Comp790-166: Computational Biology

Lecture 11

March 2, 2021

Announcements

- Reading summaries by the end of the week!
- Project proposal template is online, `https://github.com/stanleyn/Comp790-166-Comp-Bio/tree/main/Projects`
- Please sign up for a date and presentation time, `https://docs.google.com/spreadsheets/d/1puFjmjadZuuVIy6nyJJPHJbucrW4-KH-xFd2P3HjIEI/edit?usp=sharing`
- The format of presentations is each group talking for 5-7 mins with 5 mins for questions and discussion.

Project Presentations

- You will discuss what you wrote about in your project proposal for 5-7 minutes.
- Then we will discuss for 5 minutes
- It would be great to make a few slides to better illustrate your points, but you can do this presentation in whatever style you would like.

Today

- Contrastive PCA
- Combining multiple single-cell datasets and batch effects
  - Conos : graph-based integration of single-cell datasets
  - Harmony
  - CytofMerge

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

Example Question and Application

- 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

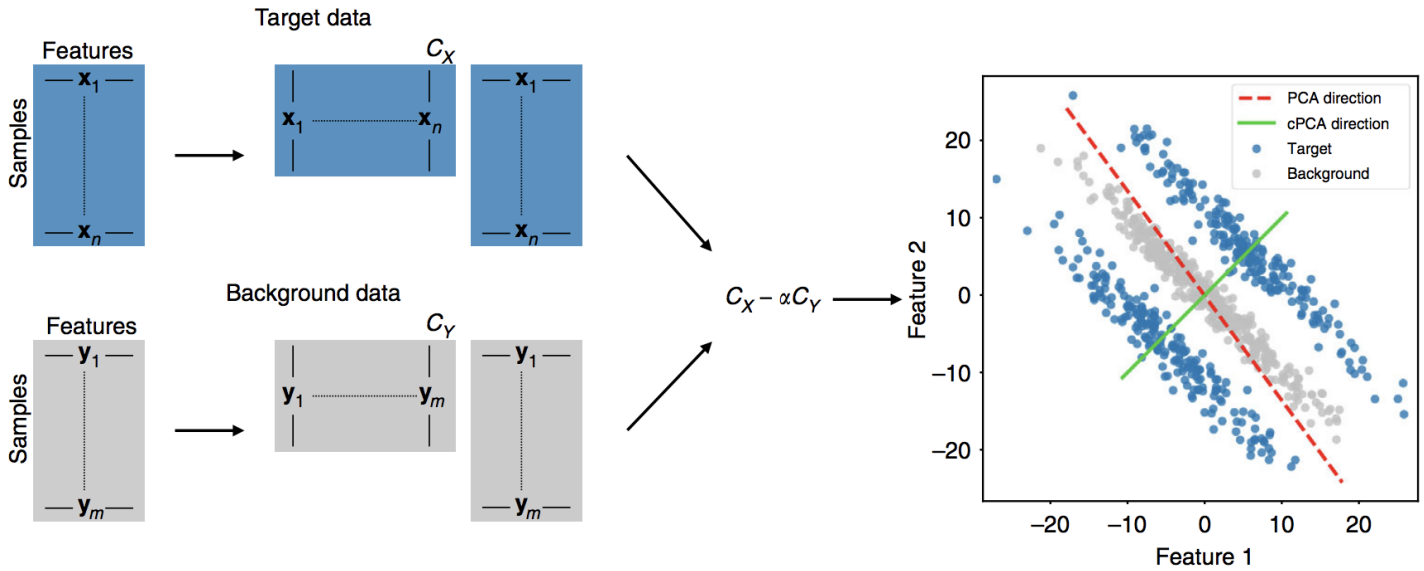

Figure: from Abid *et al.* Nature Communications. 2018. When projecting the data, the goal is to find the target direction that has the highest variance in the target data in comparison to the background data.

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 basckground dataset is important here and should ideally contain ‘structure‘ that we would like to remove form the target data.

