Comp790-166: Computational Biology

Lecture 5

February 2, 2021

Announcements

- Slack for project discussions and general discussions. Check your email for an invite link.
- Homework 1 assigned by Feb 4.
- Required reading summaries reduced in frequency from every week to a few selected weeks in the semester.

Single Cell Intro Day

- What are single-cell assays?
- Mass Cytometry Bioinformatics
- Automating Human Gating with Spade
- Practical Considerations (normalization, batch effects)

What is a Single-Cell Assay?

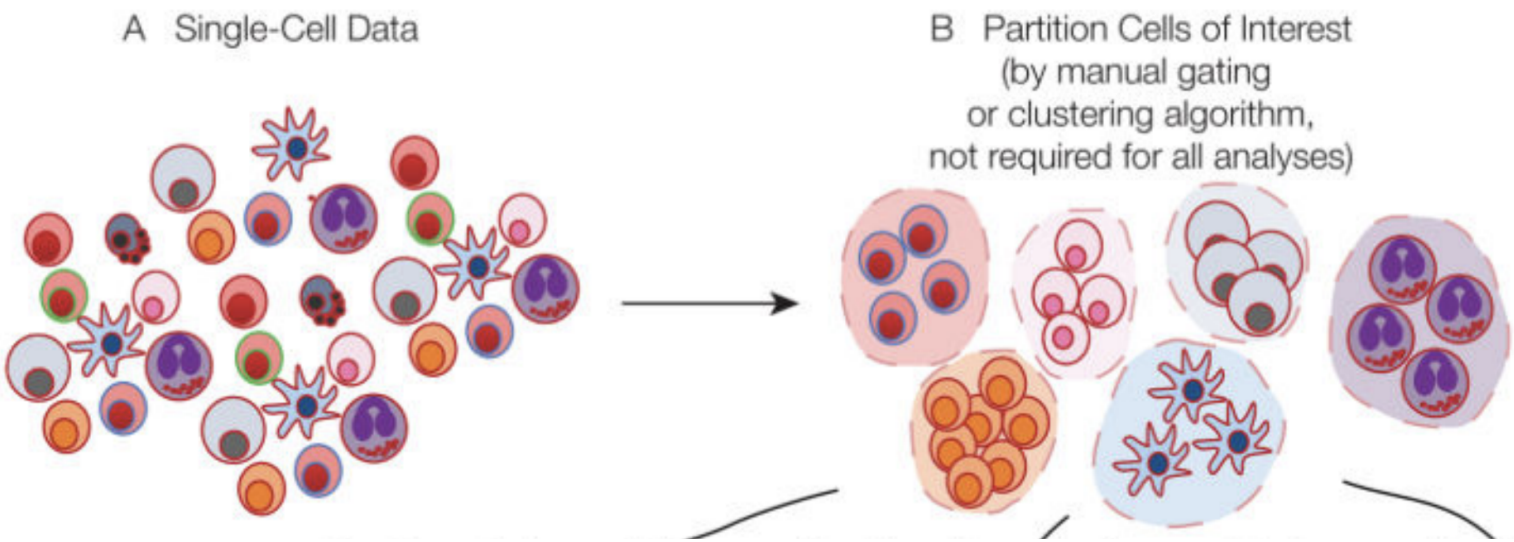

Figure: from Spitzer *et al.* Cell. 2016. In general, we are taking a biological sample (blood, tissue, etc.) and measuring properties of individual cells. We then seek to track and understand entire cell-populations.

What Level of Biology do You Want to Measure?

