A single-cell atlas of the peripheral immune response in patients with severe COVID-19

Julia Rymuza

June 2021

1 Abstract

Batter understanding of pathophysiology of COVID-19 is very important. This disease is responsible for millions of deaths and right now strongly influence our life. There fore there are many studies conducted in this area, one of them is [3]. I was able to partly reconstruct the results presented in this article. Furthermore I tried to improve the results with taking into account batch effect and also using new web server called Azimuth [5], that allows mapping to reference. Comparison of different presented approaches showed rather similar results. They were consistent in assigning labels and detecting difference cell proportion between patients with COVID-19 and healthy donors.

All scripts and results, exept seurat objects can be found at github.com/Julia820/CBS_project.

2 Introduction

Coronavirus infected over 174 million people and is responsible for over 3 million deaths [1]. We still have troubles understanding pathophysiology of Coronavirus disease 2019 (COVID-19). Approximately 20% of patients that get infected develop severe symptoms and 5% require intensive care [2]. Researchers believe that severe symptoms can be connected to changes in peripheral immune activity.

To understand pathways in peripheral immune cells scientists applied single-cell RNA sequencing (scRNA-seq) to profile peripheral blood mononuclear cells (PBMCs). scRNA-seq is a technology that allows us to examine sequences from individual cells. PBMCs on the other hand is part of blood that consists mainly of lymphocytes. It also includes monocytes and dendritic cells. The study consists of seven patient hospitalized for COVID-19, four of whom had acute respiratory distress syndrome, and six healthy controls [3].

One of tools allowing analysis of scRNA-seq data is Seurat [4]. It enables us to cluster, examine quality and visualize data. Recently there was new addition to this project, web-server called Azimuth [5]. It can be used to map data set to reference. It automates the process of visualization and annotation.

In my project I aim to partly reproduce results from [3]. Additionally I want to check whether taking into account batch effect will change the result. Since data was taken from different patient we can expect technical noise to influence the outcome of the analyze, something that was not considered by researchers in original article. Moreover I want to test performance of new web-server Azimuth and compere it with results of original workflow.

3 Methods and materials

3.1 Data preparation

Due to limited computational power I had to limit data set. I took into consideration only samples from six people. Out of them two were healthy, four have been hospitalized for COVID-19, two of whom had acute respiratory distress syndrome. Furthermore I was forced to limit number of analyzed genes. I took only genes witch had different expression based on supplementary tables of [3]. I also check for missing data.

After limiting data size I performed ComBat from sva library on data with samples as batches. Next I used both data sets to create two seurat objects. Afterwards I added metadata based on file *meta_data.csv* and calculated the percentage of cell features. That allowed me to check the quality of the data and filter outliers. Seurat objects after such a reprocessing were saved to a file so they can be used with Azimuth workflow.

3.2 Classic workflow

Next step of classic analysis using Seurat was normalizing, finding variable features and scaling data. Matrices prepared this way were ready for running PCA and UMAP. Afterwards I created Shared Nearest Neighbor (SNN) Graph and found clusters. For each cluster I looked for most variable features. Results of that are saved to file *cluster_quality.csv*.

Following that I used SingleR for annotation. I took HumanPrimaryCellAtlasData() as reference. Using function provided in that library I annotated type to each cell. Outcome of that was saved to *seurat_cell_ad.csv*. Subsequently I wanted to assign the cell types to clusters. For that I checked most frequent annotation present in given group. Results of that are saved to *seurat_cluster_ad.csv*. Finally I assign to each cluster most frequent cell type.

Results of presented workflow were saved to file. All mentioned tables can be found in my github repository. Due to limited file size I was not able to upload rds files. Mentioned procedures were performed using *data_prep.Rmd* markdown.

3.3 Azimuth workflow

Owing to automatized nature of Azimuth no additional steps were necessary. All I had to do was upload rmd file to web server and download results. As Azimuth uses an automated process for mapping and assigning cell types, we have various quality control options. We get information about the percentage of query cells with anchors and the cluster preservation score. The first gives us information about the percentage of query cells that can be identified as ”anchors” or correspondence pairs between cells that are predicted to be in a similar biological state, between query and reference data sets. On the other hand, the cluster preservation score result represents the behavior of the unsupervised cluster structure and is based on the entropy of the unsupervised cluster labels in the local neighborhood of each query cell after mapping.