Motivating Biological Examples

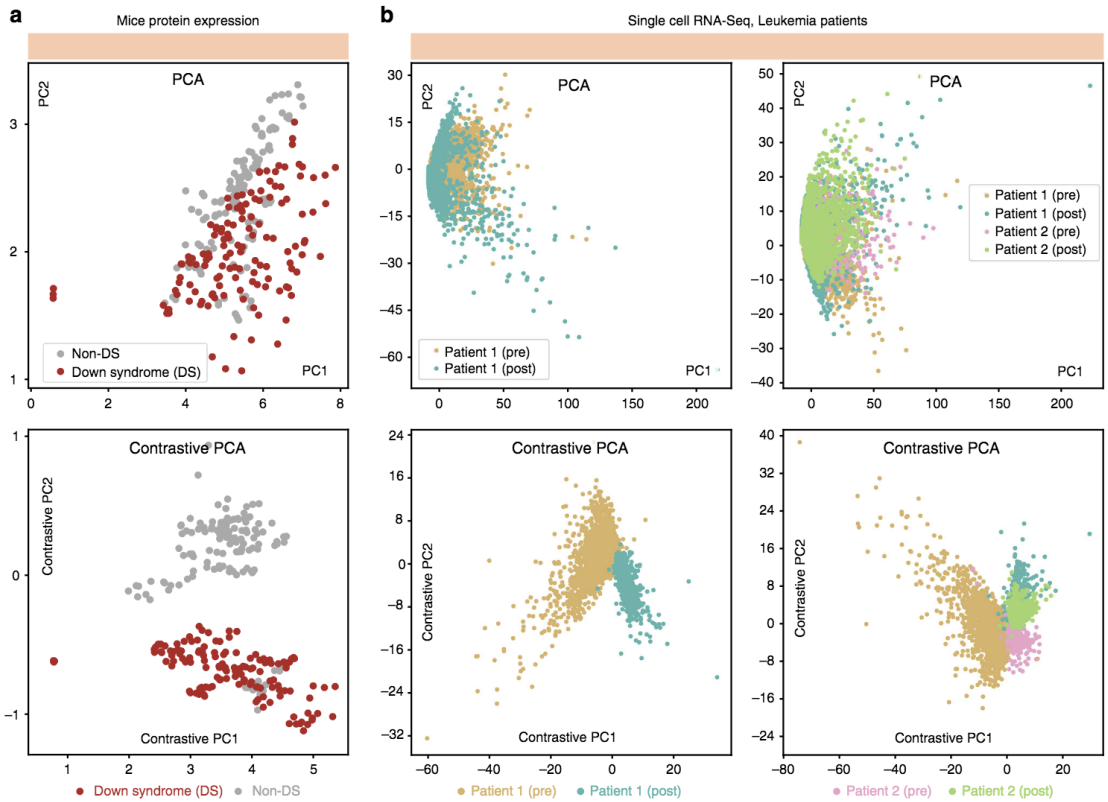

Figure: from Abid *et al.* Nature Communications. 2018. (Left) : Protein expression in Down Syndrome vs Non Down Syndrome Mice. Single cell data pre and post transplant.

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 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 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 **C** and consider the eigenvectors corresponding to the top k eigenvalues of 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.

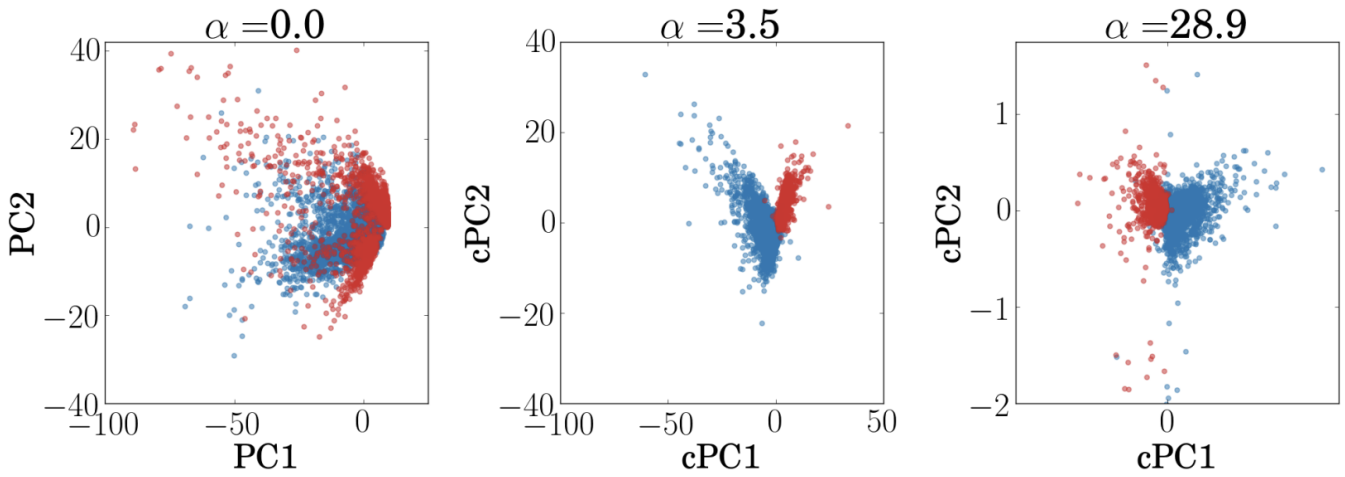

Figure: from Abid *et al.* Nature Communications. 2018. This dataset is visualizing cells from two different samples.

