Analysis of Cell Type Frequencies in Single Cell Data as a Preparation for Bachelor Thesis

Michal Hruzik

Introduction

The motivation for this project is the topic of immune ageing. What do we know about immune ageing? From previous research (A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring - A. Alpert et al.) we know that there are about 30 sub-cell types of immune cells whose frequencies in blood change as we age. But what we don't know is why the frequencies of these sub-cell types change but in others we don't observe similar phenomena.

My goal in this project is to prepare the data of gene expression so that later on we can efficiently subset it and work with the relevant cell types. Specifically, we aimed to determine the positivity thresholds for various cell markers, count cells based on these thresholds, visualize the results, and perform statistical analysis to understand the distribution and significance of cell type frequencies.

The dataset we have used is a multimodal h5mu format dataset that contains single cell data obtained by SITE-seq. It was originally in R-format but we have converted it into format that is accessible with muon and scanpy.

With knowledge gained from this project I should be able to continue this project and utilize the results I have obtained in my bachelor thesis.

Method Description

Methodology involved several key steps:

1. **Data Loading**: The dataset was transformed from R-format to muon format for analysis. We primarily focused on the ADT part of the data for marker analysis.

2. **Positivity Threshold Determination**: Gaussian fitting was performed on the histograms of each cell marker to determine positivity thresholds. In cases where Gaussian fitting was not feasible, manual thresholding was used.

3. **Cell Counting**: Using the determined thresholds, cells were counted based on their marker expressions. Markers from the gatingStrategyAyelet were used as references for cell type definitions.

4. **Visualization**: Histograms and swarm plots were created to visualize the frequency of each cell type across multiple donors.

5. **Statistical Analysis**: ANOVA was performed to analyze the variance in cell type frequencies based on metadata variables such as gender, CMV positivity, and treatment status. This step further influences the bachelor thesis as we can identify possible bias.

Results and Discussion

Initial Data Exploration

Before starting the analysis we experimented with the data we have obtained. The main part of this exploration was learning to work with h5mu format. We had to determine where all the values were placed, how to index the dataset, and what metadata was present or in other tables. We also had to substitute markers not present in the dataset for those that could be used as a replacement. With missing markers we also had to update markers that defined the desired cell types.

Positivity Threshold Determination

Using Gaussian fitting, we determined positivity thresholds for various markers. The graph in `fig.1` visually represents this process. The x-axis represents the marker expression levels, while the y-axis represents the count of cells. The intersection points of the Gaussian curves indicate the positivity thresholds.

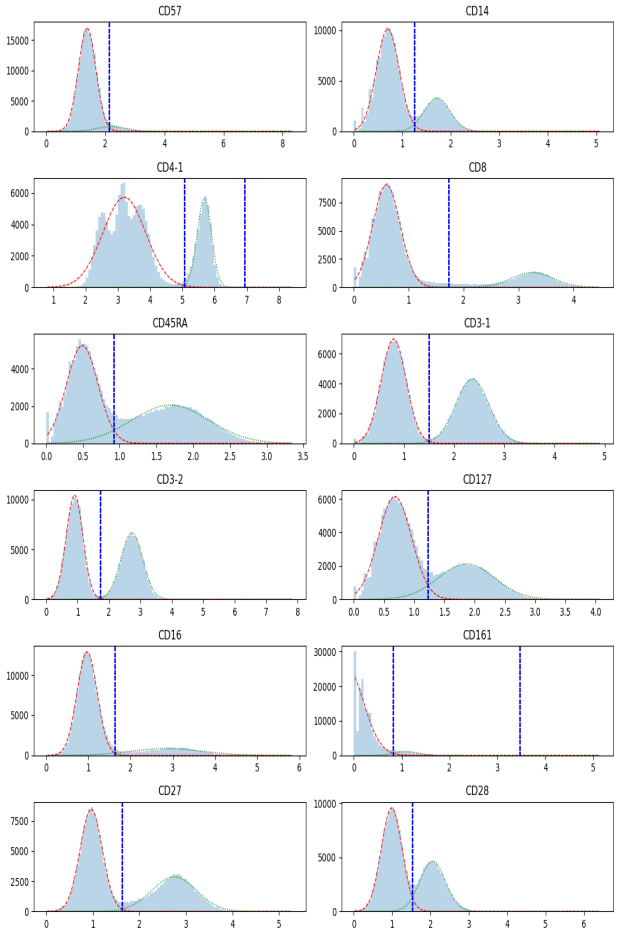

Figure 1: Gaussian fitting on histograms to determine positivity thresholds. The x-axis shows marker expression levels, and the y-axis shows the count of cells. The intersections of the Gaussian curves indicate the thresholds.

