

Comp790-166: Computational Biology

Lecture 9

February 17, 2021

Announcements

- Homework is due by the end of today!
- Gust lecture on Tuesday by Maria Brbic.
- Project proposal template coming soon....
- We will do a reading summary next week. I will send an email soon of suggested papers.

Today

- Differential cell-population analysis
 - Cydar
 - DiffCyt
 - Milo
- Contrastive PCA

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.

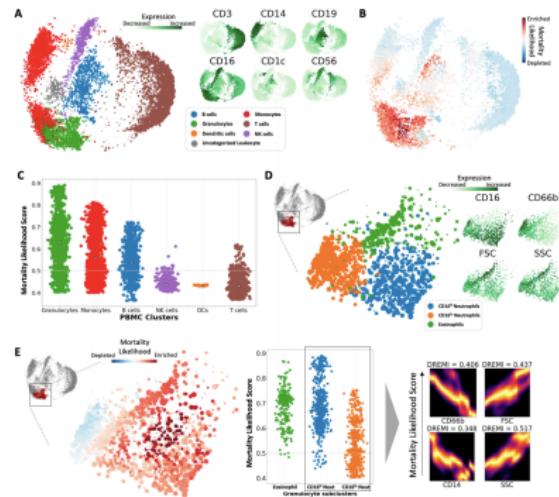


Figure: from Kuchroo et al. 2020. <https://www.biorxiv.org/content/10.1101/2020.11.15.383661v1.abstract>

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!

Welcome Cydar (Nature Methods 2017).

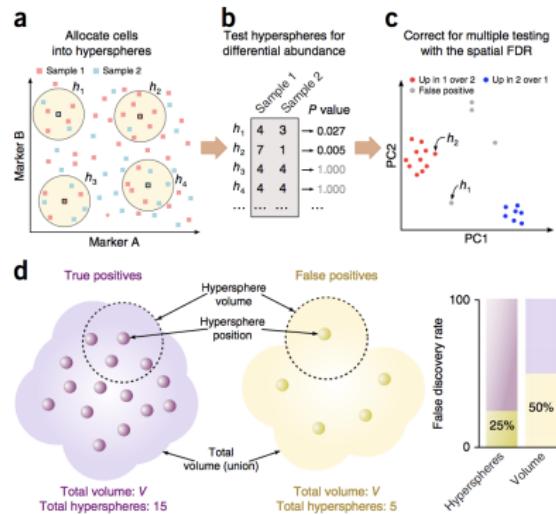


Figure: from Lun et al. Nature Methods. 2017.

Clustering → hypersphere

- Cells from across all samples are clustered using k -means with $k = \sqrt{N}$, where N is the number of cells
- From here, cells will be assigned to a hypersphere, based on the initial coarse-graining through clustering
- Hyperspheres are established around randomly chosen cells, and are particularly established around every 10th cell.

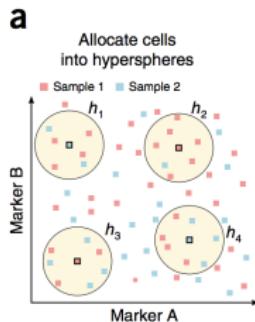


Figure: from Lun *et al.* Nature Methods. 2017.

Assigning Cells to Hyperspheres

- Let $|j - t|$ denote the Euclidean distance (in marker space) between cell j and the center of cluster t
- Let $|h - t|$ denote the distance (in marker space) between the center of cluster t and the center of hypersphere h .

A cell j in cluster t was only considered for assignment in hypersphere h if,

$$r + |j - t| \geq |h - t| \tag{1}$$

Unpacking...

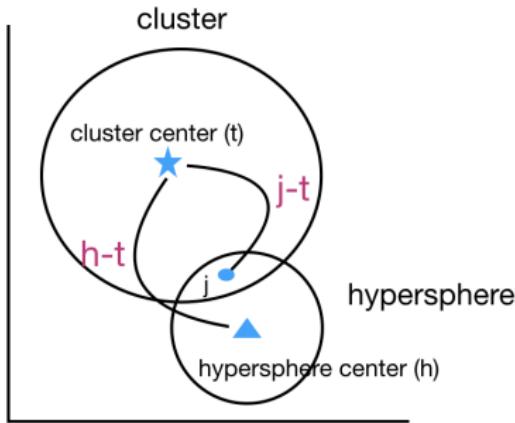


Figure: The cluster and hypersphere centers should be sufficiently close

Hypersphere business and choosing r

- Centering hyperspheres on existing cells ensures that there will be no empty hyperspheres.
- Based on this definition, many hyperspheres will be overlapping.
- What is r
 - The radius, r of each hypersphere dictates the trade-off between the number of cells assigned there and spatial resolution (distinguishing between neighboring hyperspheres).
 - They have chosen $r = 0.5\sqrt{M}$