3.4 Creating plots

For creating plots I created three files. First *plots.R* contains row functions necessary for achieving results. Other two *azimuth.Rmd* and *creating_plots.Rmd* create plots respectively for Azimuth and original workflow results. The purpose of those scripts is reconstructing plots 1b,c,d from original paper. Additionally I created cell proportion plots for each cluster and cell proportion group by donor but with separation to clusters.

4 Results

4.1 Data preparation

In original study they analyzed 44 721 cells with an average of 3 194 cells per sample, for me it was 25 886 cells with an average of 4 324. Furthermore I took only 2 140 genes out of 37 811. The dataset did not contain missing data.

Results of checking data quality can be seen in Figure 1 and Figure 2. We can see that plots for both of this groups are quiet different, especially for number of features. I decided to filter cells that had nCount_RNA > 10000, percent.mt > 18 and nFeature_RNA > 1200 for data from original workflow. For ones corrected for batch effect I limited nCount_RNA < 10000 and percent.mt < 18. That resulted in analysing 25 849 cells in original workflow and 25 812 in one with correction.

4.2 Classic workflow

4.2.1 Original data

For data from original workflow analysis resulted in 26 clusters. The adjusted p-value from differential expression, based on Bonferroni correction using all features in the data set for groups can be seen in Figure 3a. For most of them almost all values are close to zero, so we can think of them as statically significant.

Results of assigning cell types can be seen in Figure 4a. It shows that although belief in classification was rather strong other labels were also had high scores. The final type was chosen thug based on most abandoned group in the cluster. The contribution distribution can be seen on additional plots that can be found in *tables/org/plots/*. Even though for most clusters almost all of the cells were assign the same label, it is not always the case. For example in cluster 12 we have many T cells, monocytes and B cells. Despite that fact it was labeled as monocytes.

Dimensionality reduction presented in Figure 5, indicated substantial phenotypic differences between patients with COVID-19 and health ones, predominantly in T cells, B cells and monocytes. Is is also worth noticing that not that many different cell types were found and most of them are rather general.

Cell proportion plots can be seen in Figure 7. We can clearly see that patients with COVID-19 have depleted levels of NK cells and T cells. On other hand proportion of monocytes is higher for them. When we look directly on clusters not summary for patients we do not see this dependencies. Additionally from those plots we can see that T cells, NK cells and monocytes contain the biggest amount of clusters.

4.2.2 Corrected data

For data from original workflow including correcting for batch effect I got 25 clusters. The plot of corrected p-values can be seen in Figure 3b. Although for most of the clusters the plots look very good, it is not as good as for uncorrected data. The amount of clusters with genes that have high p-values is bigger.

Heat map of scores of assigned cell types is showed in Figure 4b. The belief seems to be higher in this case then the previous one but plots are not very different from each other. The same as in previous case the final cell type was assign based on most frequent type in the cluster. Proportion of different labels in given cluster can be found in *tables/combat/plots/*. As in original data here also we have many groups with clear labeling but also some with many different cell types. For instance in cluster 3 we have T cells, monocytes and B cells in almost exact promotion.

The results of UMAP are shown in Figure 6. They show phenotypic differences between patients with COVID-19 and health ones, predominantly in T cells, B cells and monocytes. It is worth noticing that same type cells are not grouped together well in the plot. Also again we were able to find only seven different cell types.

Plots showing cell proportions can be seen in Figure 8. This time we see only low levels of NK cells and elevated of monocytes for patients with COVID-19. Looking directly at clusters this trend seems to be preserved in NK cells. Moreover from the plot we can see that B cells, T cells and monocytes comprise of highest amount of clusters.

4.3 Azimuth workflow

4.3.1 Original data

For this data we have 8.36% of query cells with anchors, which is marked as possibly problematic. Although that dos not mean that the mapping was unsuccessful. We also have cluster preservation score equal to 5/5, which is the best one possible.

Figure 9 shows the results of mapping to the reference. The separation between patients with COVID-19 and healthy donors can be seen only for monocytes. Unlike in the classical workflow we do not see such a clear separation between health and infected. Using this approach allowed us to find many more cell types and also we have even more detailed information about them.

Information about cell proportion is shown in Figure 11. We observe depleted levels of CD4 T cells and NK cells and elevated of monocytes for infected donors. Looking into more detailed information about cell types we see high levels of CD14 monocytes and low of B memory, B intermediate, CD4 TEM and NK cells.