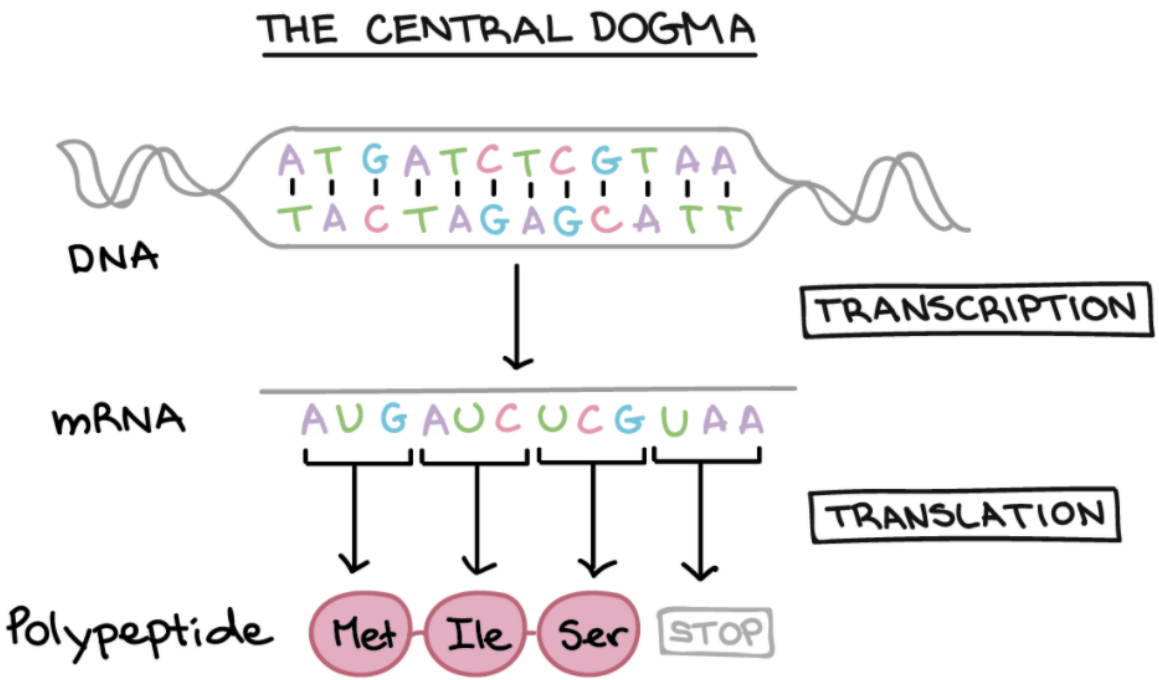

Figure: from khanacademy.org. Remember, DNA $\rightarrow$ RNA $\rightarrow$ protein

Disclaimer

I am a protein snob. My own research is in single-cell proteomics assays, hence this content will comprise the majority of the discussion. Techniques to analyze single-cell gene and protein expression assays are still quite similar and equally important.

scRNA seq for measuring gene expression

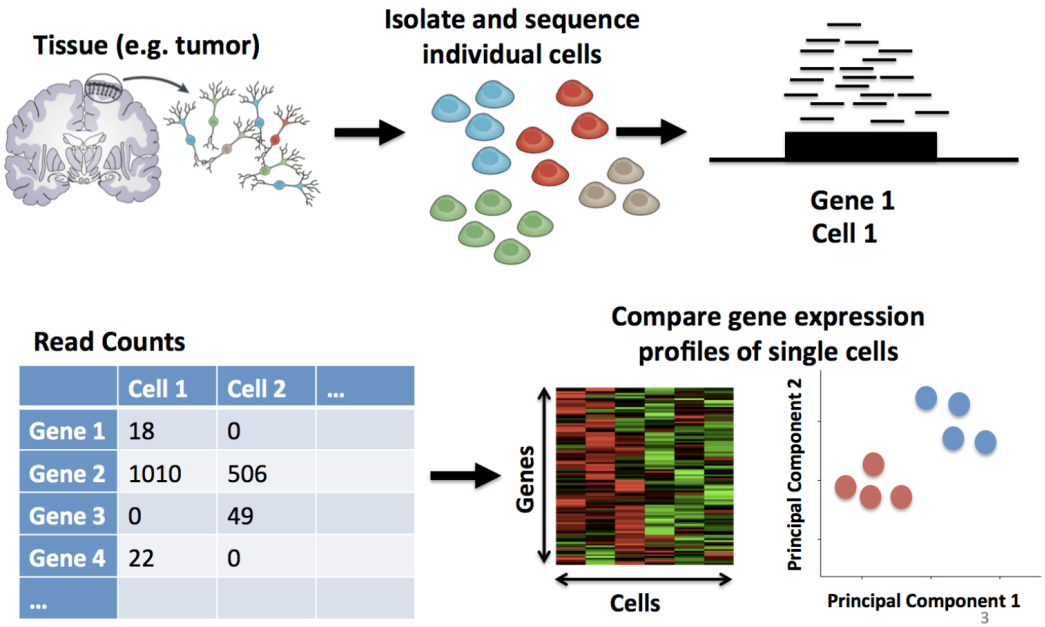

Figure: from `https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/`. Measure the expression of 10s of thousands of genes in thousands of cells.

