Comp790-166: Computational Biology

Lecture 6

January 29, 2023

Single Cell Intro Day

- Mass Cytometry (CyTOF) Bioinformatics
- Automating Human Gating with Spade
  - What is a gate?
- Practical Considerations (normalization and transformation)
- Basic featurizing

Do You Remember Questions

- What are the two approaches that node2vec uses to bias the embeddings? By bias the embeddings I mean cause nodes with similar embeddings to mean different things.
- What do you think of embedding nodes with node2vec followed by clustering in the embedding as opposed to just clustering on the original graph?

What is a Single-Cell Assay?

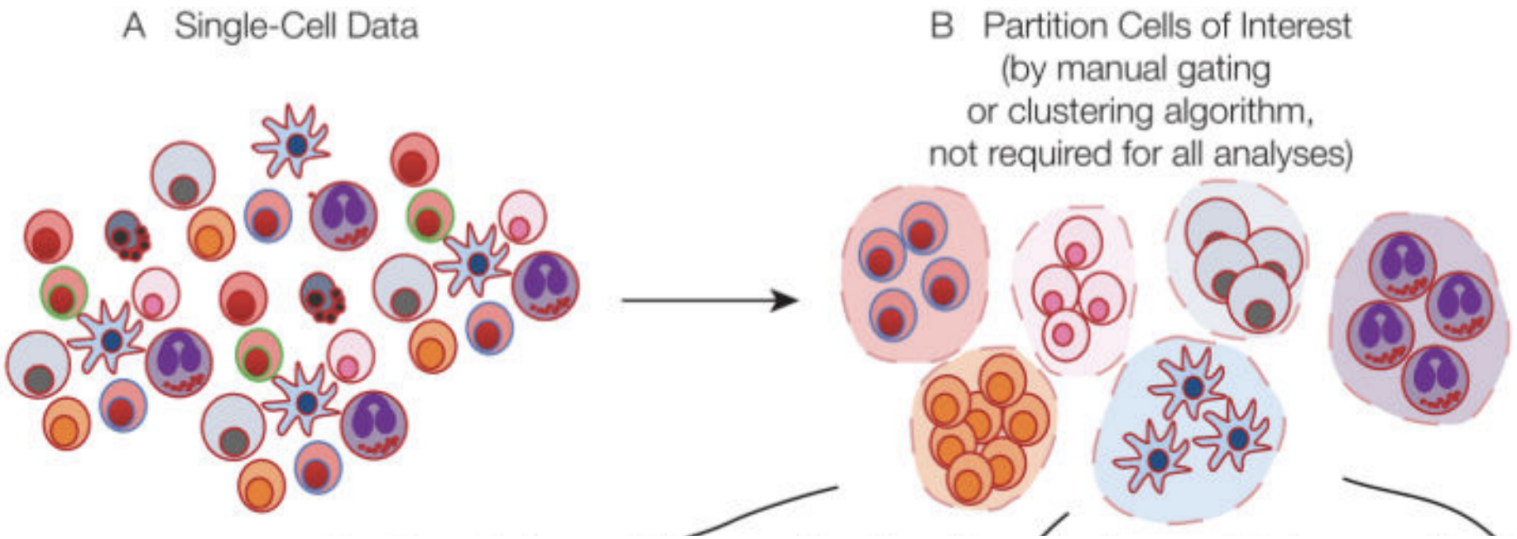

Figure: from Spitzer *et al.* Cell. 2016. In general, we are taking a biological sample (blood, tissue, etc.) and measuring properties of individual cells. We then seek to track and understand entire cell-populations.

Disclaimer

Here we will discuss analysis of single-cell proteomics assays because they are useful for immunophenotyping, or identifying, tracking, and characterizing the diverse immune cell-types.

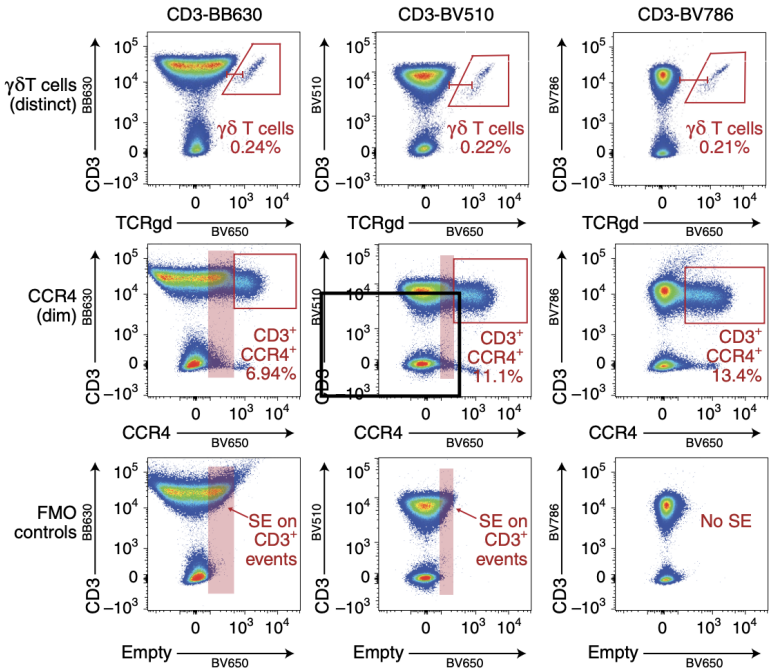

Figure: from Nature Immunology. 2021. We use protein expression in cells to figure out what they are.

Why Do we Care About Immune Profiling?

Example 1: The immune system adapts during pregnancy. We can use it to predict someone's likelihood of preterm birth.

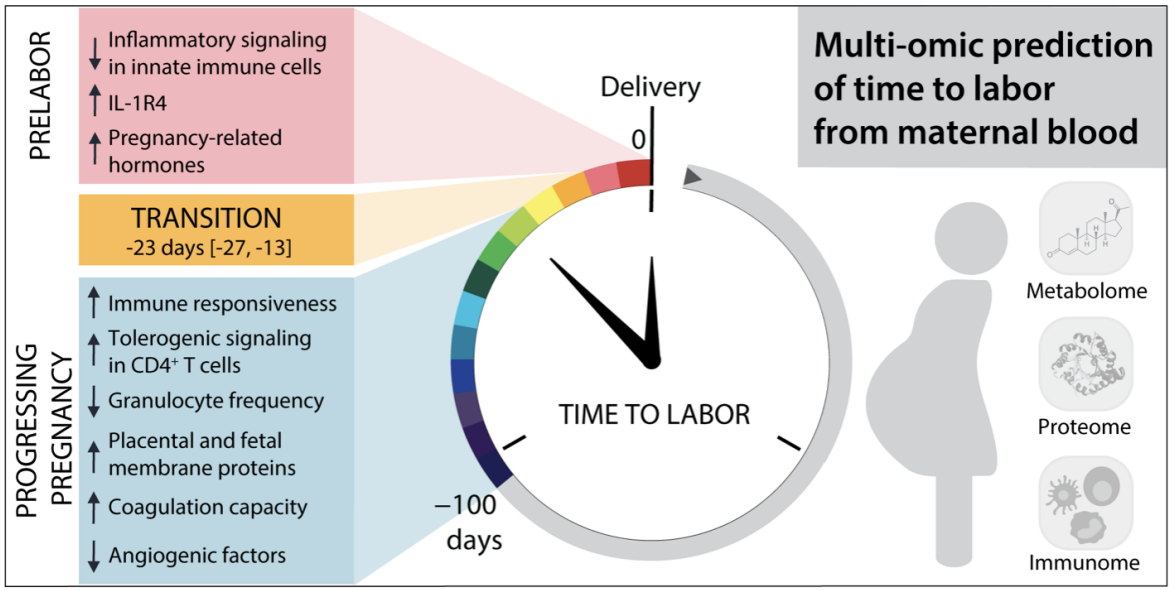

Figure: from Science Translational Medicine. 2021.

SC Protemics: Flow and Mass Cytometry

Flow $\rightarrow$ 18-20 proteins per cell at a rate of 10,000 cells per second!