Combining Multiple Single-Cell Datasets

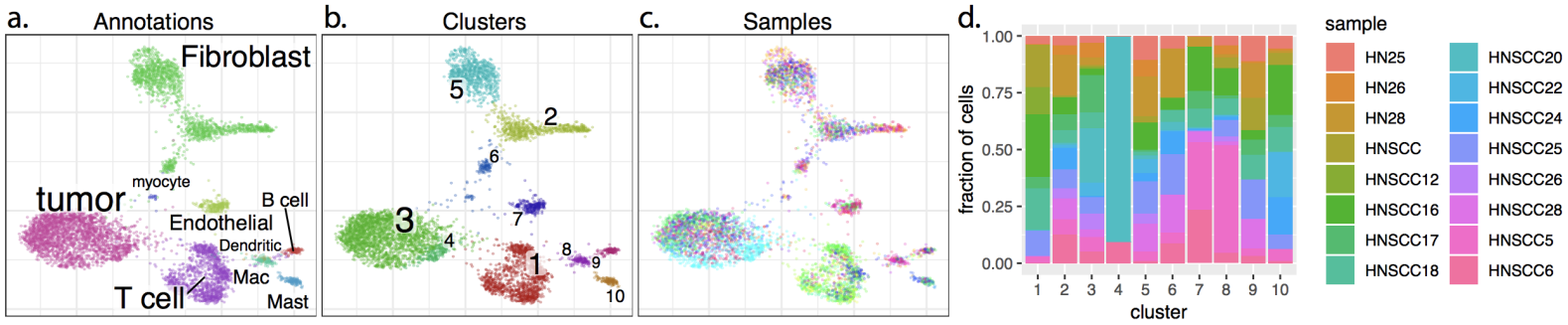

Figure: from Barkas *et al.* Nature Methods. 2019. Conos looks at how to integrate cells from multiple datasets (patients, tissues, etc.)

- 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.

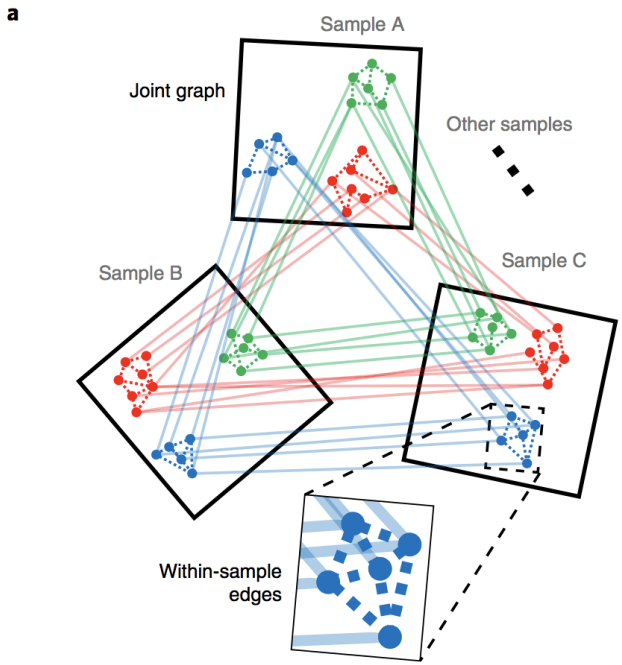

Figure: from Barkas *et al.* Nature Methods. 2019.

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 using a graph-based community detection method.

Rebalance Edge Weights

- Since samples are often collected across conditions, the authors wanted to provide flexibility to control how likely pairs of cell populations are to be mapped to each other, between conditions. Specifically, balance edge weights between cells connected between the same or different values of a factor.