SC Protemics: Flow and Mass Cytometry

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

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

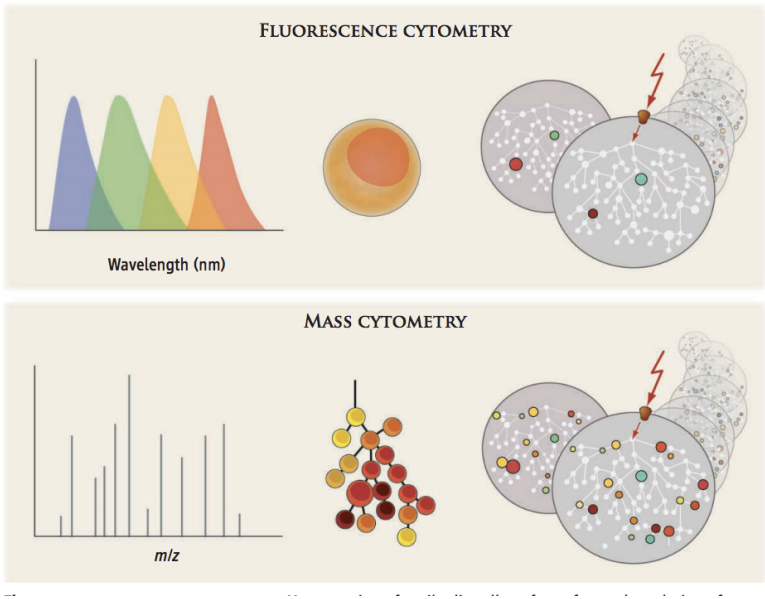

Figure: from Benoist and and Hacohen. Science. 2011

Mass Cytometry

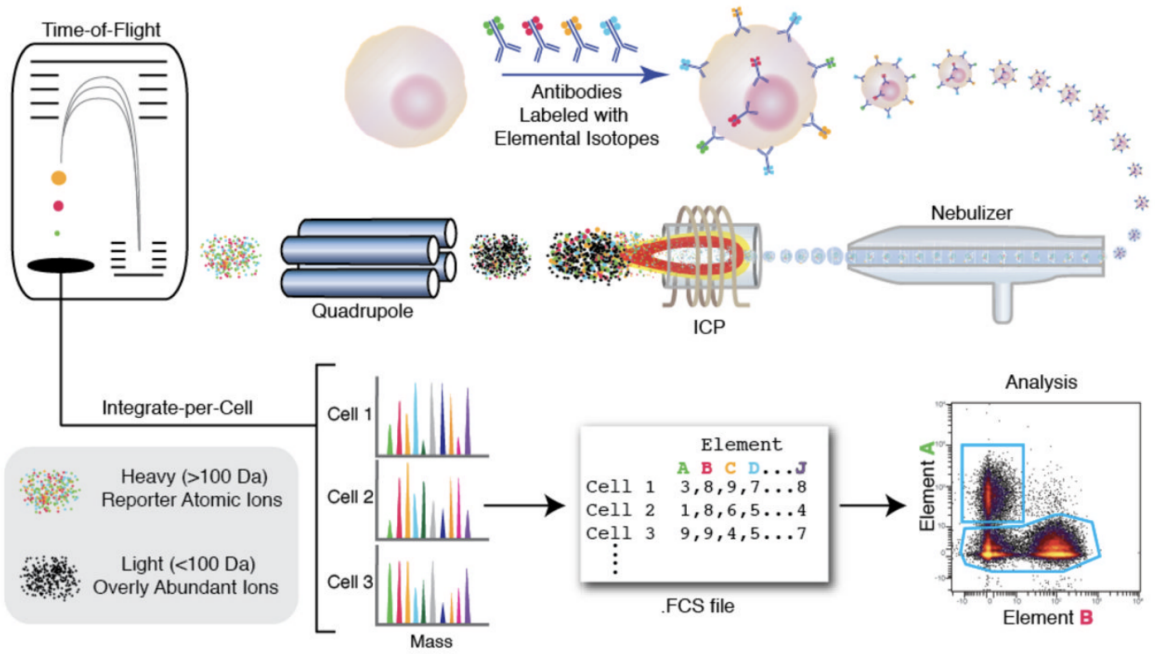

Figure: from Bendall *et al.* Trends in Immunology. 2012

CyTOF : the specific technology for mass cytometry

Manual Gating

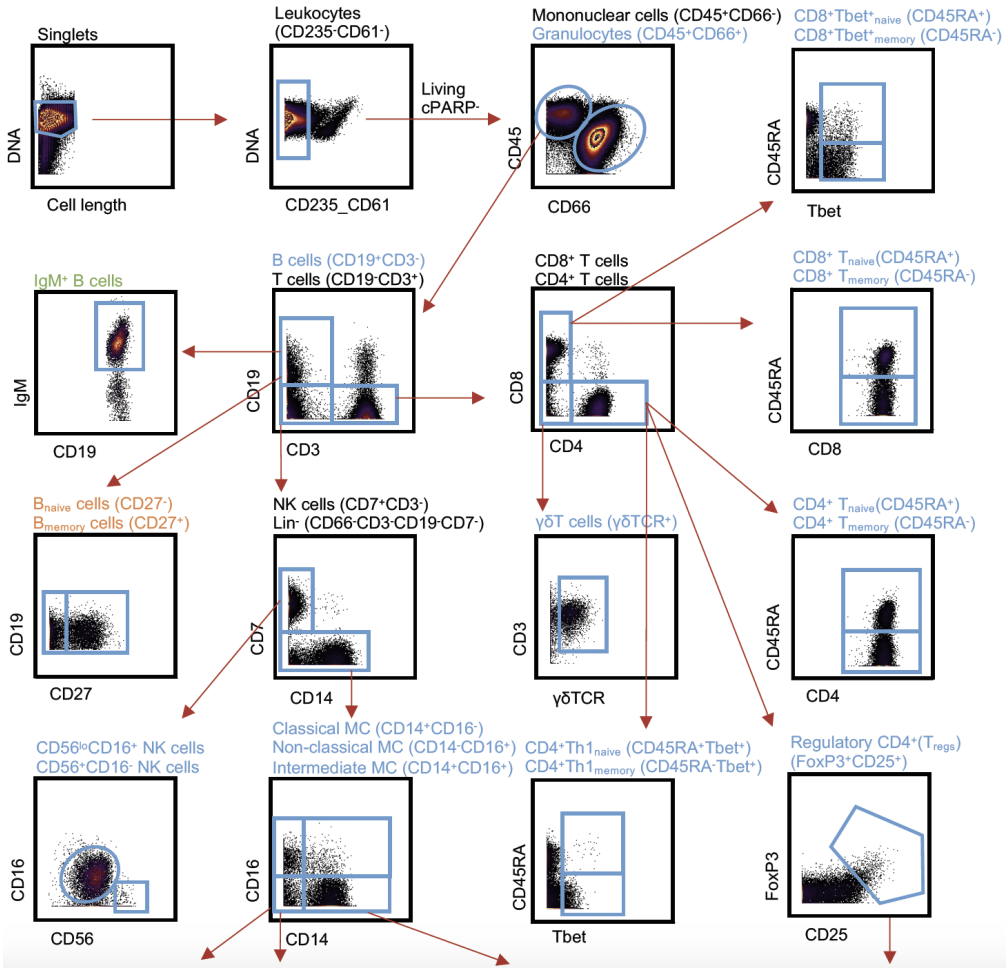

Figure: from Stanley *et al.* Nature Communications. 2020

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).