4.3.2 Corrected data

The percentage of query cells with anchors in this case is equal to 5.14%, which is marked as possibly problematic. Although that does not mean that the mapping was unsuccessful. We also have cluster preservation score equal to 1.26/5, which is likely problematic. As before that do not have to influence the quality of mapping.

The results of mapping to reference can be seen in Figure 10. They present separation between patients with COVID-19 and health ones for monocytes and B cells. Again the segregation between samples from infected individuals and others is not that clear as in classical workflow. Moreover the amount of different cell types identified is higher, but not as high as for original data.

Cell proportion plots can be seen in Figure 12. We can clearly see depleted levels of NK cells and CD4 T, also slightly elevated of monocytes for infected donors. When we look in to more detailed labeling we get lower proportion on NK cells, dnT, B memory, B intermediate and CD4 TEM and higher of CD14 monocytes for patients with COVID-19.

4.4 Comparison

I compered call annotations from classic workflow. Although the amount of cells that was analyzed is different for original data and the ones corrected for batch effect, they have 25 780 cells in common. For all of them the final annotation is the same regardless of applying ComBat.

On the other hand similar comparison for results of Azimuth is not that consistent. We have 6 892 cells with different predicted.celltype.l1 and 8 141 with predicted.celltype.l2. That shows that using ComBat change the result, something that was not observed in the classic workflow.

Unfortunately due to different cell names and my limited knowledge about PBMC I was unable to compere results between classic workflow and Azimuth.

Compering the cell proportions between methods revels that regardless of way the data was processed always the same cell types were interesting. In all the analyzes we observe depleted levels of NK cells and T cells and elevated of monocytes.

5 Discussion

To some extent I was able to reproduce the research presented in [3]. Although the limited computational power influenced my results. The classic workflow was not able to identify so many cell types. Moreover dimensionality reduction did not indicate substantial phenotypic differences between patients with COVID-19 and health ones as strongly as original Figure 1b,c. This situation is especially visible for results of Azimuth workflow. Additionally cell types that seemed to be separated the best for me are T cell, B cells and monocytes while for the researchers that was T cells, monocytes and NK cells.

Interestingly I was able to detect some of the cell proportion differences that can be seen in [3]. Same as the authors I got depleted number of T cells and NK cells but not monocytes with in my study are elevated.

It is also worth to notice that the problems with quality of Azimuth mapping according to the creators can be explained by a homogeneous group of cells or by batch effect. In light of this it is surprisings that data that was corrected by ComBat had even lower scores.

As a summary from my study we can see that correcting for batch effect did not influence strongly the outcome. It is still a very important step that should be conducted in that type of analysis. It is possible that chosen method was not best suited to the problem. In future it would be interesting to apply batch effect correction method that is implemented in Seurta. In this case we would use ScaleData() with vars.to.regress.

Moreover from presented results we can see that Azimuth handled the problem really well. It was more accurate and much faster, also it required less preprocessing. It is worth remembering that this method also has some downsides. We have less control over the process and we cannot detect cells that are not in the reference. The last cone was not observed in my study but was mentioned in [4].

References

[1] JHU CSSE COVID-19 Data
`https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6`

[2] Wu, Zunyou, and Jennifer M. McGoogan. *Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention.* Jama 323.13 (2020): 1239-1242.

[3] Wilk, A.J., Rustagi, A., Zhao, N.Q. et al. *A single-cell atlas of the peripheral immune response in patients with severe COVID-19.* Nat Med 26, 1070–1076 (2020).

[4] Hao, Yuhan, et al. *Integrated analysis of multimodal single-cell data.* Cell (2021).