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.
- 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)

	Sample 1	Sample 2	P value
h_1	4 3		→ 0.027
h_2	7 1		→ 0.005
h_3	4 4		→ 1.000
h_4	4 4		→ 1.000
...

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

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. Note that this is not just simply counting the proportion of hyperspheres where this is true.

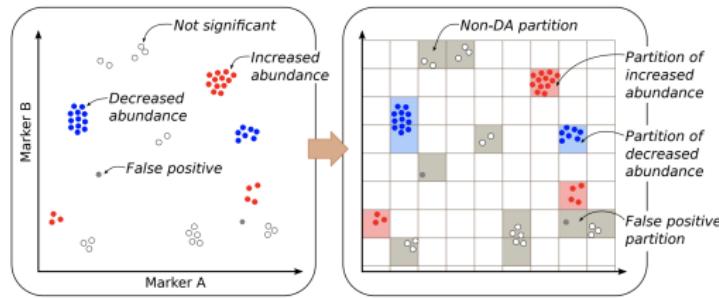


Figure: from Lun et al. Nature Methods. 2017. Each circle is representing a hypersphere colored by increase in abundance (red), decrease (blue), no change

Spatial FDR, Continued

The main idea is to partition the M -dimensional space, and control FDR across the partitions (where each partition will contain multiple hyperspheres that should be similar in terms of containing true positives vs false positives.)

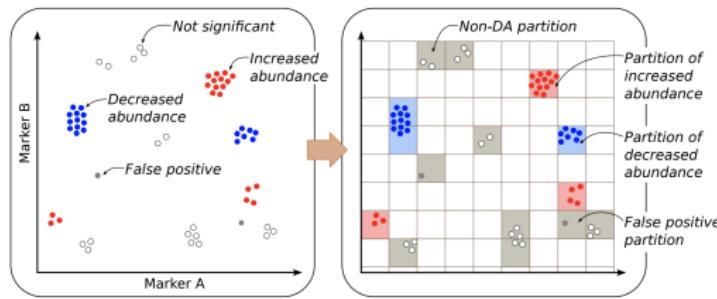


Figure: from Lun *et al.* Nature Methods. 2017.

Weighted Benjamini-Hochberg

- Assume that for n hyperspheres that there are n ordered p -values from the statistical test with $p_1 \leq p_2 \leq \dots p_n$.
- 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 . If there are few hyperspheres within a local region, there is a high probability that a particular hypersphere will be selected as representative of that partition.

Weighted BH, Coontinued

Then the weighted BH method will reject any null hypothesis where the p -value is less than the following. Here α is some threshold, like $\alpha = 0.5$.

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

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 subpopulations that change over time.

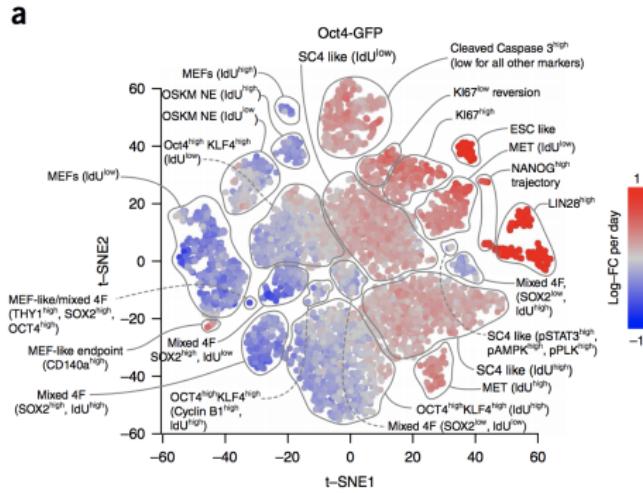


Figure: from Lun *et al.* Nature Methods. 2017. Each plotted point represents the median position of differentially abundant hyperspheres at an FDR of 5%

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.