Mass $\rightarrow$ 36-45 proteins at a rate of 1,000 cells per second

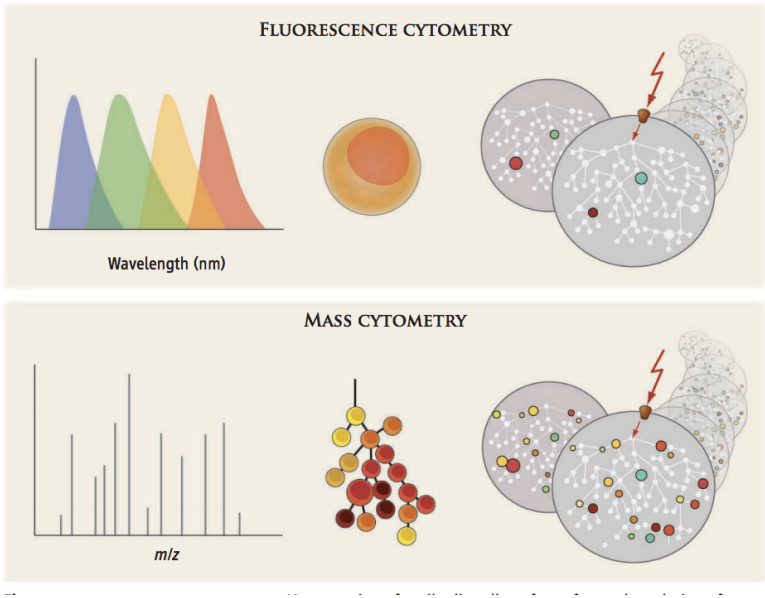

Figure: from Benoist and and Hacohen. Science. 2011

Cytometry by Time of Flight (CyTOF)

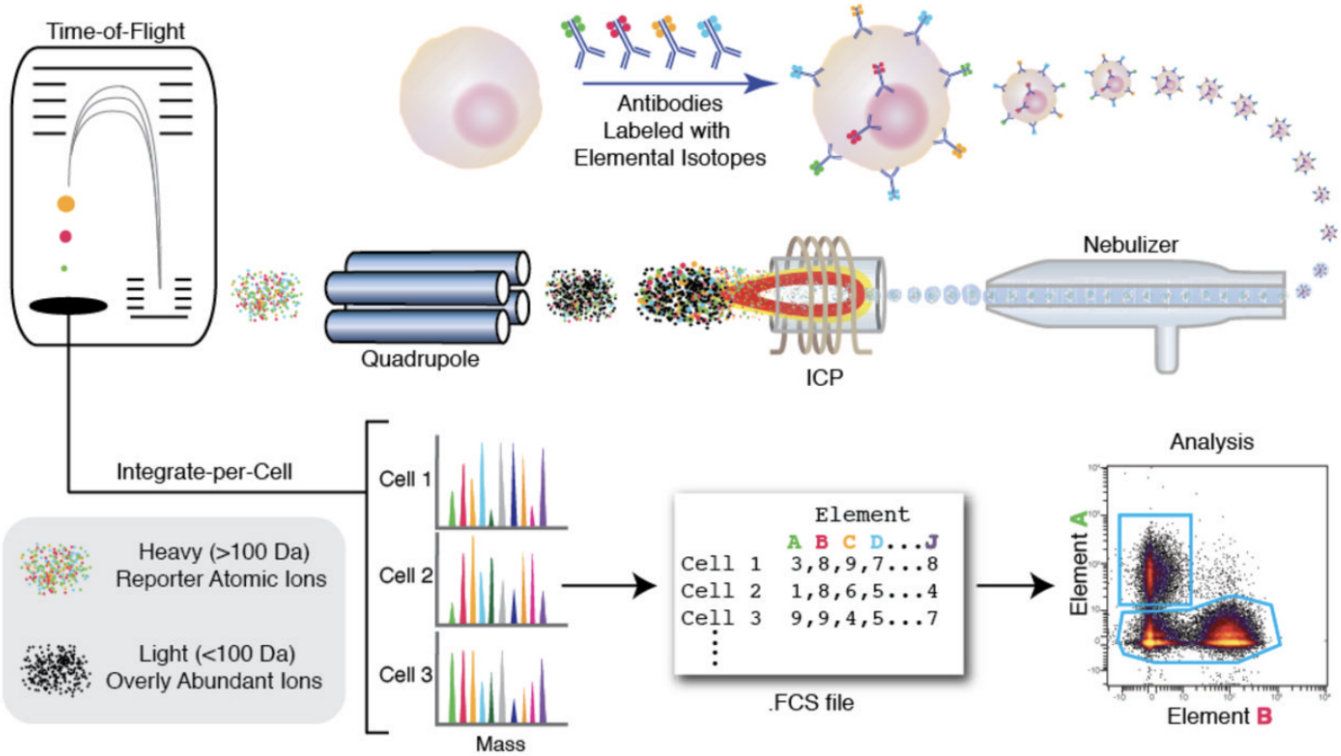

Figure: from Bendall *et al.* Trends in Immunology. 2012

CyTOF : the specific technology for mass cytometry

Manual Gating

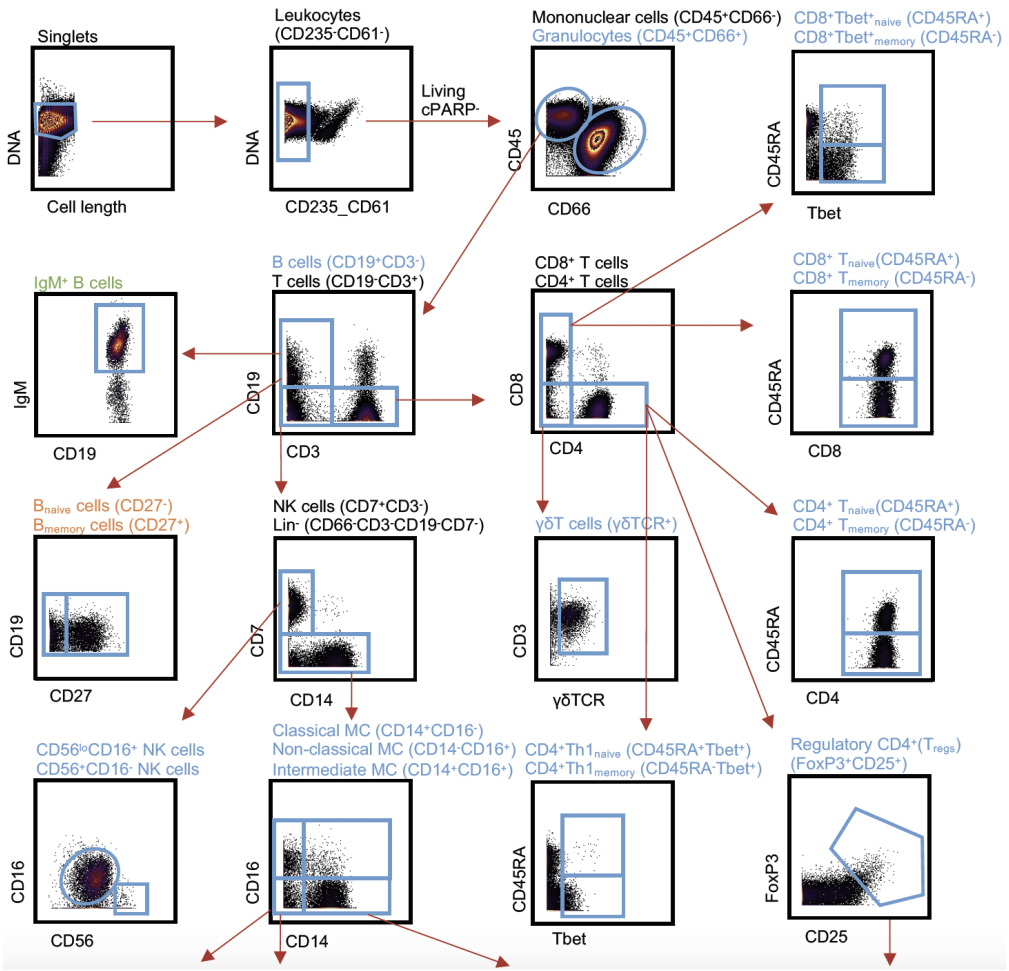

Figure: Nature Communications. 2020

Problems with Manual Gating

- Requires a human expert with extensive knowledge about how to characterize cells
- Time consuming. If you measure 36 proteins, then you have 36 choose two scatterplots. Not all combinations of markers is meaningful and the gating is hierarchical in nature.
- Biased towards characterizing cell-populations that have already been well described (e.g. you only find what you are looking for).
- A coming attraction : fully automating this analysis!

I have my manually gated cell-populations, now what?

We look at particular immune cell-types to understand how the immune system is responding in particular circumstances (for example, pre-term birth).