The solution is to minimize the following,

$$\sum_{l=1}^{N_{\text{factors}}} \sum_{s=1}^{N_{\text{cells}}} \left| \frac{\sum_{t \in \text{adj}_l(s)} w_{st}}{\sum_{t \in \text{adj}(s)} w_{st}} - \frac{1}{N^s_{\text{factors}}} \right|$$

Unpacking...

$$\sum_{l=1}^{N_{\text{factors}}} \sum_{s=1}^{N_{\text{cells}}} \left| \frac{\sum_{t \in \text{adj}_l(s)} w_{st}}{\sum_{t \in \text{adj}(s)} w_{st}} - \frac{1}{N^s_{\text{factors}}} \right|$$

- N_{factors} is the total number of factor levels
- N_{cells} is the total number of cells.
- $adj(s)$ is the set of cells adjacent to cell s.
- $adj_l(s)$ is the set of cells adjacent to s and belong to factor level l.
- w_{st} is the weight of the edge between cells s and t
- N^s_{factors} is the number of different factors of cells connected to s.

Imbalance Between Factor l and cell s

For their minimization they first estimate the imbalance ratio for a cell s and a factor level, l as,

$$u_{sl} = N^s_{\text{factors}} \frac{\sum_{t \in \text{adj}_l(s)} w_{st}}{\sum_{t \in \text{adj}(s)} w_{st}}$$

Using Imbalance to Update Edge Weights

Edge weights are updated using the imbalance computed in the previous slide as,

$$w_{st} = \frac{w_{st}}{\sqrt{u_{sl}\, u_{tl_s}}}.$$

- Here l_c denotes the factor level of cell c.
- This process is repeated 50 times.

Effect of Alignment Strength

Here is an example varying alignment strength on a dataset containing cells from multiple technologies.

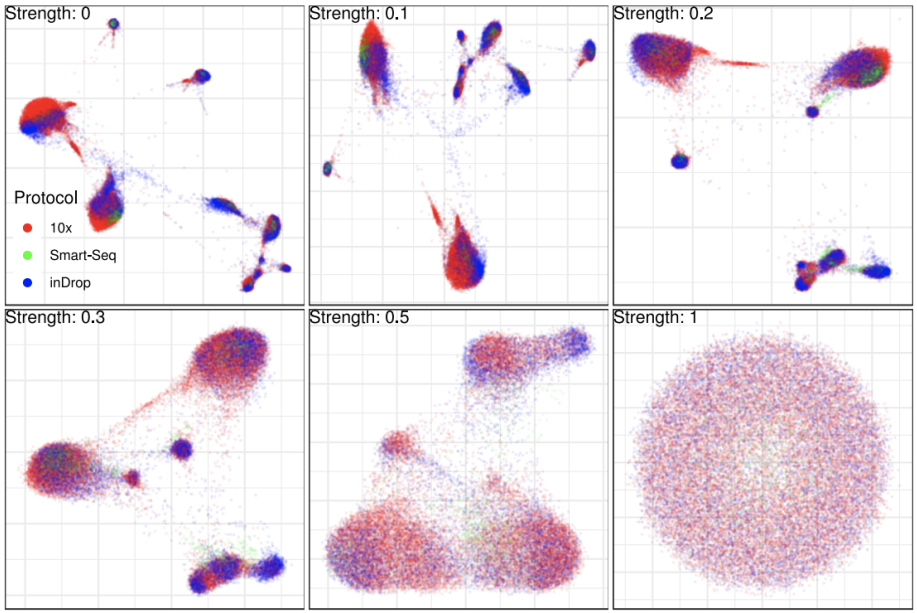

Figure: from Barkas *et al.* Nature Methods. 2019.