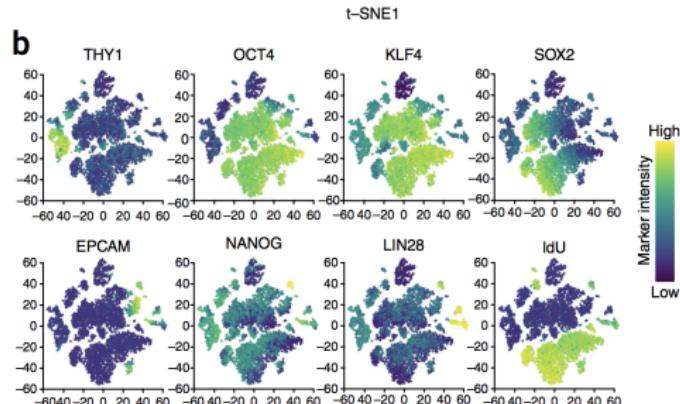


Figure: from Lun et al. Nature Methods. 2017. Color cells by the expression of individual markers.

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.

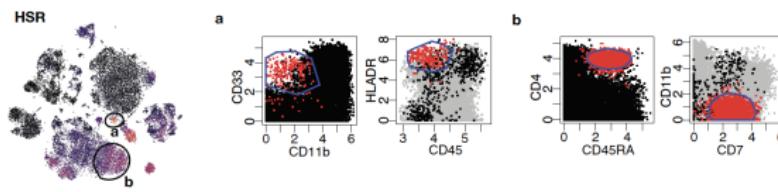


Figure: from Stanley *et al.* Nature Communications. 2020. We applied GateFinder to figure out what types of cells our prioritized cells were based on the combinations of markers that were expressed.

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 Bonferroni)
- 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 → test → 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.

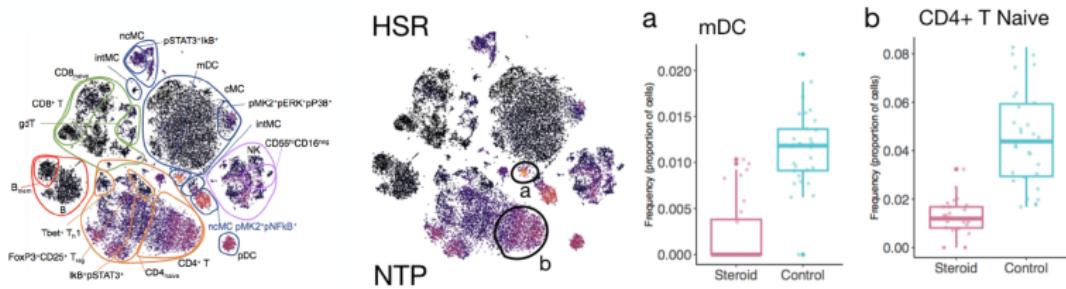
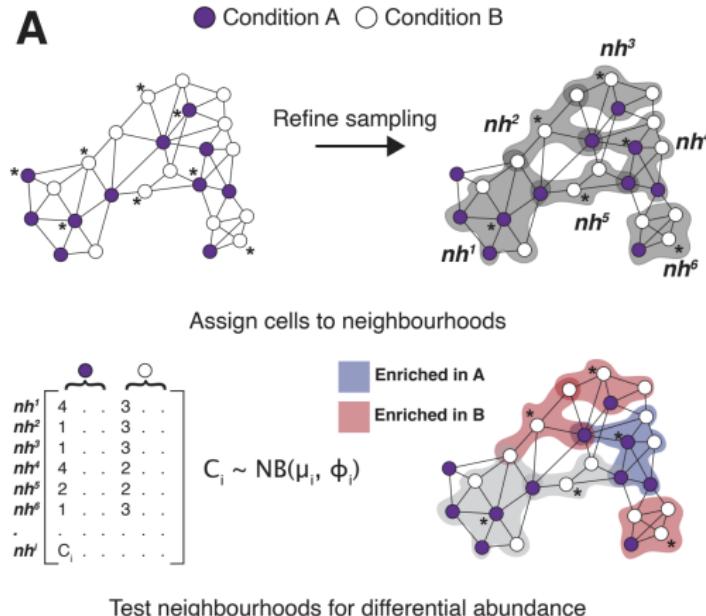


Figure: from Stanley *et al.* Nature Communications. 2020.