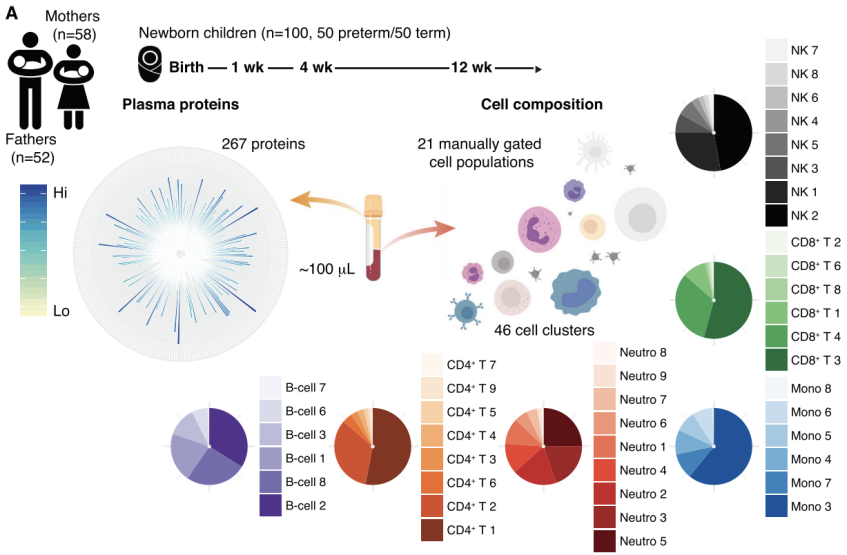

Figure: from Olin *et al.* Cell. 2018

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).

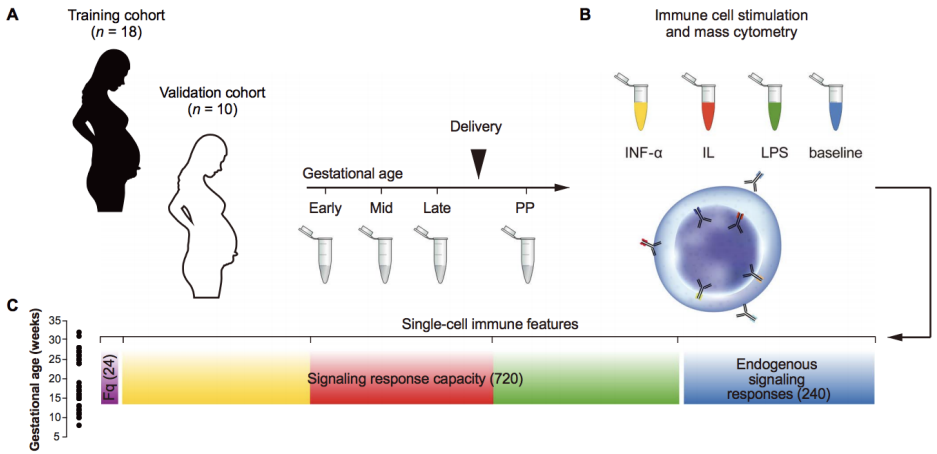

Figure: from Aghaeepour *et al.* Science Immunology. 2017

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

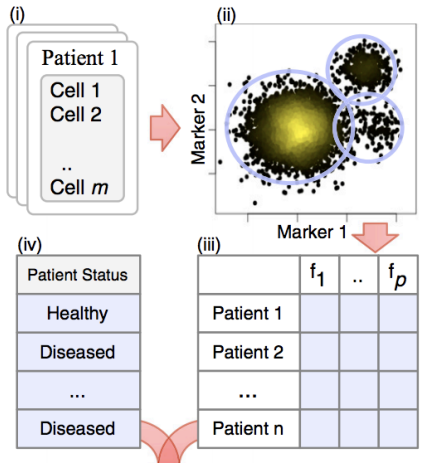

Figure: from Bruggner *et al.* PNAS. 2014

Practical Concerns. Converting data from CyTOF Machine to a Matrix.

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.