Example 1: Bone Marrow and Chord Blood

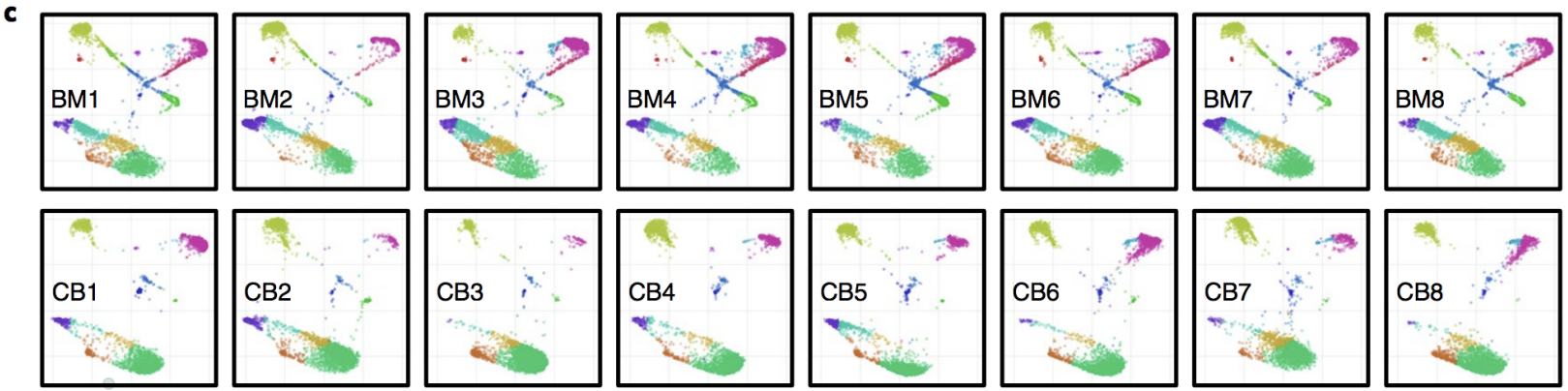

Figure: from Barkas *et al.* Nature Methods. 2019. You can see similarities and differences between cell-populations in each dataset.

Experiment 1: Adding Noise and Recovering Clusters

Noise was added to increase heterogeneity and decrease signal between samples.

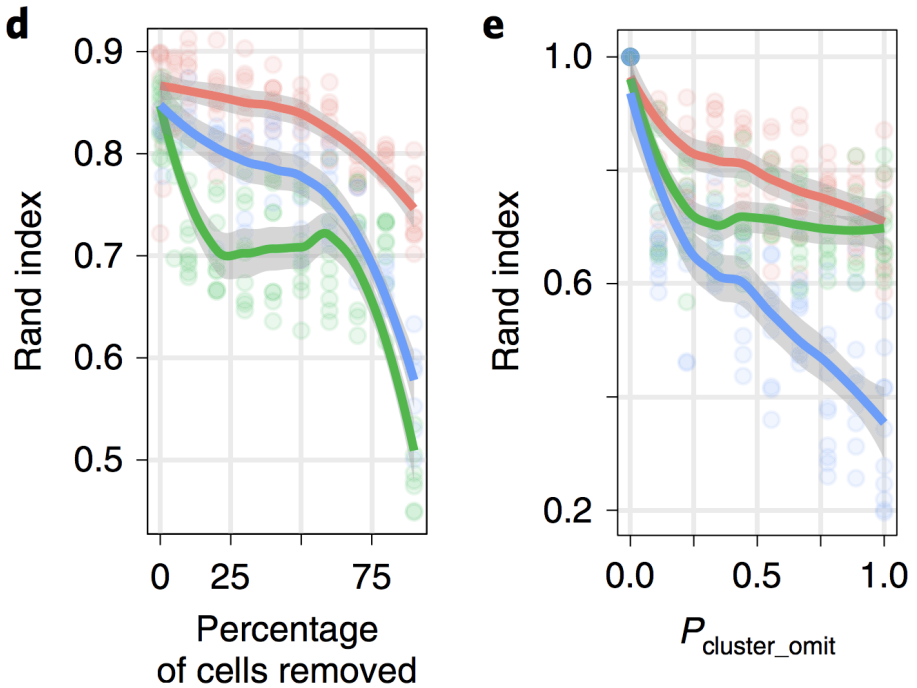

Figure: from Barkas *et al.* Nature Methods. 2019. They made perturbations by decreasing the number of cells or by decreasing the magnitude of expression-specific signatures.