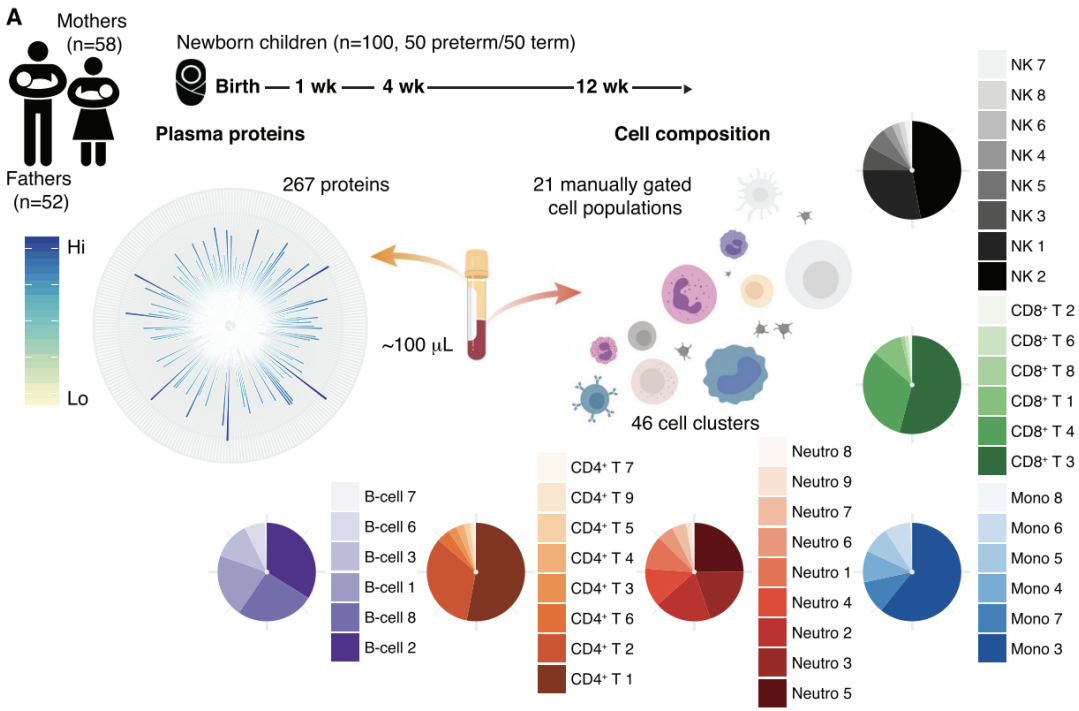

Figure: from Olin *et al.* Cell. 2018

What Properties of Cell-Populations do we Study?

You can count cells in each population (known as frequency, or fq), or characterize signaling activity across multiple stimulations (e.g. experimental perturbations).

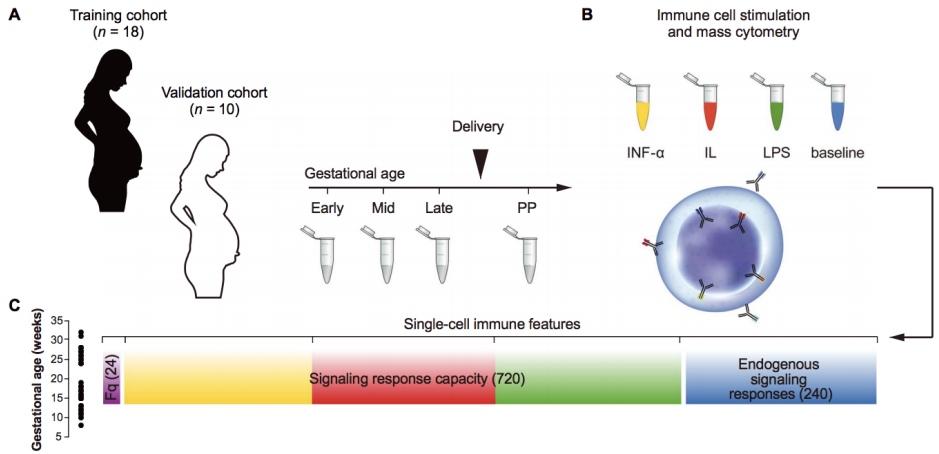

Figure: from Aghaeepour *et al.* Science Immunology. 2017

Extracting Features for Prediction Tasks

- Start with a matrix of cells $\times$ protein expression for each patient sample
- Define populations through gating
- Compute function or frequency based features for each population

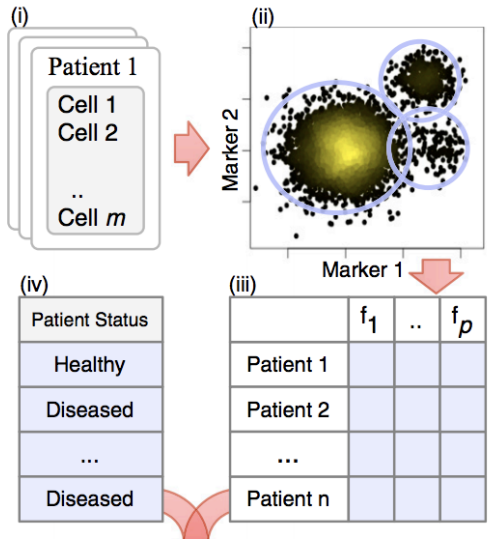

Figure: from Bruggner *et al.* PNAS. 2014

What Are Cell Frequencies?

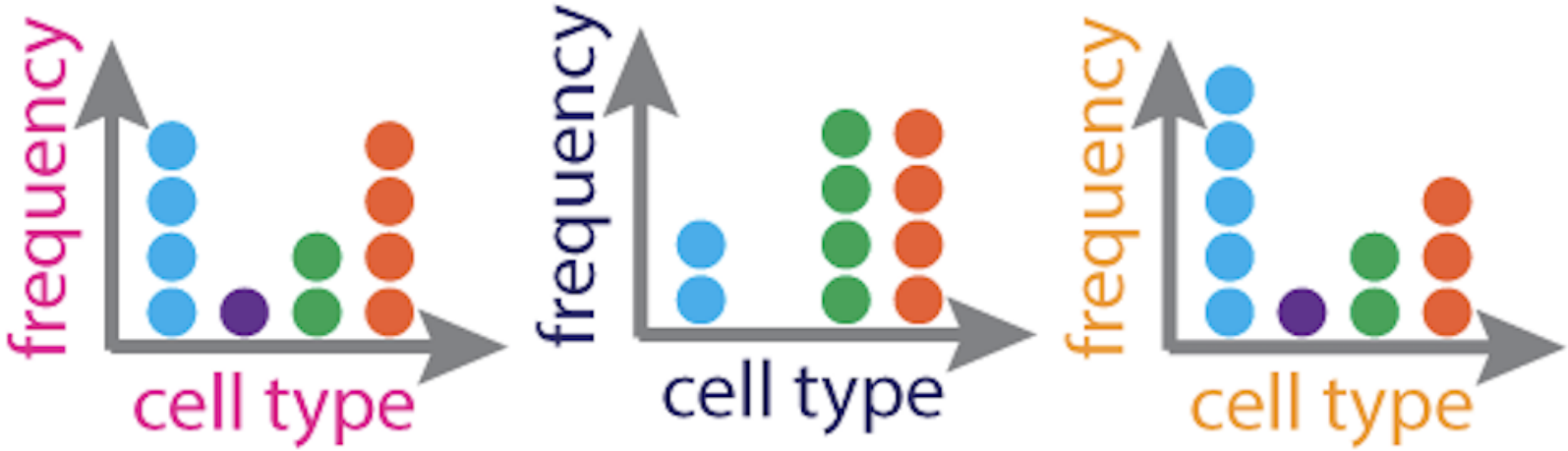

Figure: Simply count the number of cells in each profiled sample assigned to each specific type. This is a very simple hand engineered feature.

Practical Concerns. Converting data profiled with CyTOF into a matrix that can be used to understand a biological question.

