Milestone 2

Timothy Keyes

2020-05-11

Background

In this report, I am working with data originally published in a previous paper from my lab about leukemia biology that was published in one of the 2018 issues of the biomedical journal **Nature Medicine**. The paper devised a method of "aligning" cancer cells with the healthy cell type to which they are most similar. Overall, their idea was that comparing cancer cell subtypes to healthy cell subtypes might help us to infer how cancer cells behave or where they come from. The link to the original paper can be found here.

So far, I have been able to access the raw data from their original study–which can be found on the lab's GitHub page–and read it into both R and Python for preprocessing and some initial exploratory data analyses. So far, I have been able to recreate some of the figures from the original paper just to give myself a "sanity check" that what I'm looking at is relatively similar to what the authors were looking at.

Description of the dataset

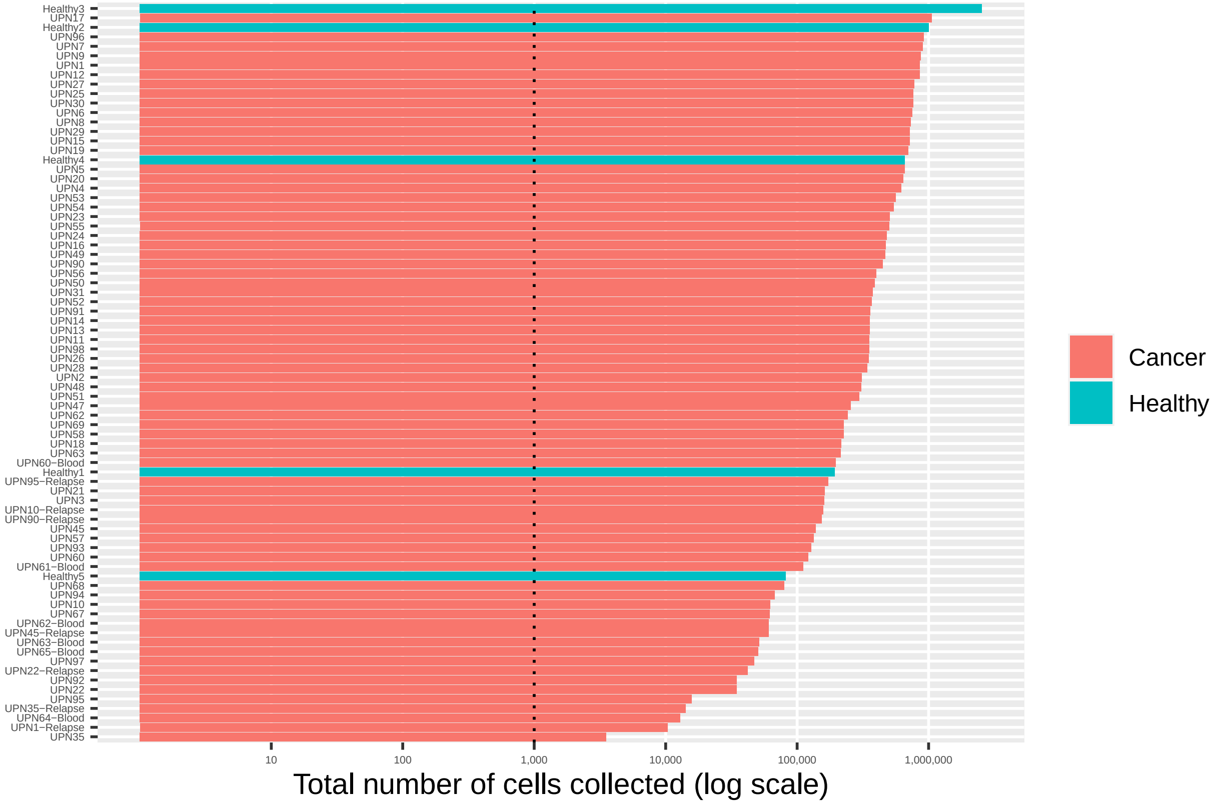
Specifically, our dataset contains mass cytometry data from 60 patients with B-cell precursor acute lymphoblastic leukemia (BCP-ALL) and 5 healthy control patients. The data were collected on a cytof2 mass cytometer and appropriately normalized, batch-corrected, and clinically annotated prior to our analysis. The data are stored as .fcs files, a file format developed by the International Society for the Advanced of Cytometry.