Evaluating Cluster Entropy

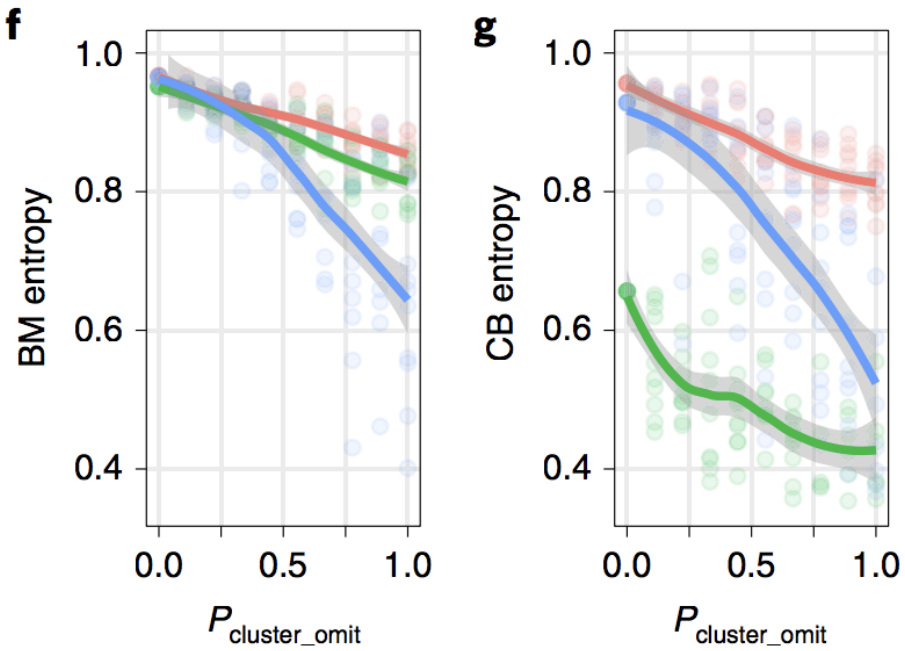

Figure: from Barkas *et al.* Nature Methods. 2019. Conos was able to maintain high entropy within clusters, or ensuring that clusters contained cells across all samples.

Visualizing the Joint Graph

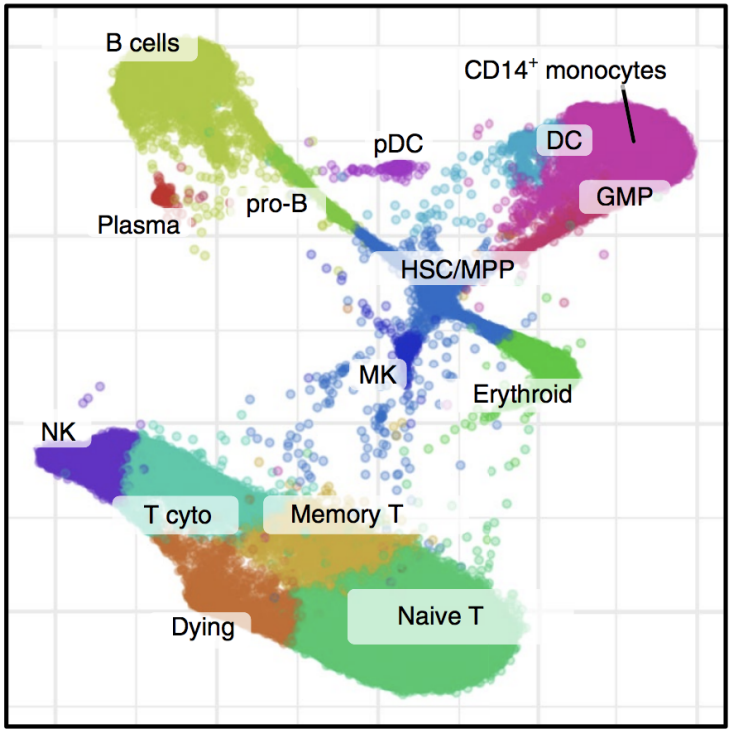

Figure: from Barkas *et al.* Nature Methods. 2019. The layout of the joint graph is determined by LargeVis.

Ensuring to Group Cells by Phenotype, not State

- Suppose you had CD4+ T cells in a cancer sample and CD4+ in a healthy sample. Because their function in the cancer sample is disrupted, it might not cause cells to cluster by phenotype. What we really want is a unified cluster of CD4+ across all samples.