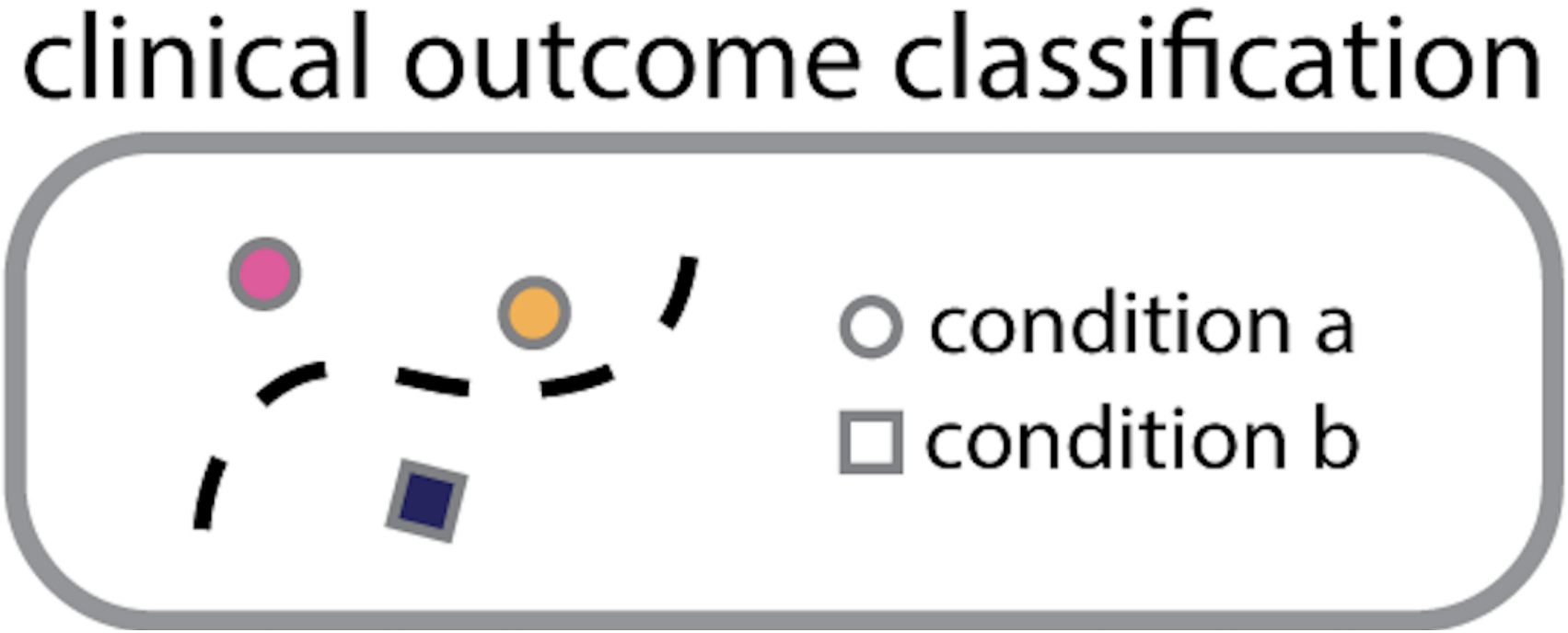

Figure: Each point is representing a different patient sample. How do we ultimately specify a set of features for each sample that can be used to classify them correctly?

FCS file

- The collection of cells for each sample are stored in an ‘FCS‘ file. Cells are also called events.
- FCS stands for ‘Flow Cytometry Standard‘ and contains information about the experiment, like human interpretable names for markers, channel names, information about any performed normalization, etc.
- You will need to process each individual sample file.

Dealing with These Data

You can find a lot of publicly available multisample flow and mass cytometry datasets on Flow Repository (`http://flowrepository.org/`)

	File	Size	MD5sum	Type
☑	Gates_PTLG021_IFNa_LPS_Control_1.fcs	27.9 MB	53e4367f87 ...	FCS 3.0
☑	Gates_PTLG021_IFNa_LPS_Control_2.fcs	37.5 MB	7aab21b8cb ...	FCS 3.0
☑	Gates_PTLG021_Unstim_Control_1.fcs	30.9 MB	b36fdfbf05 ...	FCS 3.0
☑	Gates_PTLG021_Unstim_Control_2.fcs	26 MB	42676ee1b3 ...	FCS 3.0
☑	Gates_PTLG025_IFNa_LPS_Control_1.fcs	26.7 MB	4cc5b6c07c ...	FCS 3.0
☑	Gates_PTLG025_IFNa_LPS_Control_2.fcs	35.7 MB	f87c5e6356 ...	FCS 3.0
☑	Gates_PTLG025_Unstim_Control_1.fcs	29.5 MB	1a3d2d9448 ...	FCS 3.0
☑	Gates_PTLG025_Unstim_Control_2.fcs	28.9 MB	d8b3989dfb ...	FCS 3.0
☑	Gates_PTLG026_IFNa_LPS_Control_1.fcs	32.6 MB	17ab8c9b98 ...	FCS 3.0
☑	Gates_PTLG026_IFNa_LPS_Control_2.fcs	37.9 MB	b370371ed9 ...	FCS 3.0
☑	Gates_PTLG026_Unstim_Control_1.fcs	36 MB	0f9c4dfd2d ...	FCS 3.0
☑	Gates_PTLG026_Unstim_Control_2.fcs	39.8 MB	4687536e38 ...	FCS 3.0
☑	Gates_PTLG027_IFNa_LPS_Control_1.fcs	26.8 MB	787f4a6366 ...	FCS 3.0
☑	Gates_PTLG027_IFNa_LPS_Control_2.fcs	35.3 MB	3b2ab7e2a2 ...	FCS 3.0

Figure: One FCS file per sample.

Step 1: Separating Markers into Function vs Phenotypic vs Experimental

- There are 3 types of columns in an FCS file
- Phenotypic marker columns help us to characterize particular cell-populations (hint, usually starts with **CD**).
- Functional marker columns help us to quantify signaling or other function within a cell
- 'Junk' markers- help to keep track of which cells died, etc.
- After filtering out dead cells, we will only do analysis based on functional or phenotypic markers.

First Step of Manual Gating : Extract Live Cells

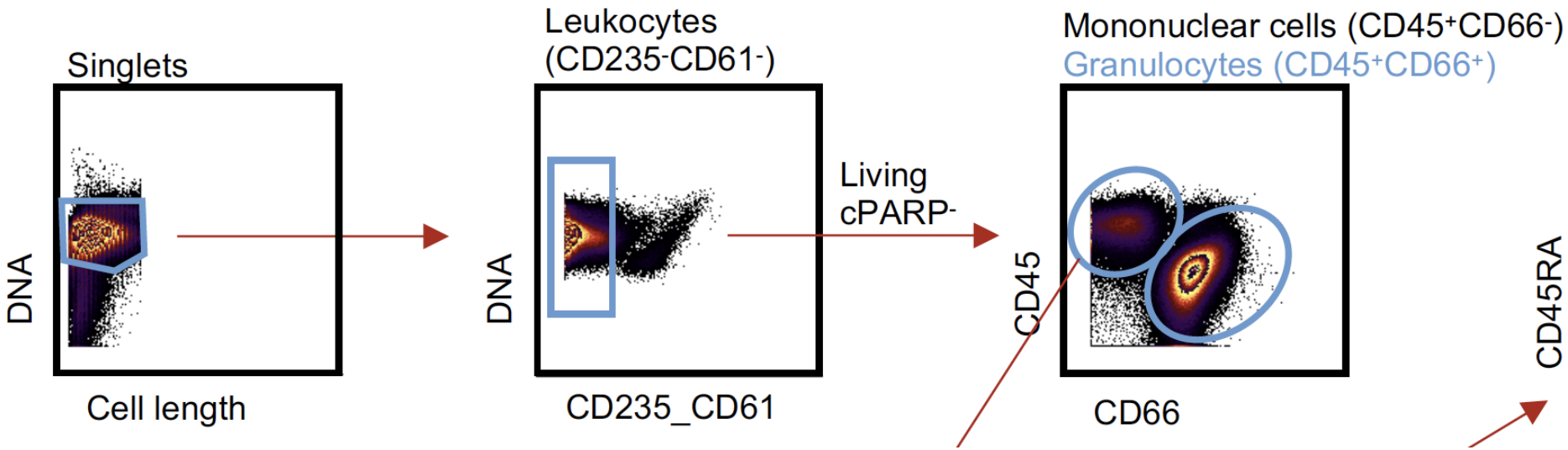

Figure: from Stanley *et al.* Nature Communications. 2020

Example Marker Names from an FCS File

```
>>> sample.pnn_labels
['Time', 'Cell_length', 'DNA1', 'DNA2', 'CD45RA', 'CD133', 'CD19', 'CD22', 'CD11b', 'CD4', 'CD8', 'CD34', 'Flt3', 'CD20', 'CX
, 'CD45', 'CD123', 'CD321', 'CD14', 'CD33', 'CD47', 'CD11c', 'CD7', 'CD15', 'CD16', 'CD44', 'CD38', 'CD13', 'CD3', 'CD61', 'C
 'HLA-DR', 'CD64', 'CD41', 'Viability', 'file_number', 'event_number', 'label', 'individual']
```

Figure: For example, DNA1 and DNA2 help us to find dead cells. Markers like 'CD19' help us to find specific cell-types. 'CD19' in particular characterized B-cells! Event number is a junk marker example, which is just counting the cells.