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

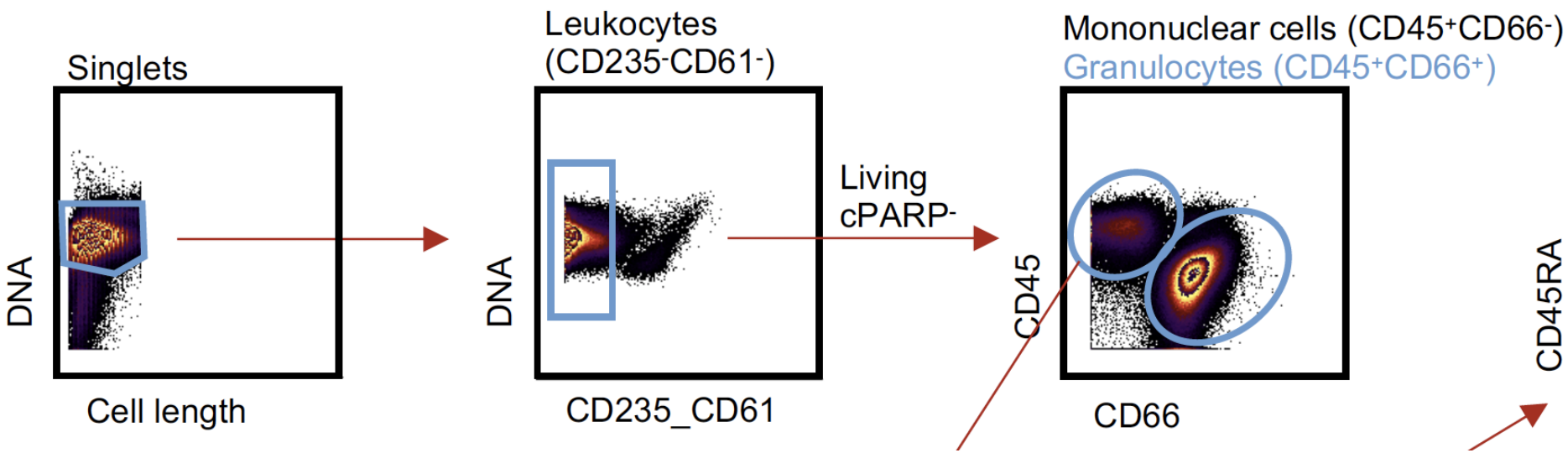

Figure: from Stanley *et al.* Nature Communications. 2020

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!

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.

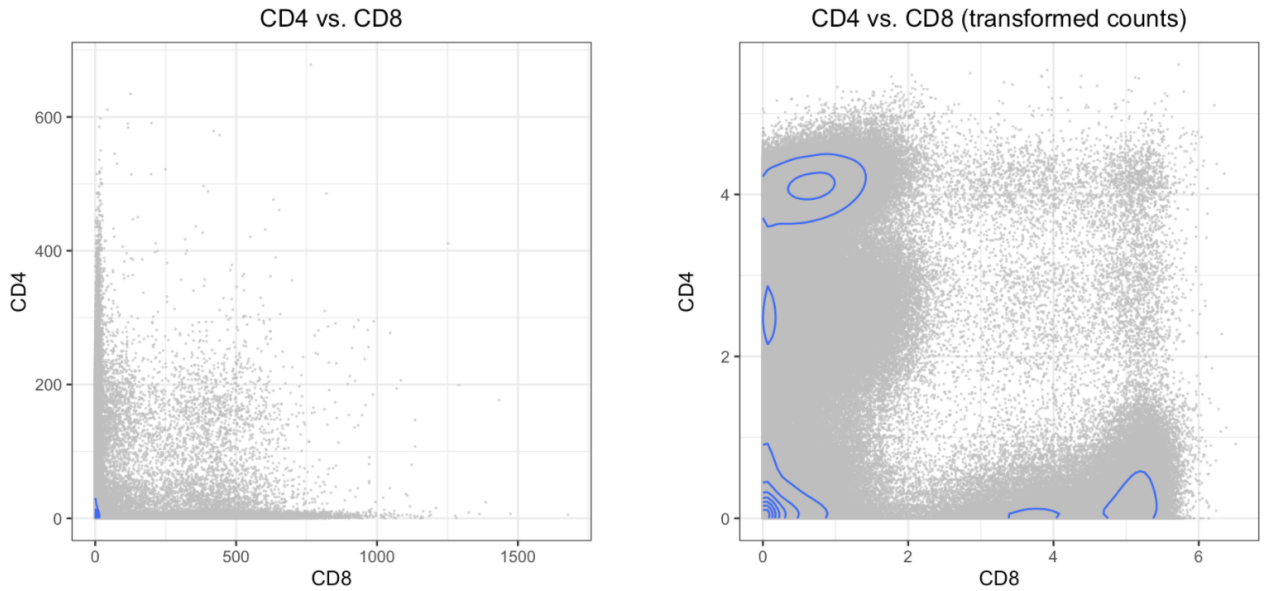

Figure: from `https://biosurf.org/cytof_data_scientist.html`

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 $x' = \text{asinh}(\frac{1}{5}x)$