[5] Azmiuth
`http://www.satijalab.org/azimuth`

Supplementary images

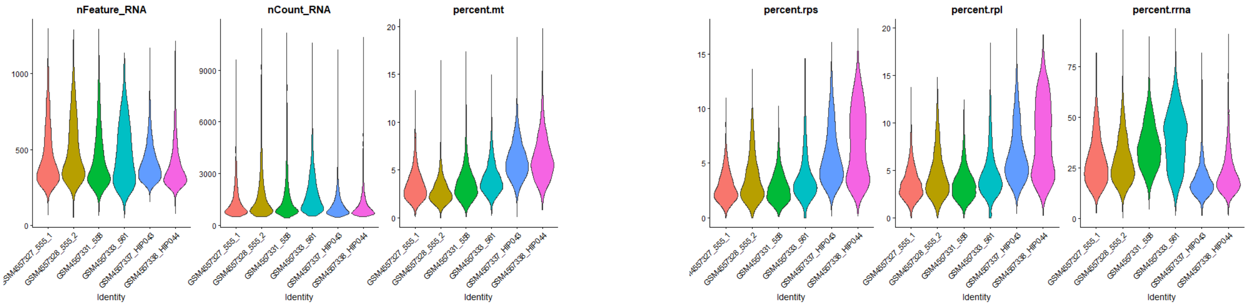


Figure 1: Vialoin plot of uncorected data

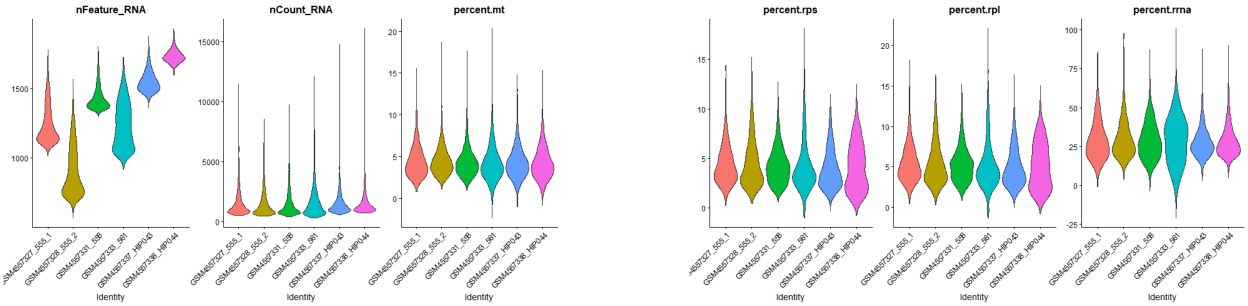
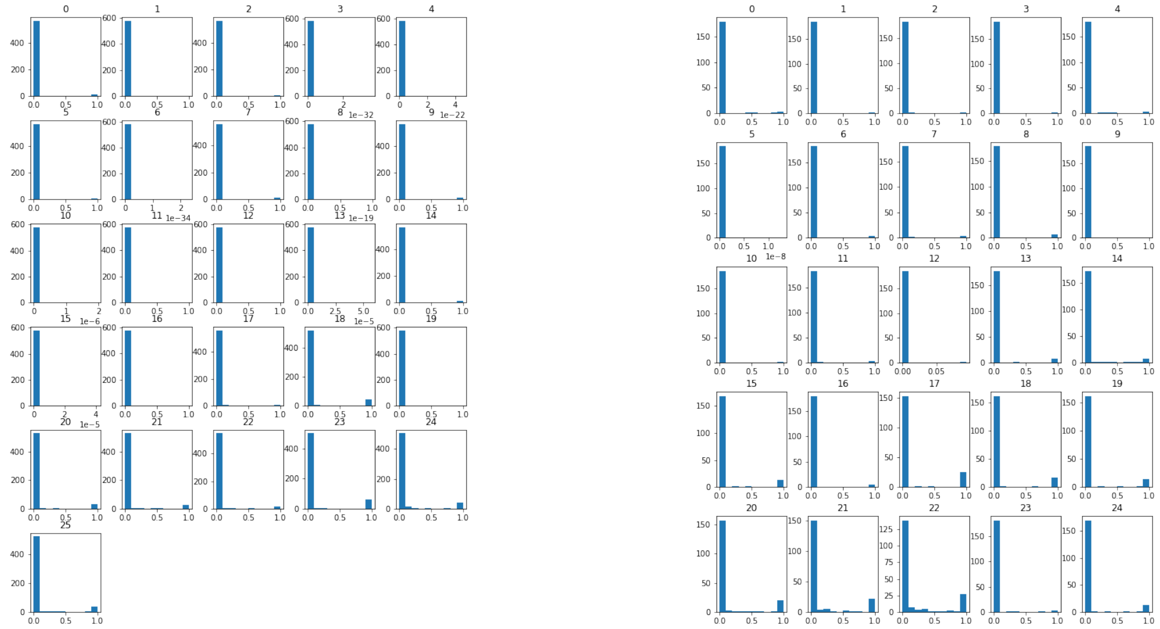