Cell Counting and Visualization

With the thresholds determined, we counted the cells for each donor and visualized the results. The bar plot in `fig.2` shows the frequency of each cell type for all donors. The x-axis represents the cell types, and the y-axis represents the cell frequency. We had to introduce some replacements for missing markers but overall this went smoothly.

The swarm plot in `fig. 3` provides a more compact comparison of frequencies. Here, the x-axis represents the cell types, and the y-axis represents the cell type frequencies. Each dot represents the value for one donor, color-coded by donor, allowing for an easy visual comparison across donors. We notice significant differences in donors which leads us to our next part statistical analysis.

Statistical Analysis

ANOVA was performed to understand the impact of metadata variables on cell type frequencies. The results, saved in `anova_res.txt`, show the p-values for each variable tested. A significant p-value indicates that the variable has a substantial effect on the frequency distribution of cell types. This file is very important for our future analysis because we want to avoid biases. In this file we observe results:

```
Now running anova for: Treatment

Null hypothesis rejected with cell type: CD57 + NK
and p value: 0.01098
Null hypothesis rejected with cell type: CD185 + CD4-1 + CD4-2 + T
and p value: 0.02537
Null hypothesis rejected with cell type: CD27 + CD8 + T
and p value: 0.03467
Null hypothesis rejected with cell type: NK
and p value: 0.02709
Null hypothesis rejected with cell type: T
and p value: 0.01297
Null hypothesis rejected with cell type: Naive CD4-1 + CD4-2 + T
and p value: 0.046
Null hypothesis rejected with cell type: CD161 - CD45RA + CD4-1 + CD4-2 + Treg_cel
```

```
and p value: 0.0382
Null hypothesis rejected with cell type: Naive CD8 + T
and p value: 0.03135
Null hypothesis rejected with cell type: Plasmablast
and p value: 0.02698

Finished running anova for: Treatment

Now running anova for: CMV_positive

Null hypothesis rejected with cell type: CD8 + T
and p value: 0.00444
Null hypothesis rejected with cell type: Plasmablast
and p value: 0.01188

Finished running anova for: CMV_positive

Now running anova for: Gender

Null hypothesis rejected with cell type: B
and p value: 0.00395
Null hypothesis rejected with cell type: CD8 + T
and p value: 0.02921
Null hypothesis rejected with cell type: Plasmablast
and p value: 0.00623

Finished running anova for: Gender
```

This implies that Treatment can introduce a bias into our analysis - and we don't want this. After consultation with Martin L., we have decided that in our future work we will exclude the patients that obtained treatment 1. Treatment 1 indicated that the donor was treated with IL-12 which is a strong cytokine.

Conclusion

This project involved several challenging and rewarding aspects:

Challenges Encountered:

- Gaussian fitting was case-sensitive and required significant tuning. I am quite disappointed that I didn't manage to automate this whole thing.
- Large data size necessitated efficient handling and processing techniques.

Easily Accomplished:

- Visualization of cell type frequencies using histograms and swarm plots.
- Manual thresholding for markers where automated fitting was not feasible.
- Using boolean mask as a tool for counting cells.

Recommendations for Future Work:

- Improve the automation of Gaussian fitting to reduce manual intervention.
- Explore additional statistical methods to further validate the findings.
- Refactor the class system.
- Create scripts that save results in a more presentable format.

Learnings:

- Handling and processing large single-cell datasets.
- Application of statistical methods to biological data.
- CITE-seq and markers.
- Immune system and immunology.
- h5mu data format.

Overall, this project provided a comprehensive analysis of cell type frequencies, highlighting the significance of marker thresholds and the variability across donors. I have learned a lot about working on an unknown datset in unknown format. The single cell data was really interesting and I could not ask for a better dataset. The project had some low points and some exciting moments.