Step 2: Transform Marker Expressions

In the cell $\times$ marker matrix, we are counting the number of detected ions or joins between protein and heavy-metal tagged antibody. We will use a transformation, to compress the upper end of the spectrum and enhance the lower end.

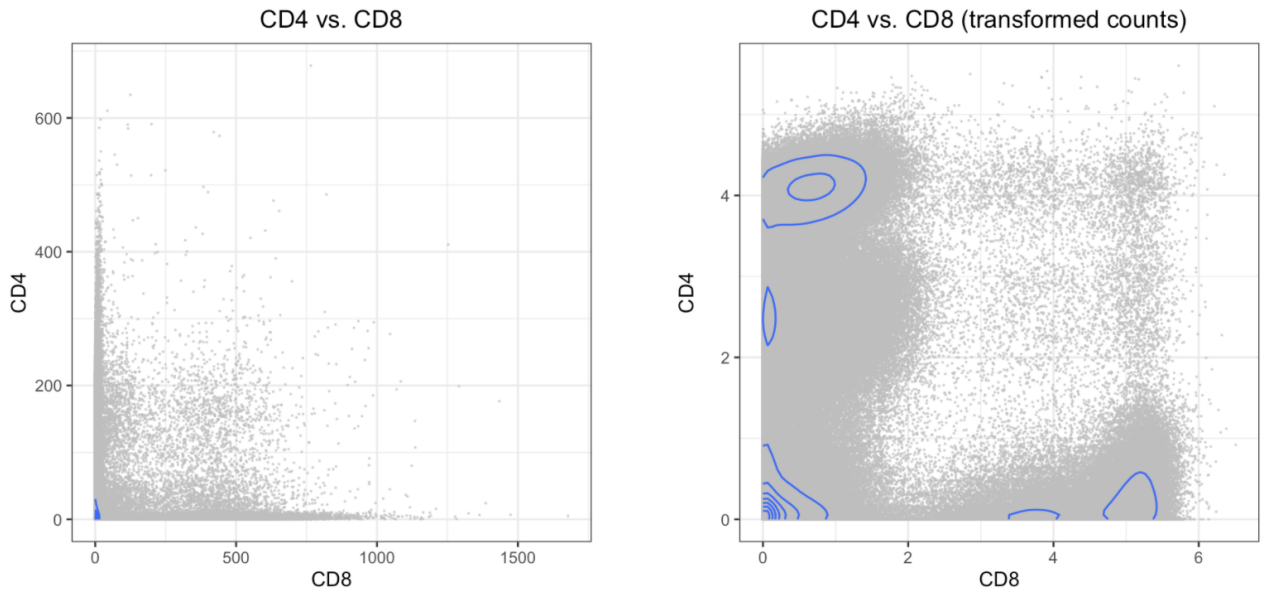

Figure: from `https://biosurf.org/cytof_data_scientist.html`

Arcsinh Transformation

In practice, especially with mass cytometry data, it is common practice to use an Arcsinh transformation, with co-factor aka scaling factor of 5. For count x, the transformed value $x' = \text{asinh}(\frac{1}{5}x)$

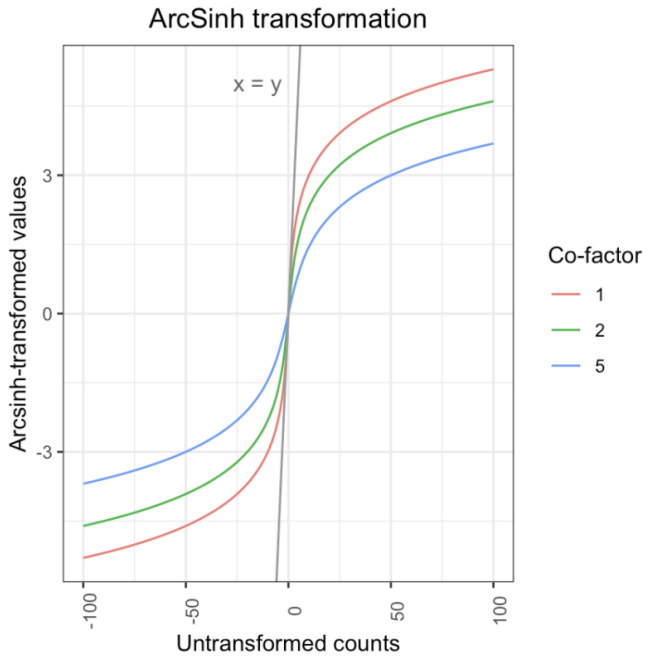

Figure: from `https://biosurf.org/cytof_data_scientist.html`

Effect of Normalization on Gating

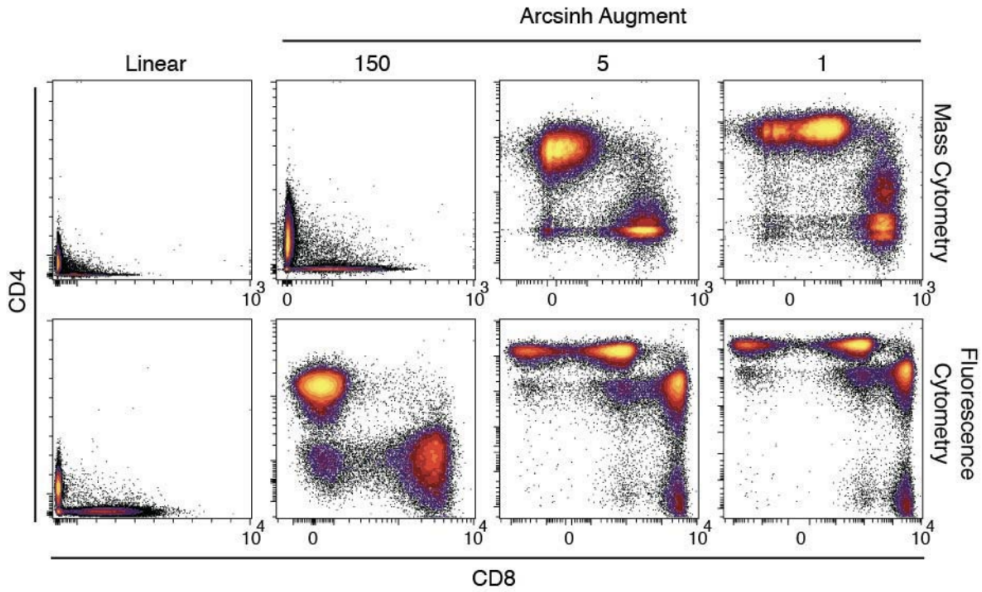

Figure: from Bendall *et al.* Science. 2011. The cofactor can cause sensitivity in the number of populations that emerge.

Tutorial

I have a tutorial for converting FCS files into matrices for both Python and R. `https://github.com/stanleyn/fcs_tutorial`.