Milo

Milo is a graph-based version of Cydar, just recently on BioArXiv.

<https://www.biorxiv.org/content/10.1101/2020.11.23.393769v1>.



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.

How Does Milo Do?

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

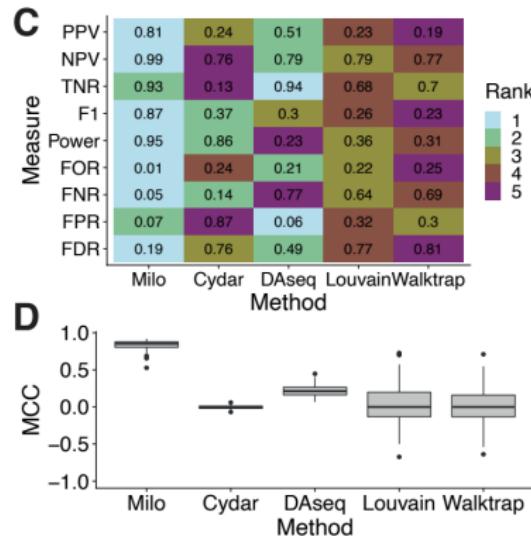


Figure: from Dann *et al.* 2020. BioArXiv

Resulting Visualization of Milo

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

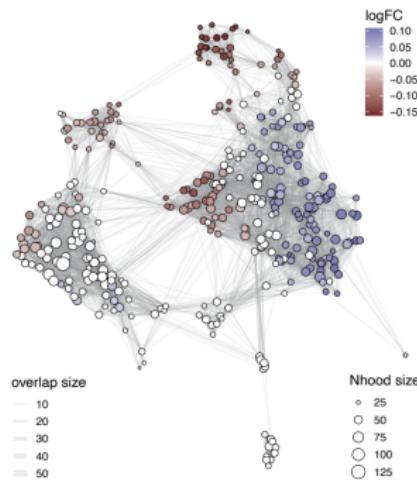


Figure: from Dann *et al.* 2020. BioArXiv. Here the data are single thymic epithelial cells sampled from mice from age 1 to 52 weeks.

Connecting to Ground Truth Labels of the Cells

Each cell has 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.

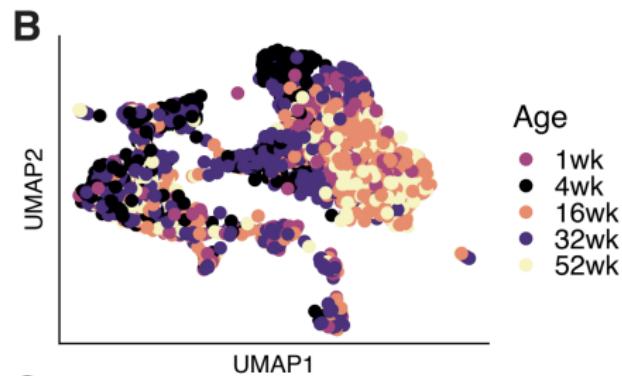


Figure: from Dann *et al.* 2020. BioArXiv. Cells are colored by the age of the mouse that they came from.

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)

Meld vs DA Style Approach

- **Meld:** Collect the set of prototypical cells from each condition (according to Meld Score). Examine differences in prototypical cells from each condition.
- **DA Approach:** Coarse grain cells according to clusters, hyperspheres, neighborhoods, whatever. Do multiple tests and correct.
- Which do you all think is better? Or how about a hybrid of the two?