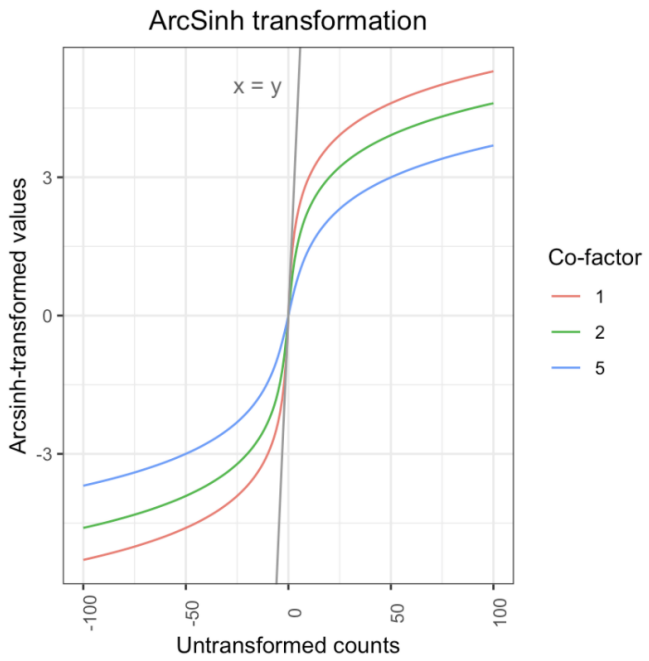

Figure: from `https://biosurf.org/cytof_data_scientist.html`

Effect of Normalization on Gating

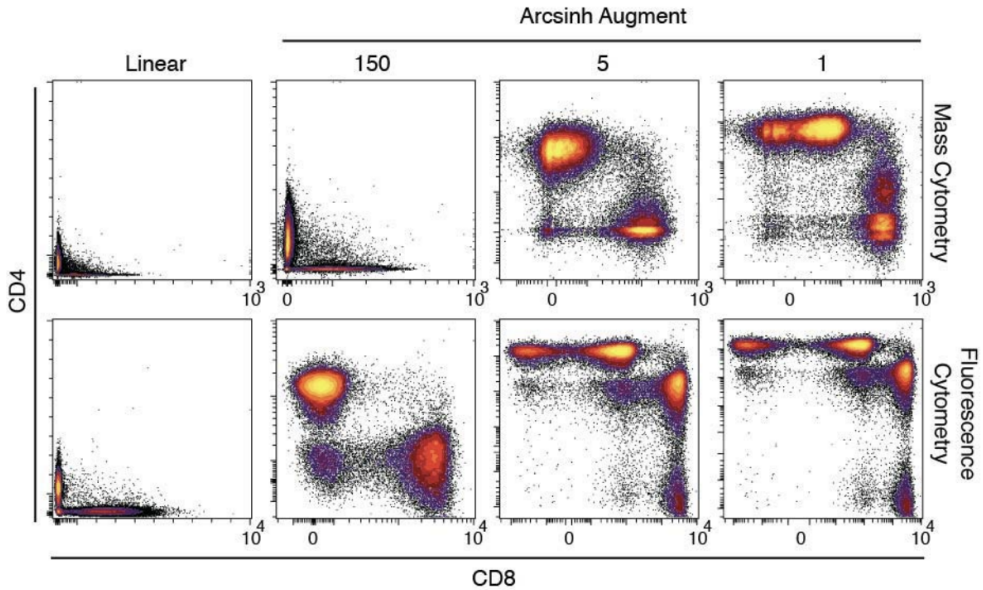

Figure: from Bendall *et al.* Science. 2011. The cofactor can cause sensitivity in the number of populations that emerge.