Recap

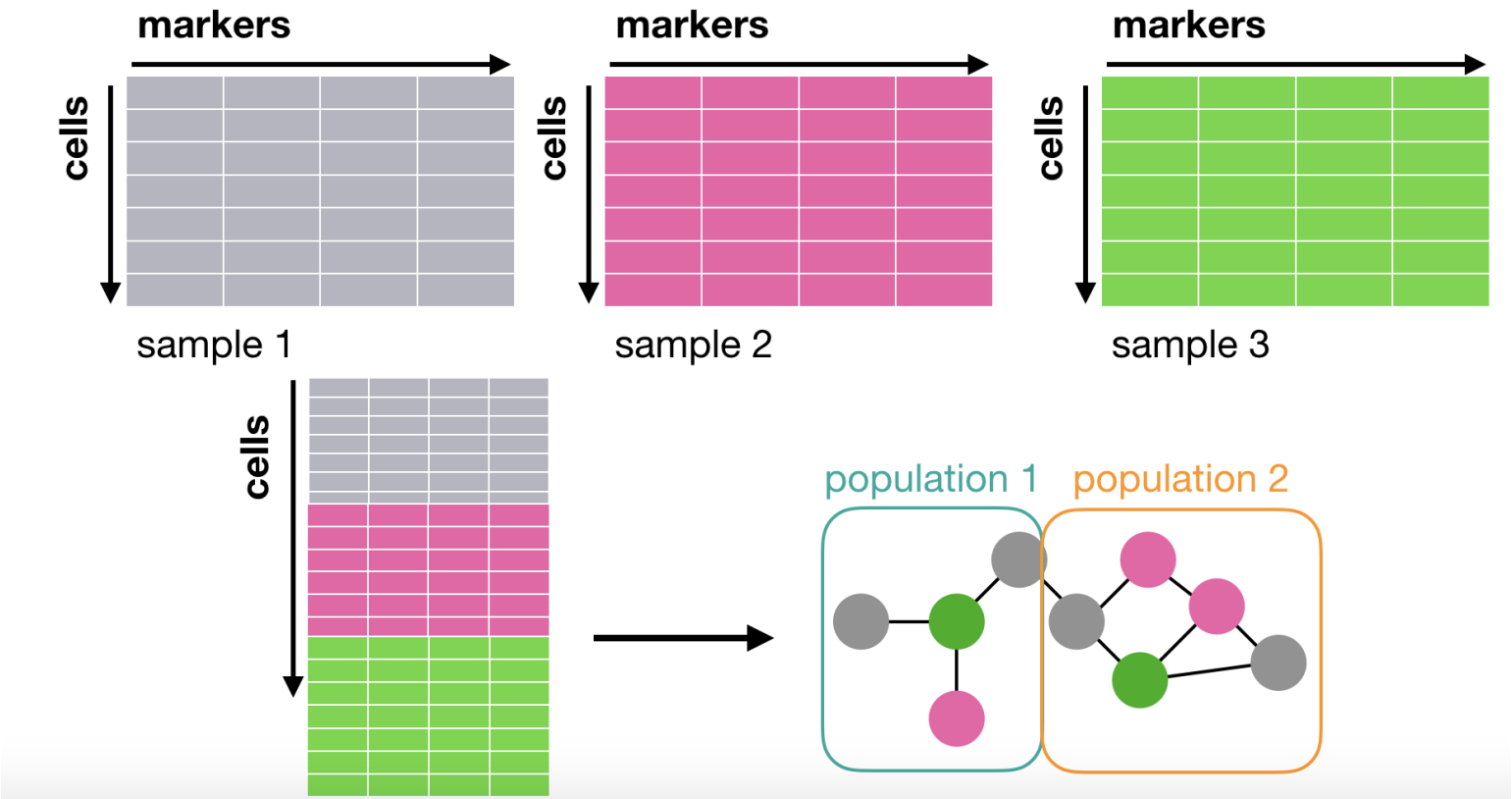

Figure: We are processing each FCS file, which corresponds to an individual sample. Ultimately cells are pooled across all samples.

To Automate Gating or to Not Automate Gating

Ask your collaborator to give you FCS files for each sample with only live cells included. If you ask them for gates, it will take them a lot of work. But you can make your own gates through unsupervised clustering!

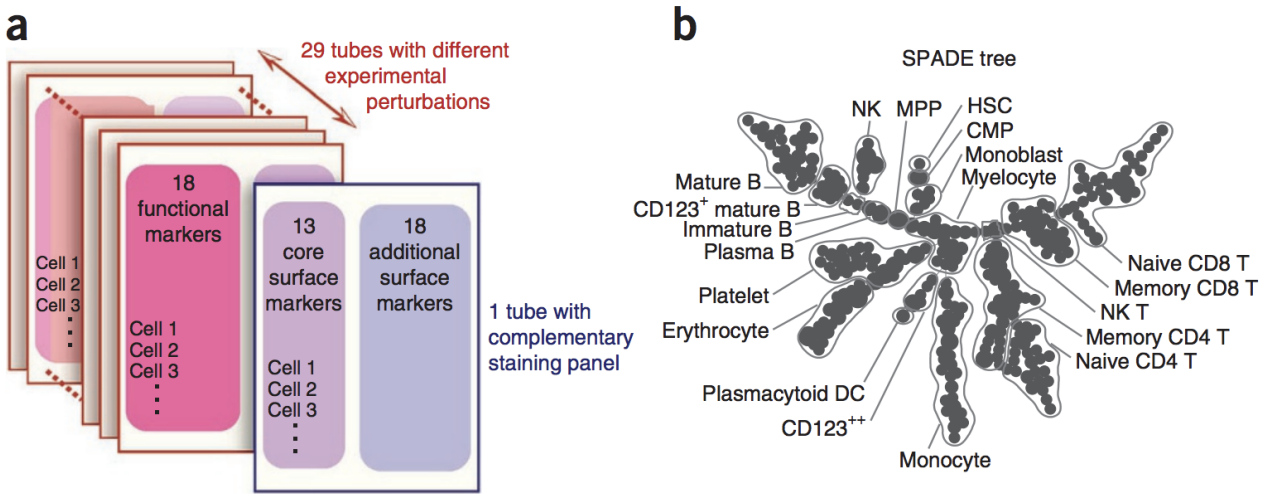

Figure: from Qiu *et al.* Nature Biotechnology. 2011. Do agglomerative hierarchical clustering based on the expression of measured markers.

Why Hierarchical Clustering?

It recapitulates our general understanding of cellular differentiation.

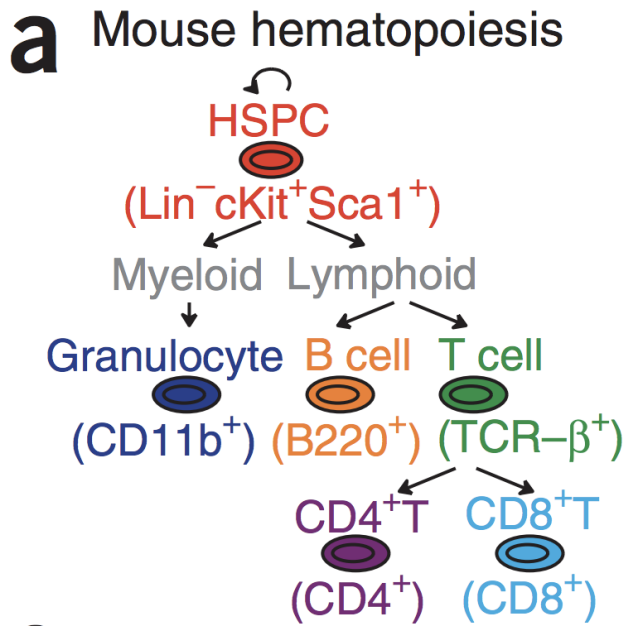

Figure: from Qiu *et al.* Nature Biotechnology. 2011. Cells can differentiate from Stem Cells to more specialized cell-types.

A Practical Consideration

- For N cells (and recall that N is large, like $>$100K), we need to compute an $N \times N$ distance matrix.
- If we have 100 samples, we will have around 100K $\times$ 100 total cells.
- We can't calculate all of those distances! So, right now, hierarchical clustering fails...

Question for You

Do you have any ideas about how you might reduce the size of this problem of the analysis of such a large number of cells across multiple samples?

Density-Dependent Downsampling

- Downsampling is a popular approach where a limited number of samples is samples for each FCS file.
- This is not ideal, because downsampling causes information loss.
- Spade does this is a density-dependent way, by sampling subsets of cells, and computing densities of their neighborhoods. Hence, Spade ensures that cells are represented across neighborhoods, especially in sparse neighborhoods.

Visualization in Spade : Cluster-Level

Construct a minimum spanning tree between clusters, based on the median expression of the markers.