We report some basic information about the data we've obtained:

- Number of cells collected in each sample in total
- Number of cells within
- Some basic observations about the characteristics of the main cell populations in each sample

Number of cells

In most single-cell experiments, it is customary to count how many cells you were able to analyze for each patient enrolled in your study as a general measure of data quality. This is because it is easy to lose cells during the data collection process, and it is generally not advisable to analyze samples with fewer than several thousand cells. Here, we make a plot counting the number of cells collected for each patient in the dataset:



Source: Good et al., 2018, Nature Medicine

Thus, we can see that none of our samples have below 1,000 cells (and thus none need to be thrown away using this rule of thumb), but we also see that the number of cells collected between samples was NOT equal at all. Some samples had way more cells collected than others. In the original paper, the predictive model of relapse that the authors used incorporated features representing the measures of central tendency (medians, means, and percent "positive" cells over a certain threshold) within different cell populations within each of these samples. This is important because, given that the number of cells actually measured was very different across patients, the interpretation of these measures of central tendency shouldn't necessarily be the same (does the mean of a cell population with several thousand cells mean the same thing as the mean of a cell population with several hundred thousand cells?).

Importantly, there are not the same number of samples that come from healthy people and leukemia patients in the dataset at all…with only 5 healthy samples and 60 leukemia samples. The dataset's leukemia samples are not all the same - 17 of the patients from which the samples were collected ended up developing relapsed disease and 43 of them did not.

While these is not a particularly larger numnber of patients, there are many, many cells in the dataset in total and so I am interested in building a model that operates on the single-cell level (predicting if a cell is associated with relapse or not).