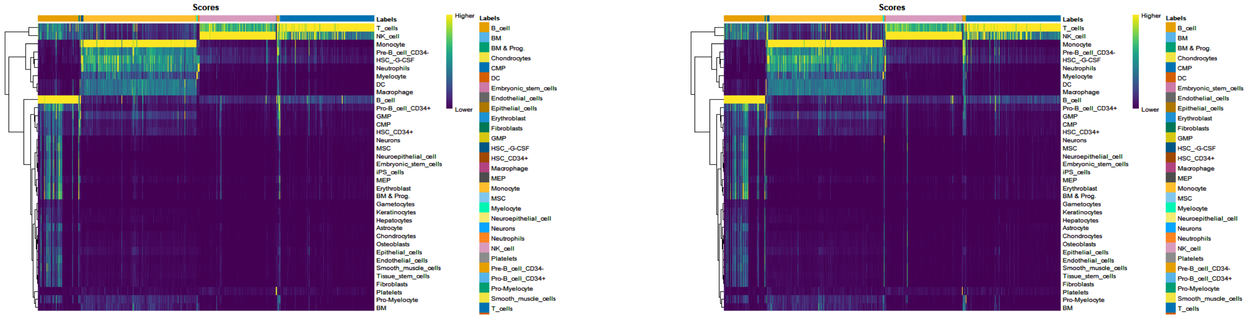
Figure 2: Vialoin plot of corected data



(a) original data

(b) corected data

Figure 3: Hisogram of p-values for clusters



(a) original data

(b) corected data