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

Intro to Contrastive PCA

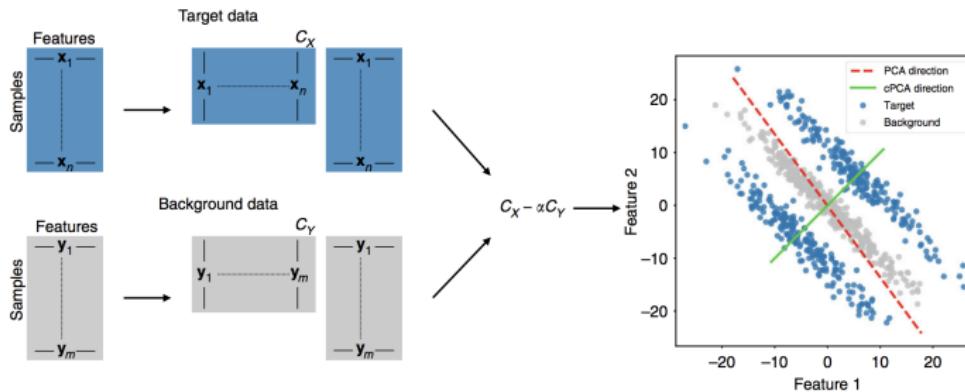


Figure: from Abid *et al.* Nature Communications. 2018. When projecting the data, the goal is to find the target direction that has the highest variance in the target data in comparison to the background data.

Thinking About Background Data

- Given two groups of datapoints, 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.
- Correcting for background will cause hidden structure to emerge

Motivating Biological Examples

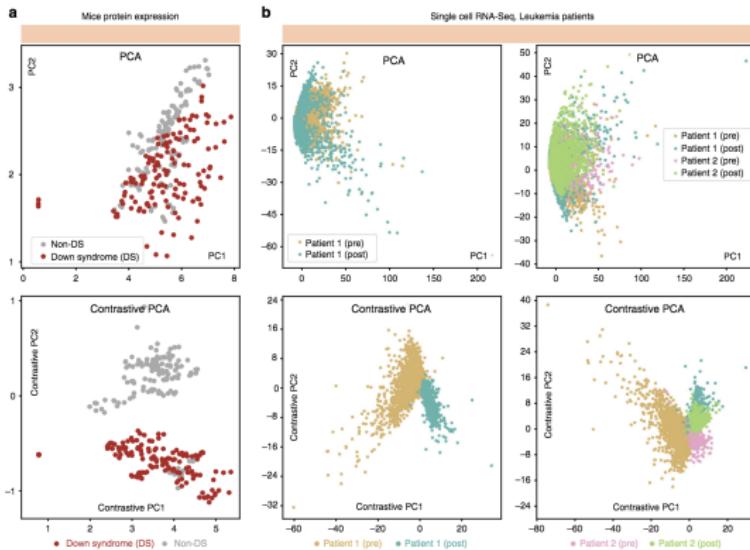


Figure: from Abid et al. Nature Communications. 2018. (Left) : Protein expression in Down Syndrome vs Non Down Syndrome Mice. Single cell data pre and post transplant.

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 any $\mathbf{v} \in \mathbb{R}_{\text{unit}}^d$,

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

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

cPCA is Quite Simple!

Algorithm 1 cPCA for a Given α

Inputs: target data $\{\mathbf{x}_i\}_{i=1}^n$; background data $\{\mathbf{y}_i\}_{i=1}^m$; contrast parameter α ; 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 \mathbf{C} and consider the eigenvectors corresponding to the top k eigenvalues of \mathbf{C} .

Effect of Varying α

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

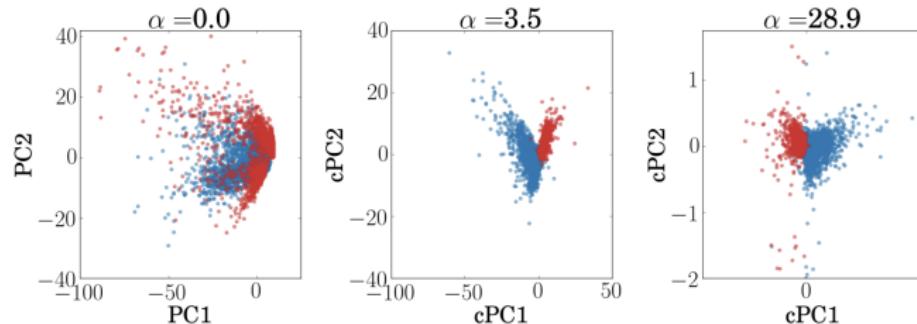


Figure: from Abid et al. Nature Communications. 2018. This dataset is visualizing cells from two different samples.

Recap

- DA analysis with Cydar and Milo
- Thinking about Meld vs DA style analyses
- Contrastive PCA