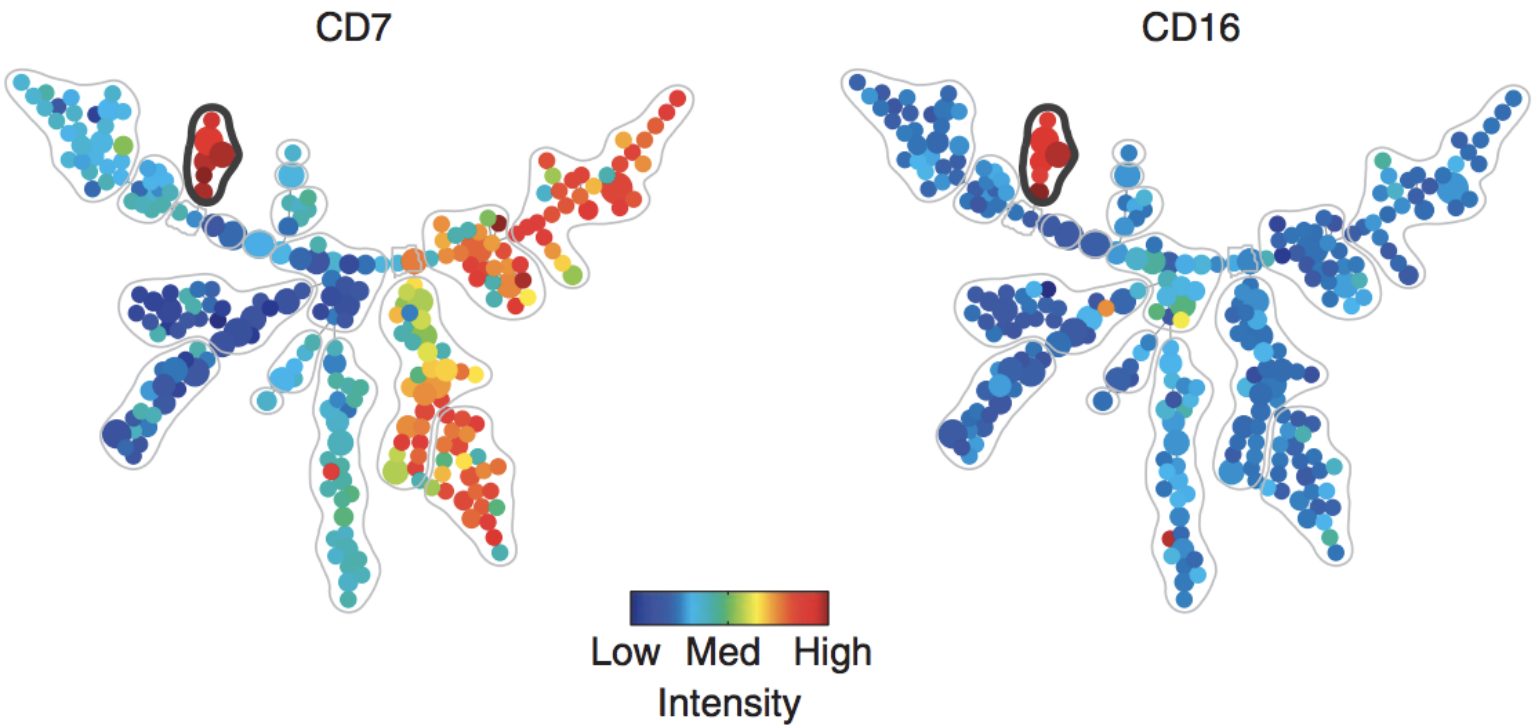

Figure: from Qiu *et al.* Nature Biotechnology. 2011.

Warning : Batch Effects

As with every biological assay, there is technical variation. For example, half the samples run on a machine in California, and the other half run on a machine in North Carolina. NC and CA samples might look very different from each other.

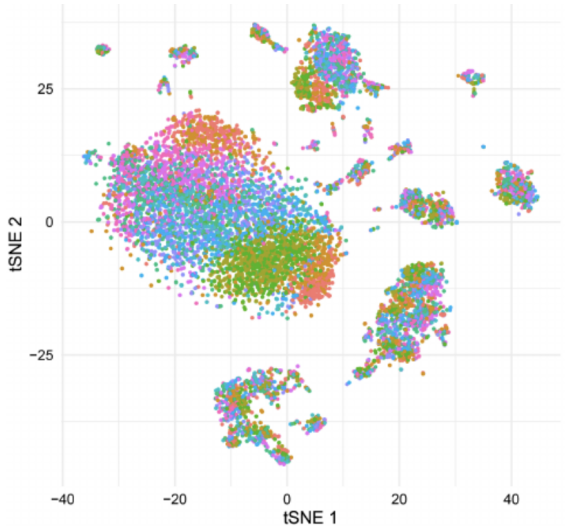

Figure: from Van Gassen *et al.* Cytometry A. 2019. Here cells are colored by batch.

Your job as a collaborator

- Batch effects are the most problematic when they correlate with the patient label that you are hoping to predict.
- If you could perfectly separate patients, it would be difficult to know if the success was because of a batch effect, or true biological signal.
- As a collaborator, you need to suggest for as much randomization as possible! E.g. have batches that have healthy and sick, for example

All of the Graph Knowledge Finally Applied....

We saw Spade last time. The purpose of PhenoGraph is also to define clusters of cells that recapitulates manual gating results

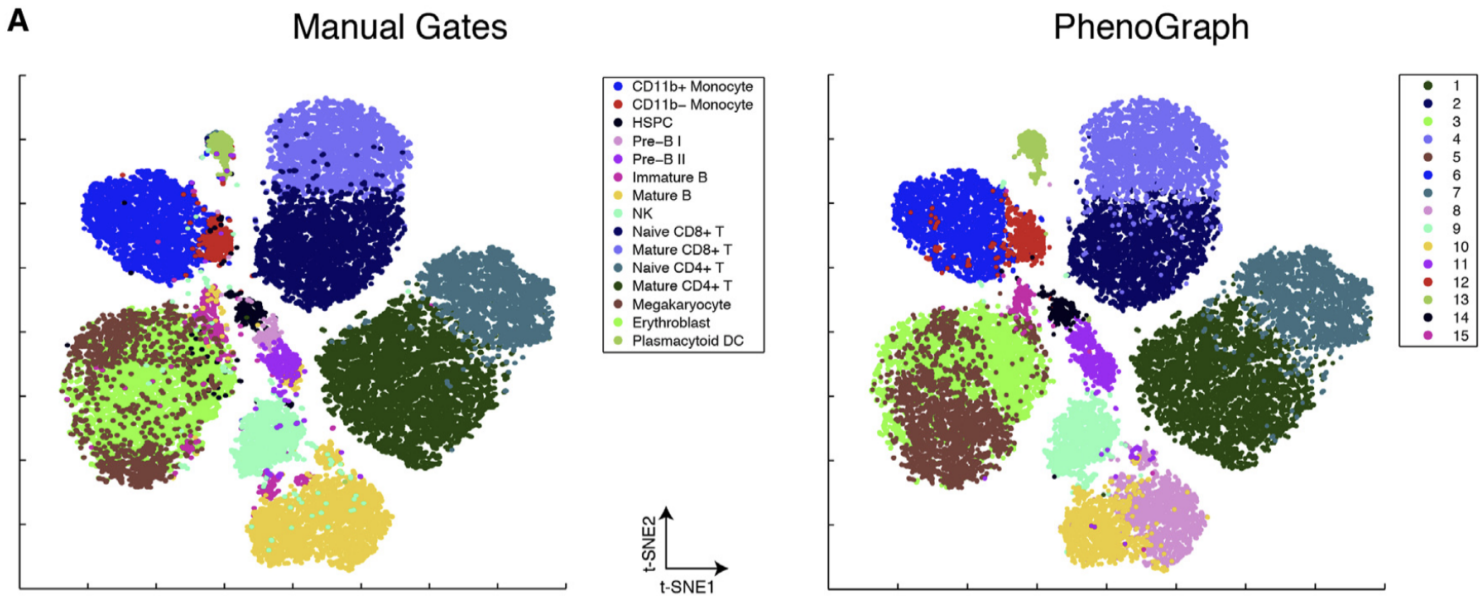

Figure: from Levine *et al.* Cell. 2015

A Direct Application of Louvain

- PhenoGraph is sold as being able to capture cell-populations of various sizes.
- A graph of cells is constructed for each samples
- Edges that exist are then turned into weighted edges, based on shared neighbors (a good contribution)
- The graph for each sample is partitioned with Louvain
- Clusters between samples are mapped with a metaclustering approach.