Tutorial

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

Recap

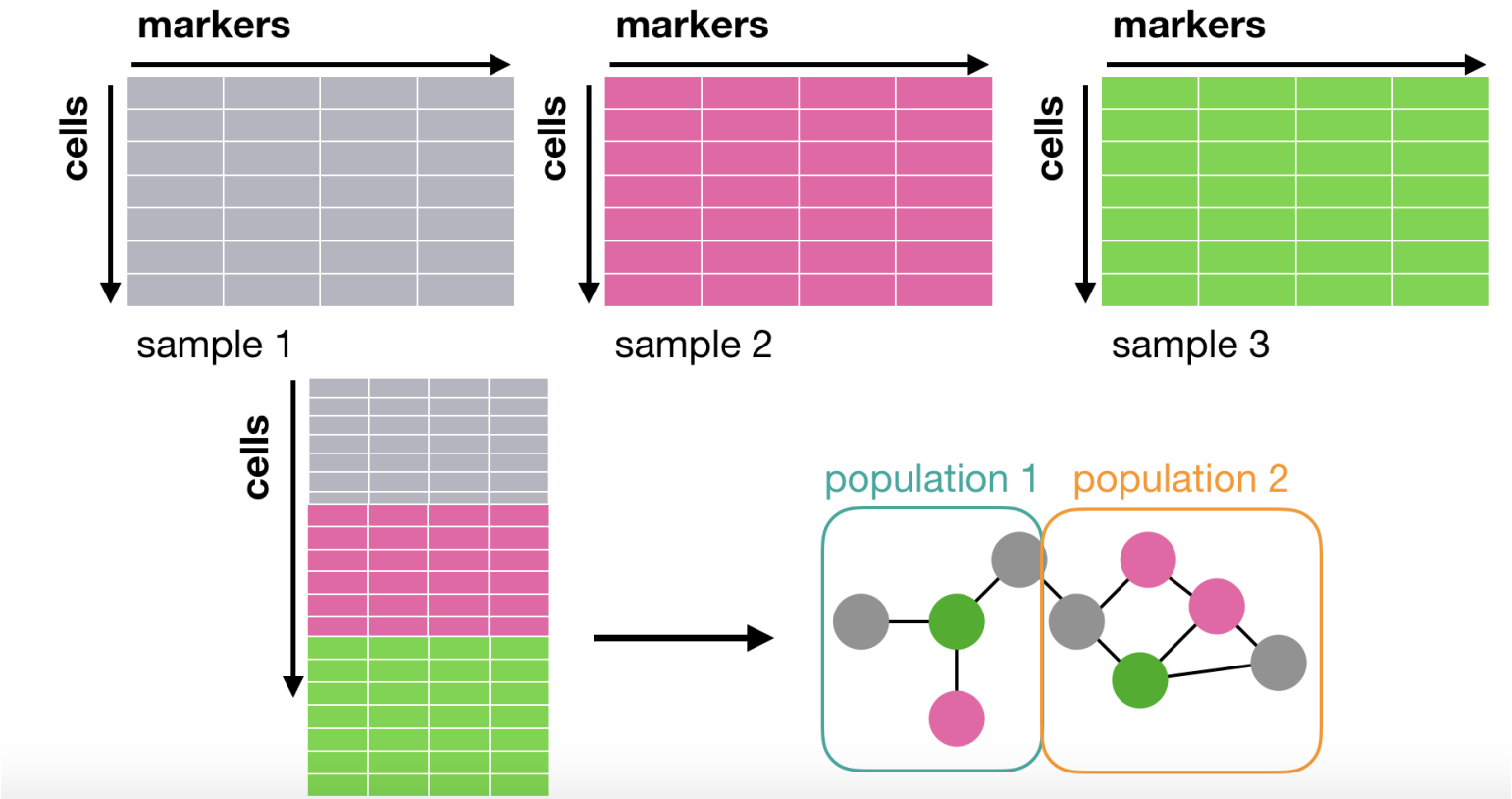

Figure: We are processing each FCS file, which corresponds to an individual sample. Ultimately cells are pooled across all samples.

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!

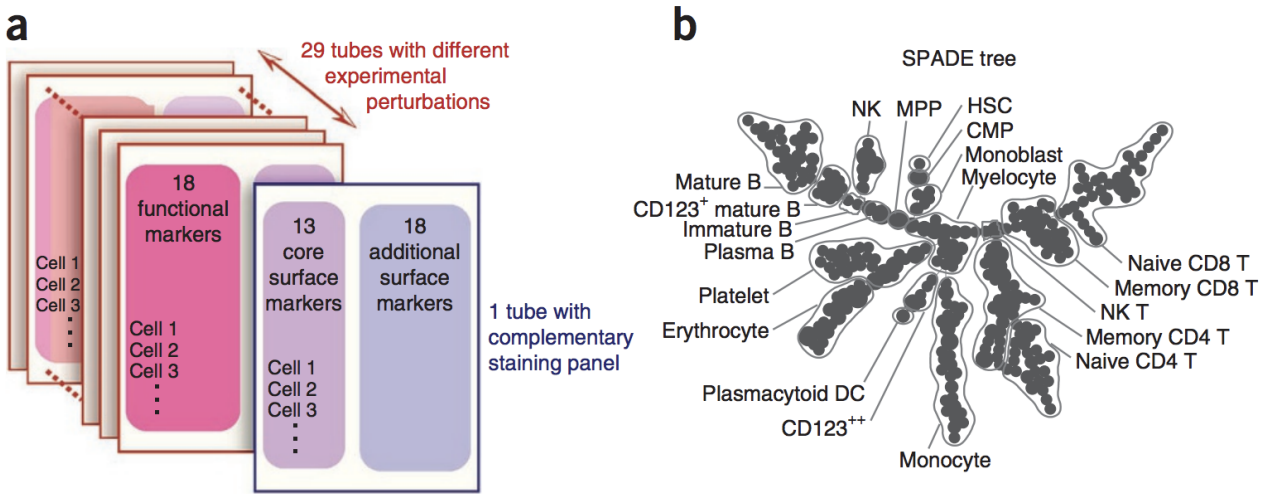

Figure: from Qiu *et al.* Nature Biotechnology. 2011. Do agglomerative hierarchical clustering based on the expression of measured markers.

Why Hierarchical Clustering?

It recapitulates our general understanding of cellular differentiation.

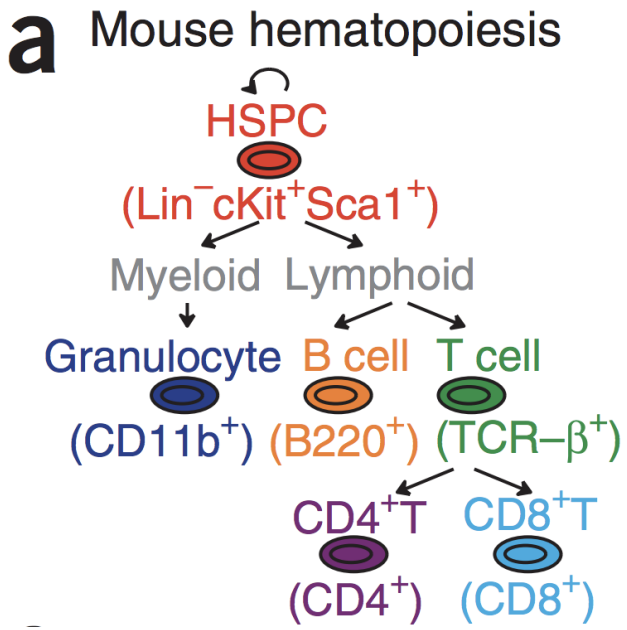

Figure: from Qiu *et al.* Nature Biotechnology. 2011. Cells can differentiate from Stem Cells to more specialized cell-types.

