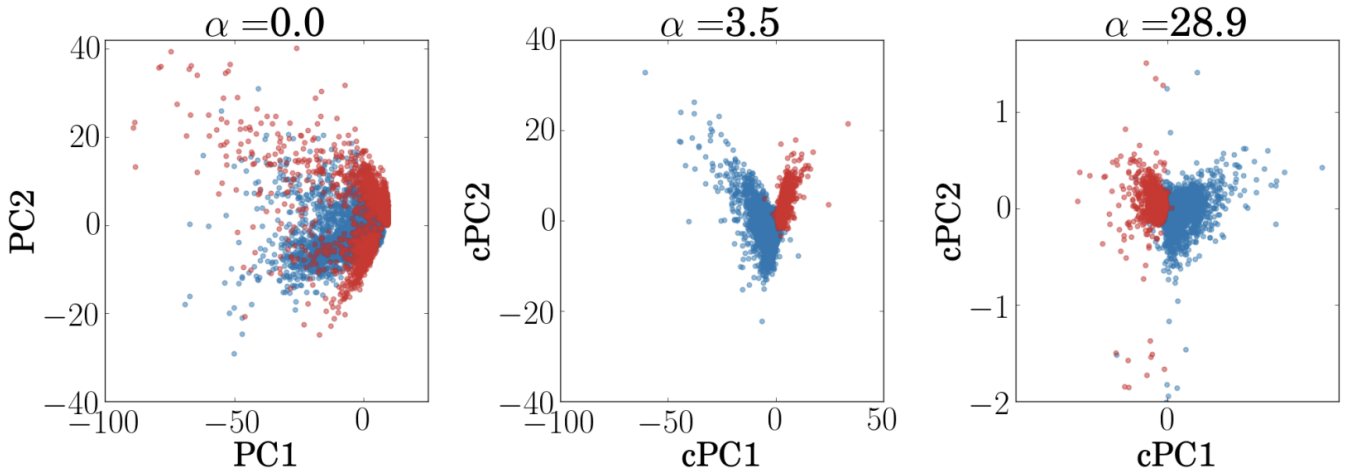



Figure: from Abid *et al.* Nature Communications. 2018. This dataset is visualizing cells from two different samples.

# Recap

We have now seen Meld vs Cydar vs Milo vs cPCA (if we have a background). Thoughts? What would you do to compare your samples from disparate clinical or experimental groups?

# Combining Multiple Single-Cell Datasets

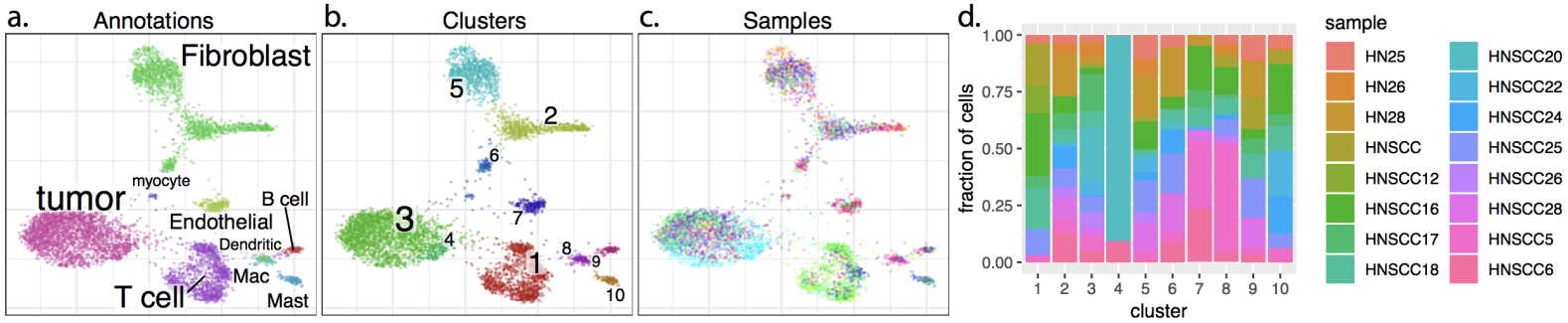



Figure: from Barkas *et al.* Nature Methods. 2019. Conos looks at how to integrate cells from multiple datasets (patients, tissues, etc.)

- The problem is a bit different from batch effect effect correction where you can identify technical artifacts and get rid of them. Cell-populations might be completely missing from particular datasets.

# Conos Overview: Construct a Joint Between-Cell Graph

The goal is to establish a unified graph representation of the multiple single-cell datasets. Specifically, to infer cell-populations across all datasets, Conos seeks to infer inter-cell edges between datasets.

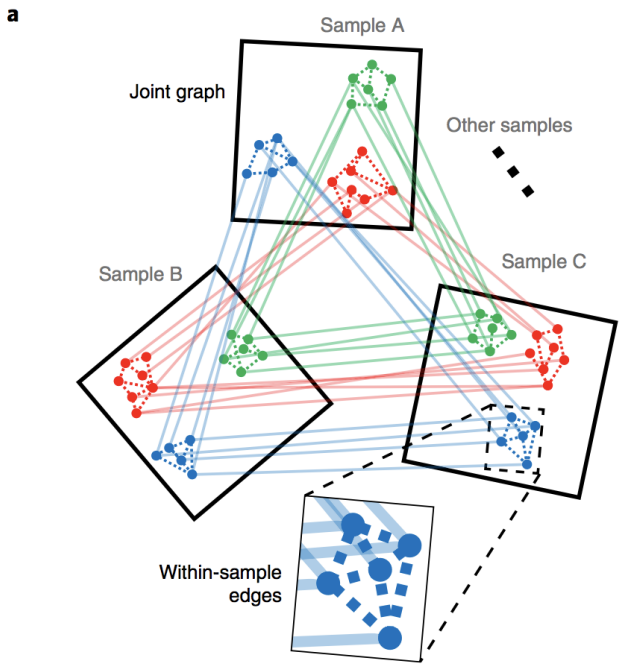



Figure: from Barkas *et al.* Nature Methods. 2019.

# Pairwise Dataset Alignment

- As a pre-processing step, choose a set of high-variance genes. (The authors use 2,000).
- For a pair of datasets, $i$ and $j$, let $G_i$ and $G_j$ denote their corresponding set of features measured per cell. Then consider only features that are measured in both datasets (so $G_i \cap G_j$)
- The similarity between cells $K$ and $l$ in datasets $i$ and $j$ is $w_{kl} = \exp(-\frac{||M_k^i - M_l^j||}{\sigma})$

# Creating the Joint Graph

- Use $w_{kl}$ for $k$-NN gaphs
- For **inter-sample edges**, connect each cell to its 15 nearest neighbors by default
- For **intra-sample edges**, connect each cell to its 5 nearest neighbors.
- Create joint clusters by clustering the graph with a graph-partitioning method.