Figure 4: Heatmap of cell type annotation

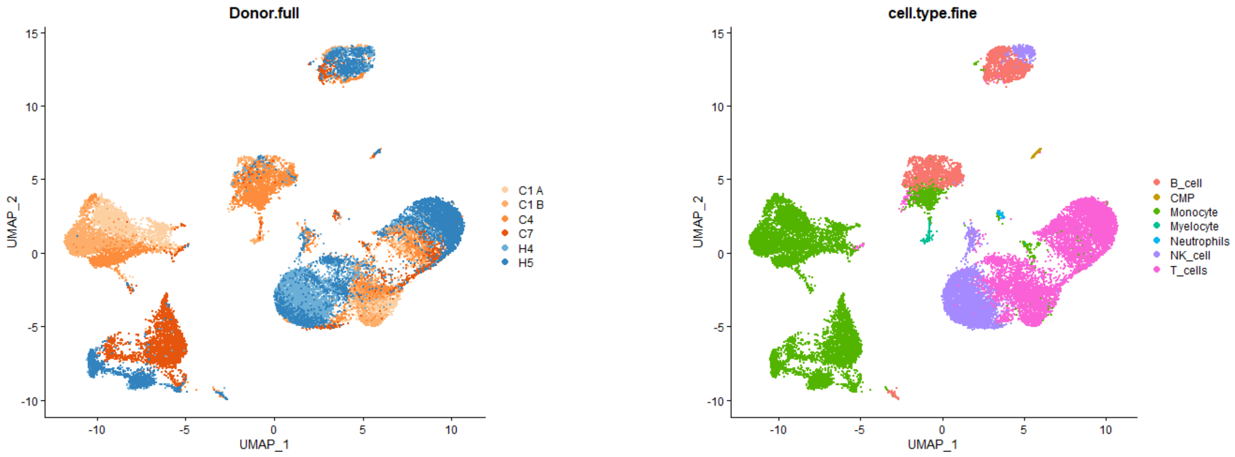


Figure 5: UMAP of PBMCs for uncorected data

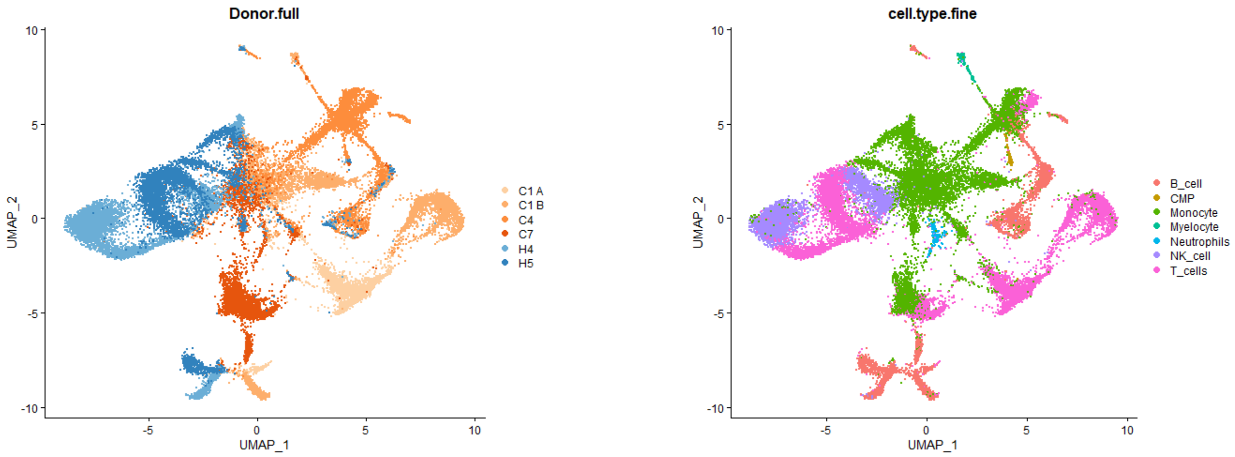
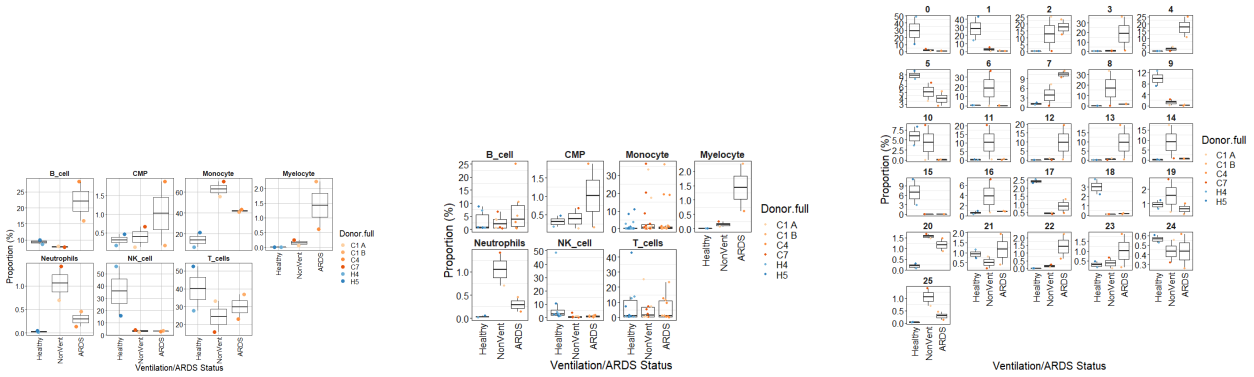