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...

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.

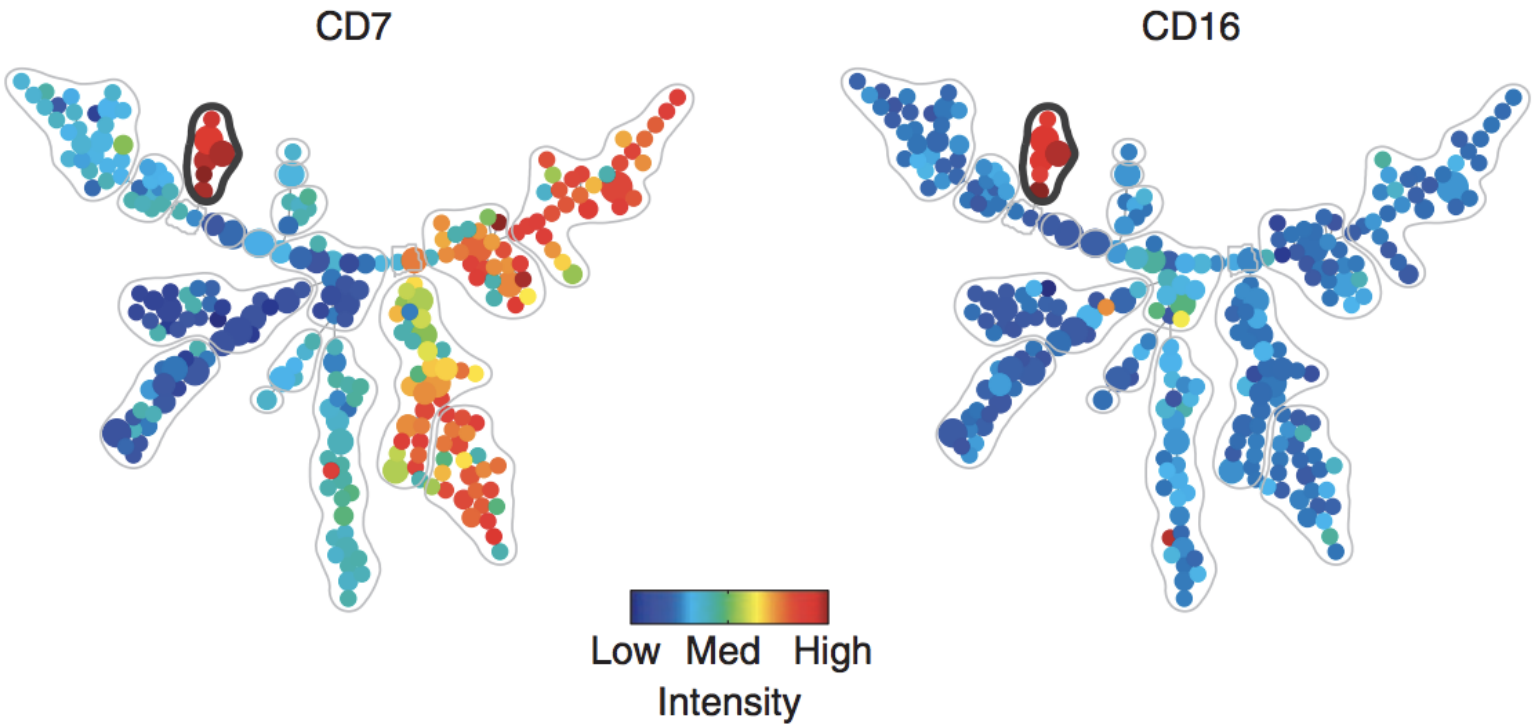

Figure: from Qiu *et al.* Nature Biotechnology. 2011.

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.

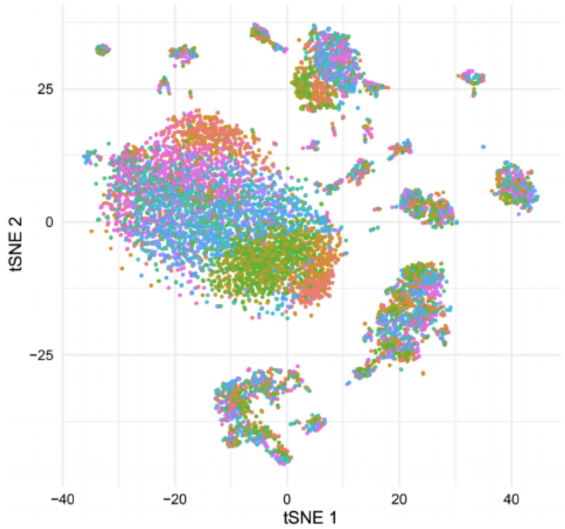

Figure: from Van Gassen *et al.* Cytometry A. 2019. Here cells are colored by batch.

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

Next time

- Unsupervised automated cell-population discovery methods
- Imputation in Single-Cell Assays
- Visualization