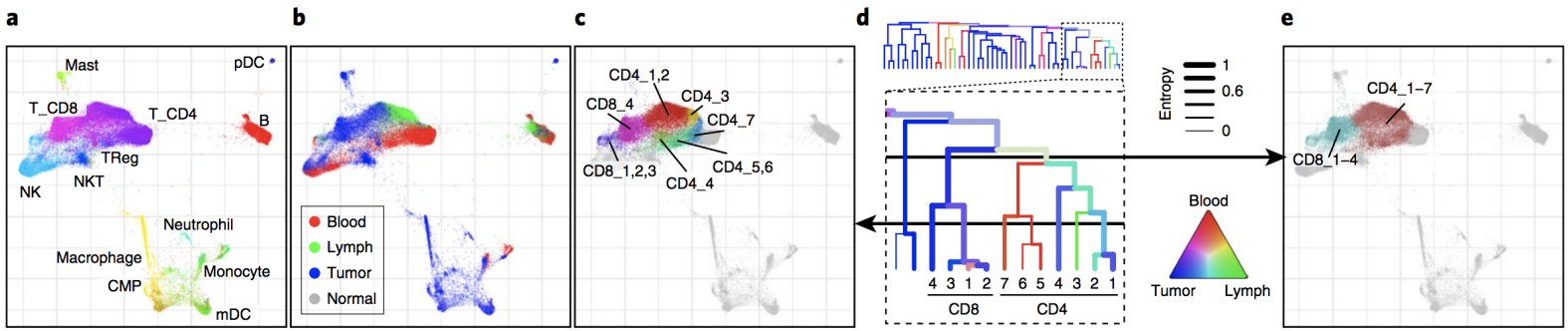

Figure: from Barkas *et al.* Nature Methods. 2019.

Predicting a Cell's Label from the Graph Structure

Label propagation can be used to predict a cell's label based on the labels of neighboring cells.

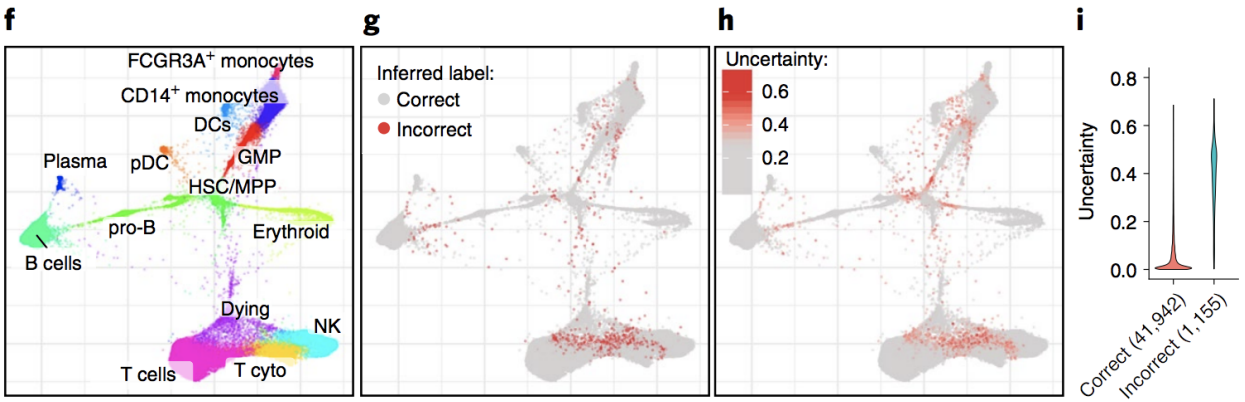

Figure: from Barkas *et al.* Nature Methods. 2019. Cells colored red represent those with incorrect predictions.

Another Approach : Harmony

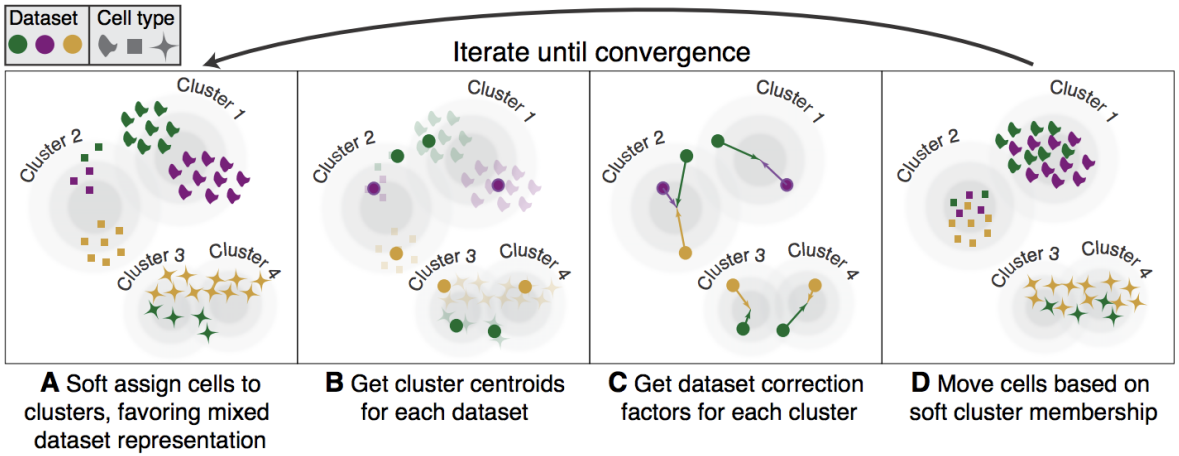

Figure: from `https://www.biorxiv.org/content/10.1101/461954v2.full.pdf`. Harmony uses a soft clustering approach and specifically favors clusters with representative members across all datasets. It also performs a dataset-specific correction per cluster.