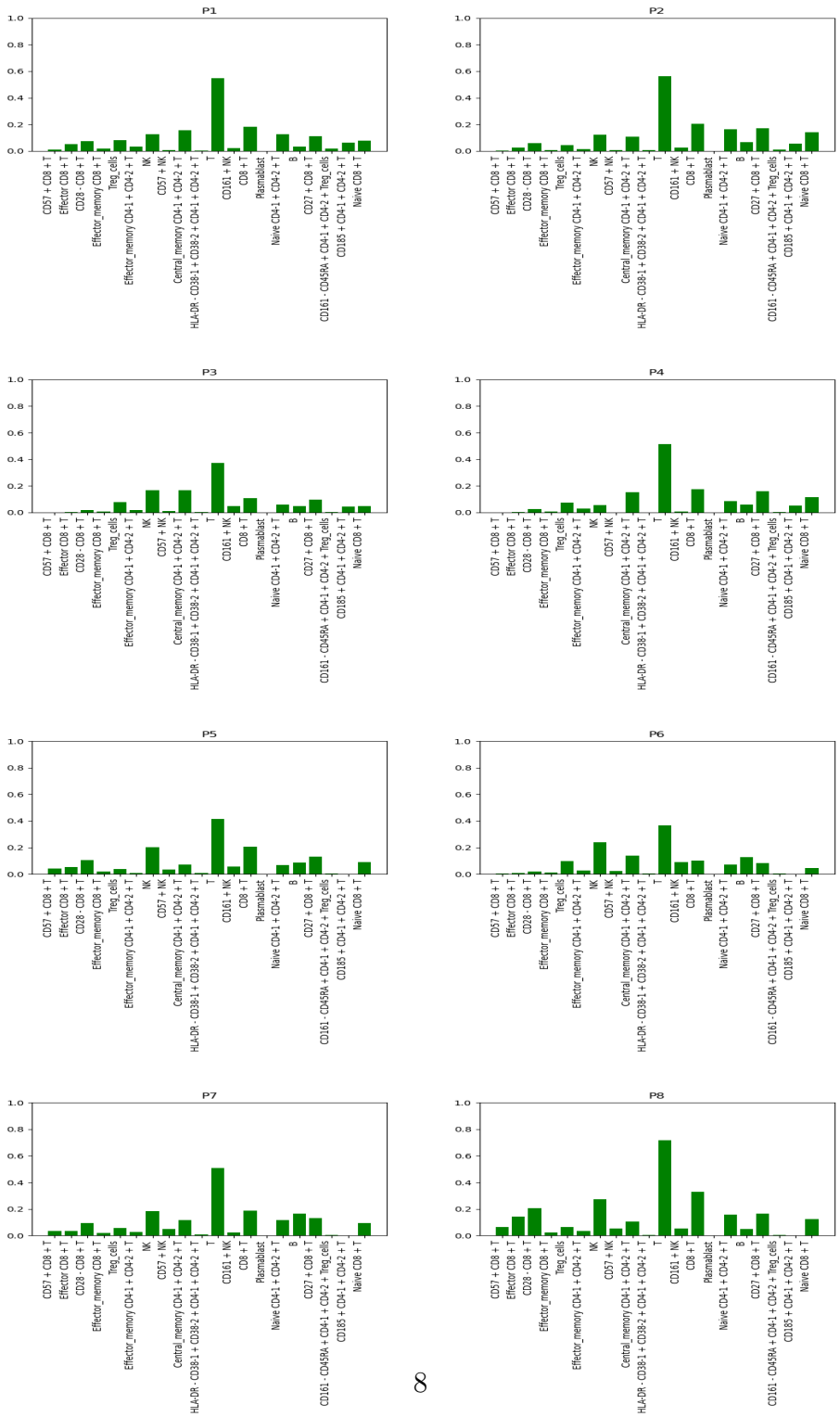

Figure 2: Bar plot showing the frequency of each cell type for each donor (P1-P8). The x-axis represents the cell types, and the y-axis represents the cell frequencies.

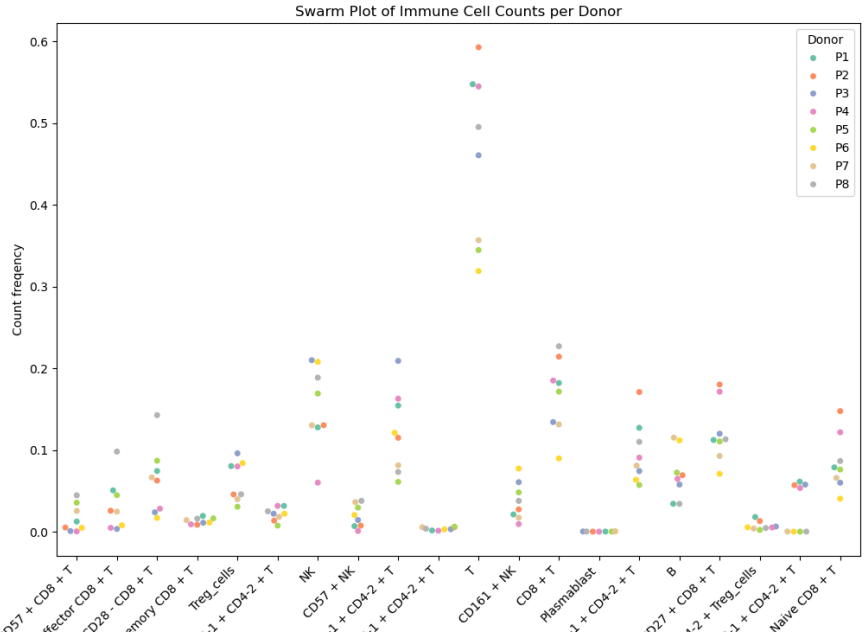

Figure 3: Swarm plot showing the frequency of each cell type for each donor. The x-axis represents the cell type, and the y-axis represents the cell type frequencies. Each dot represents the value for one donor, color-coded by donor.