Communities Map to Cell-Populations

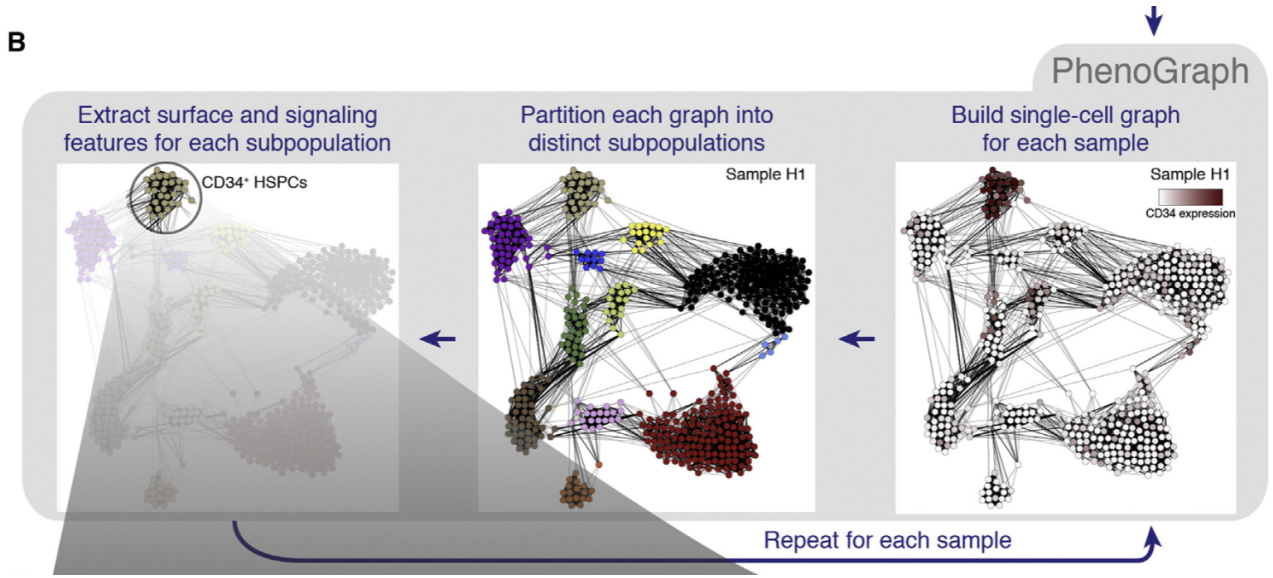

Figure: from Levine *et al.* Cell. 2015

Refining the kNN graph

Suppose an edge exists betweens nodes i and j. Then their edgeweight, w_{ij} is defined as the Jaccard Score of shared neighbors as,

$$W_{ij} = \frac{|v(i) \cap v(j)|}{|v(i) \cup v(j)|} \tag{1}$$

Here, $v(i)$ is the set of neighbors of node i. $|\cdot|$ is cardinality (number of neighbors)

Evaluating Similarity to Manual Gates

It's a bit reassuring to know that the performance (in terms of F-measure) is stable, regardless of the choice of k. [1]

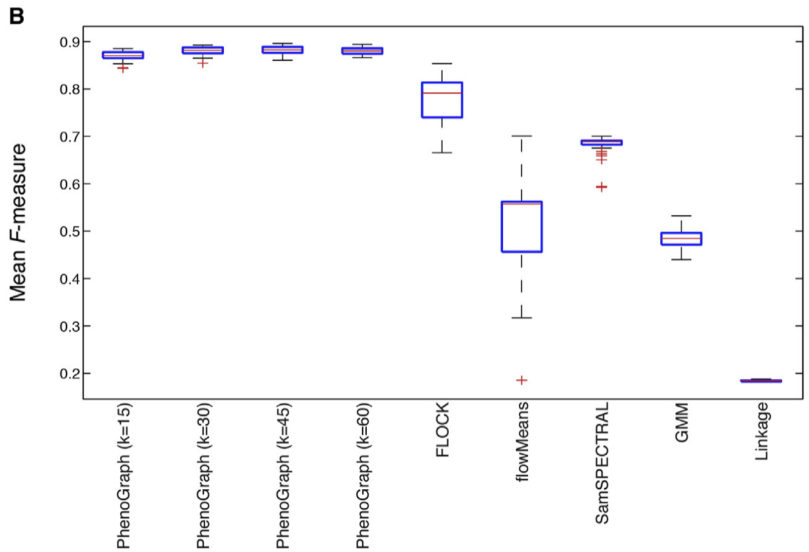

Figure: from Levine *et al.* Cell. 2015

[1] discussion of F measure `https://www.nature.com/articles/nmeth.2365.pdf`

Healthy vs AML with Cell Frequencies

Calculate the proportion of each sample's cells assigned to a particular metacluster. As you can see, there are differences in frequency between healthy (first few rows) and AML.

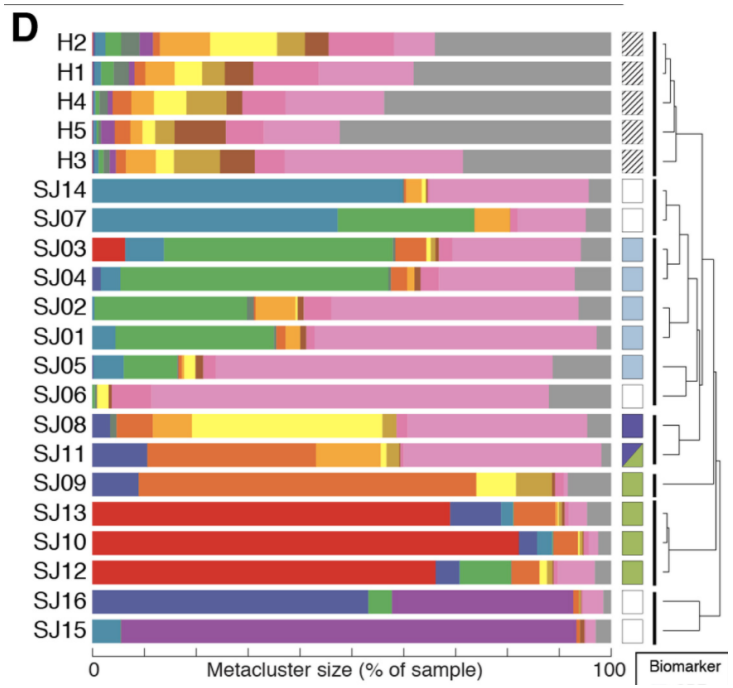

Figure: from Levine *et al.* Cell. 2015

You are ready for your first CyTOF dataset

If the data have signal, cell frequencies are a pretty powerful biologically interpretable feature that can allow us to classify patient samples.
To recap,

- Define per sample clusters
- Define metaclusters representing all samples
- Map cells from individual cells to metaclusters
- Compute frequency features
- You have now build a feature matrix that can be used for classification!

Example of a Powerful Frequency Feature

Frequency differences can be quite prominent in clinical settings. In this example, we are studying differences between patients who received steroid treatment after surgery, versus not.

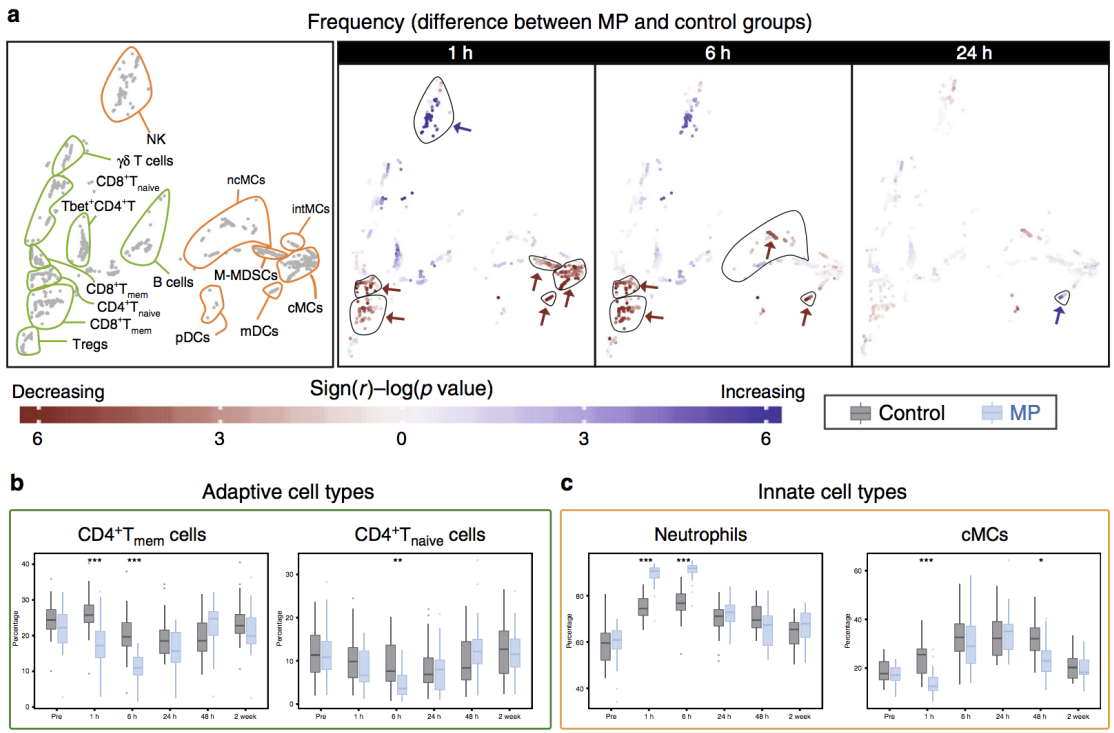

Figure: from Nature Communications. 2019.