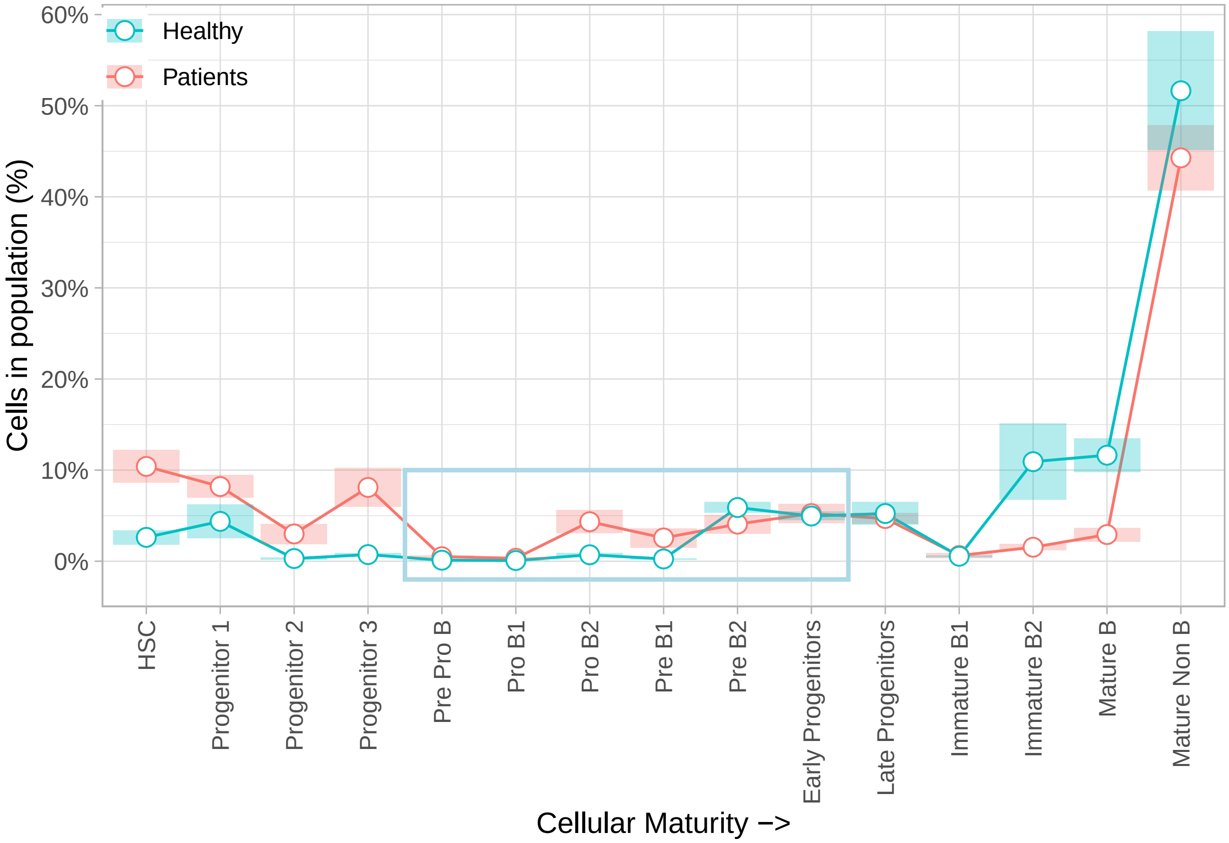
Number of total cells:
30480925

Abundance analysis of cancer cell subpopulations

In the original paper, the authors' first major finding was that certain subpopulations of cells are "expanded"–or present in higher proportions–in cancer cells relative to healthy cells of the same lineage. This result was summarized in Figure 3a of the manuscript:

In this figure, they found that several populations in the middle of the lineage's development were expanded relative to healthies. I was able to run the public implementation of the single-cell classifier (which was a custom in-house algorithm that they developed similar to linear discriminant analysis) that sorted cancer cells into different cell subtypes after the classifier was trained on healthy cells. The idea behind this classifier was to identify healthy-like cell subpopulations in cancer by applying an algorithm that was trained on healthy cells to cancer cells. On the original healthy samples, the classifier that they developed was about to sort cells into the "correct" (after gold standard manual identification by a biologist) subtype in nearly 90% of cases. On cancer cells, it is not possible to compute the accuracy because no real gold-standard exists.

The results from this classifier, at least compared to the original authors' reported results, were as follows:



From the above plot, we can see that our results are somewhat similar to the original results of the *Nature Medicine* paper, with some minor differences.

- Using their classifier algorithm, there are in fact several expanded "immature" cell populations in the cancer cells relative to the healthy cells. However, the boxed region indicates the cell populations that were noted to be expanded in the original paper (with the two demarked by arrows being the most important, or the most indicative of relapse in the model that the authors ultimately built). While we do see an expansion in the cancer cells relative to the healthy cells in the `Pro_B1` and `Pre_B2` populations as in the original paper, it does not seem to be as large of an expansion as the original authors found. In addition, most of the other populations in this region of the plot were not expanded despite the authors' original observations.
- Instead, we see that we have progenitor expansion earlier in the developmental "phase" of the cells, with expansions in the `HSC` and `Progenitor 1-3` populations (which were not observed in the original paper).
- In addition, we find a smaller spike in `Late Progenitors` in the healthy patients relative to the cancer patients than was found in the original paper.

There are several potential explanations for why our findings may differ from those in the original paper. First, the original paper did not include references to the healthy data used to perform the original classification (i.e. the "training set" for the developmental classifier). Thus, our classification was performed to a different dataset of healthy data collected by the same lab group (personal correspondence), but on a different machine. So, it is possible that there are some training set-test set differences in our classification that would lead to slight differences.

Thus, we were able to reproduce some of the main findings of the original paper while also being able to appreciate some of the difficulties of the reproducibility crisis in scientific research (particularly in the case where methdolody is not rigorously documented and provided to the public.)

Future directions

After some initial exploratory data analysis of the available data, I have gotten a bit lost of where to go next. I originally proposed using an autoencoder called SAUCIE, a multitasking network previously developed to batch normalize, denoise, and cluster single-cell data collected from human patients in the context of viral infection. I am still interested in potentially applying this network to the cancer data set to see if it identifies similar cell populations as the single-cell classifier that the **Nature Medicine** authors did, but after taking a closer look at the SAUCIE paper I am worried that they use more advanced methods than we have been able to discuss in our class (and there are several parts that I don't understand fully).

I've also become potentially interested in potentially using a convolutional neural network approach that has been applied to similar data as mine to predict clinical outcomes on the single-cell level (which could be useful for this data, seeing as there is such a limited number of patients but such a larger numnber of cells). This approach is called CellCNN and I am trying to decide if I am going to pivot my project to focus more on adapting this methodology for my application.