Recap

- We just saw a couple of ways that we can combine multiple datasets
- These multiple datasets can correspond to **batches, technologies, or different tissue samples.**
- Conos tries to define a graph such that its structure (in this case, communities) contain cells that are mixed across datasets
- When integrating multiple datasets, we must be careful that cells are grouping together because of phenotype, not because of the function that might have been perturbed under a particular condition.

Another Kind of Dataset Integration Problem, Multiple Panels

Suppose you are trying to analyze a dataset with somewhat overlapping, but unique panels. This could happen for example if different sets of samples were profiled in different labs.

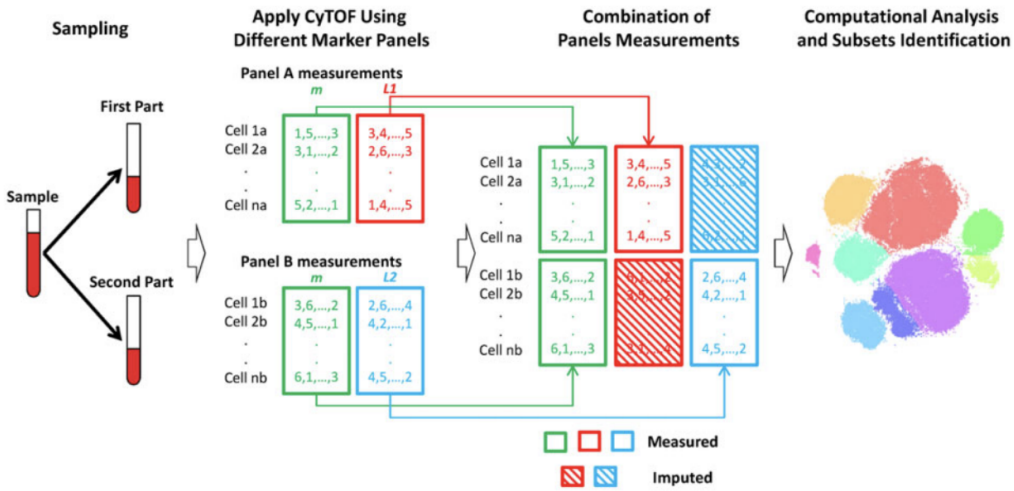

Figure: from Abdelaal *et al.* Bioinformatics. 2019.

Problem Formulation

- Imagine having two panels, where these panels share m markers.
- Assume that panel 1 has $L1$ unique markers and panel 2 has $L2$ unique markers.
- One goal is to determine an informative set of markers that overlap between the two datasets
- Another goal is to impute missing markers.

Imputation

- Starting in panel A, find the k most similar cells in panel B, using the set of m shared markers
- For each cell in panel A, impute missing values for the markers by finding the k most similar cells from panel B and computing missing features as the median of the corresponding features in these k cells.

Computing an Importance Score for Overlapping Markers

Among features overlapping between panels, an importance score for each feature can be calculated to help minimize redundancy across feature measured in panels etc. The importance score for marker p is computed using the first m PCs as,

$$i_p = \sum_{q=1}^{m} \beta_{pq}^2 \times \lambda_q \tag{1}$$

- β_{pq} is the loading of marker p to the qth PC
- λ_q is the variance explained by the qth PC.

You could use the top scored markers either in the design of a new panel or to do your imputation.

Evidence that Imputation is Working for Certain Cell-Populations

	Imputed data
CD4+T cells	0.78
CD8+T cells	0.79
B cells	0.83
CD3–CD7+cells	0.78
TCR$\gamma\delta$ cells	$0.77 \pm 8e{-5}$
Myeloid cells	$0.82 \pm 7e{-5}$
All cells	0.81

Figure: from Abdelaal *et al.* Bioinformatics. 2019.

Recap

- cPCA for comparing to a background.
- Combining multiple datasets
  - Different measured features (CytofMerge)
  - Different tissues, conditions, etc