Figure 6: UMAP of PBMCs for corected data

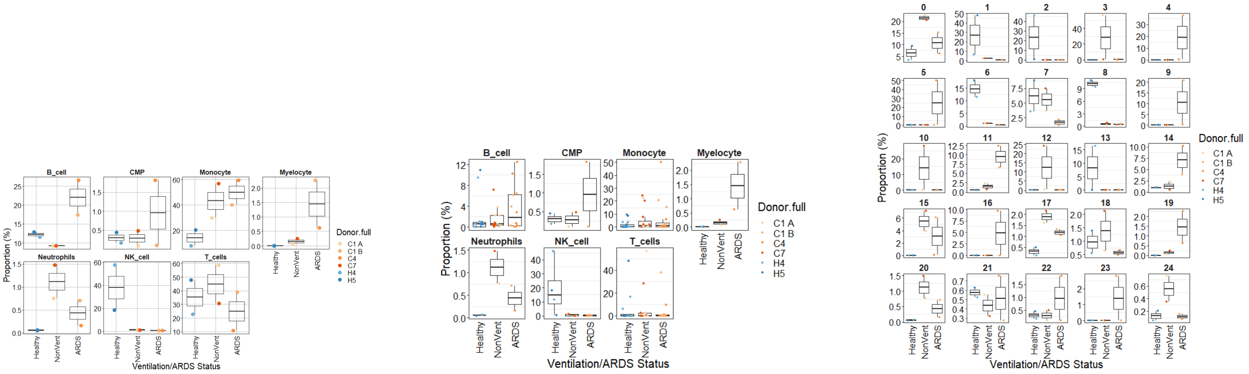


(a) gropu by donor of origin

(b) gropu by cluster

(c) gropu by donor of origin

Figure 7: Cell proportion for uncorected data



(a) gropu by donor of origin (b) gropu by cluster (c) gropu by donor of origin

Figure 8: Cell proportion for corected data

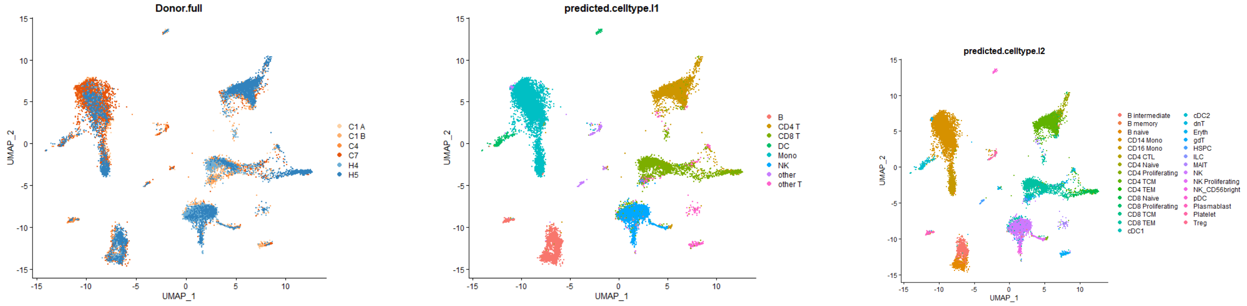


Figure 9: UMAP of PBMCs for uncorected data prossesed by Azimuth

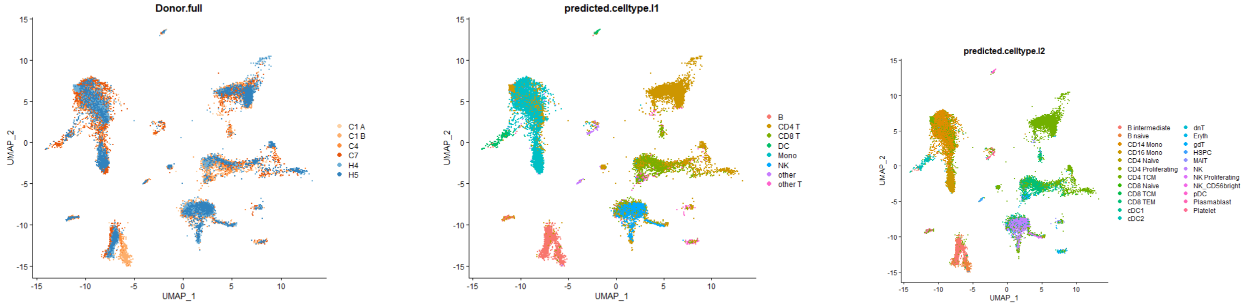


Figure 10: UMAP of PBMCs for corected data prossesed by Azimuth

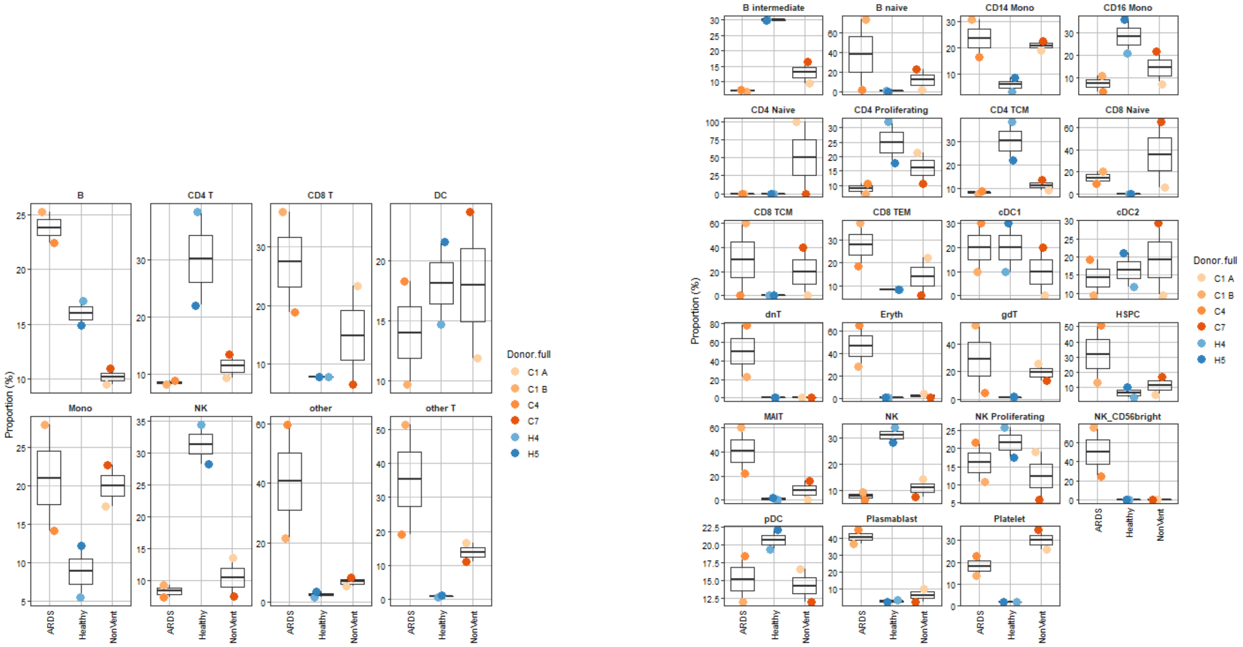


Figure 11: Cell proportions for uncorected data prossesed by Azimuth

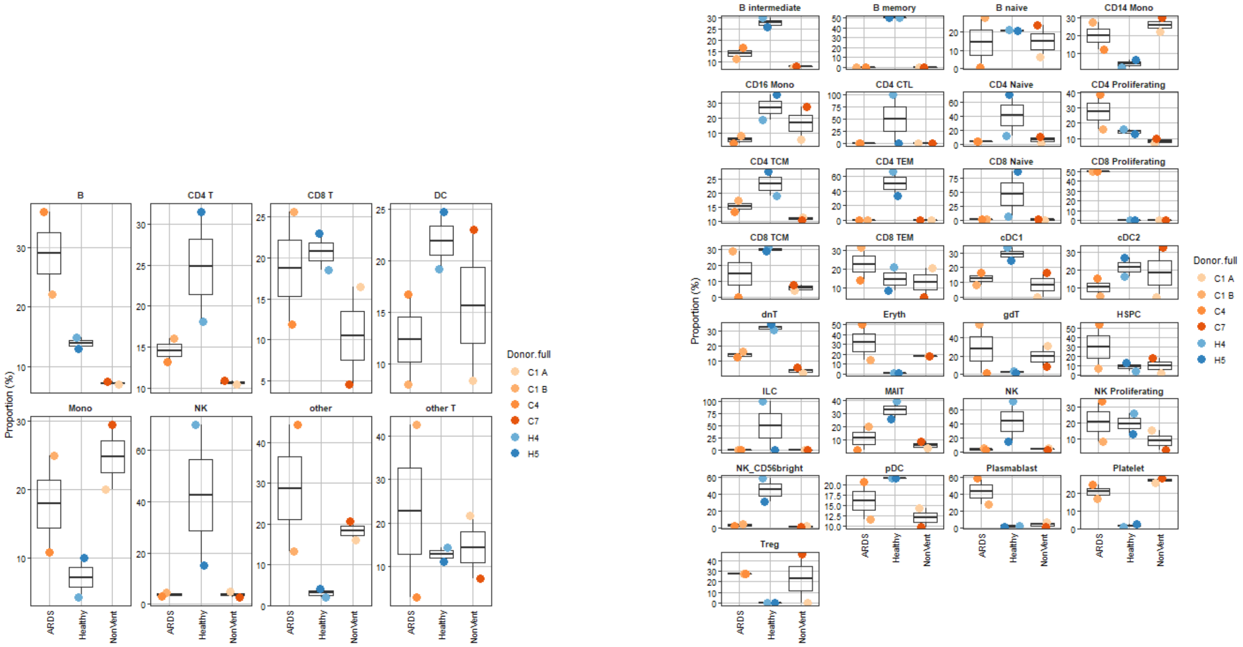


Figure 12: Cell proportions for corected data prossesed by Azimuth