**Title:**

Analysis of lung single-cell sequencing leads to the characterization of heterogeneity of macrophages in lung fibrosis

**Background:**

In this day and age, lung fibrosis is one of the prominent causes of deaths in the world. This is especially relevant in the case of COVID which is known to have symptoms revolving around one's respiratory system. Lung fibrosis is a disease that is caused from damaged or scarred lungs which can inhibit an individual’s ability to use these organs properly. Macrophages are phagocytic cells found in tissues, especially when there is an infection. Since macrophages are so deeply interconnected with this disease, it was only reasonable to find the characterization and the data from them to deeply analyze this topic. The macrophage in question which is deemed a “disease associated subgroup” is the alveolar macrophage which is seen as the root cause of fibrosis diseases in the paper of interest.

To study the heterogeneity of macrophages, single cell data is needed because it resolves diversity within cell populations. In the dataset studied in this analysis, there is a mixture of different macrophages. Looking at single cell data should enhance understanding of macrophage heterogeneity and specify subsets for functional analysis in fibrosis, which is the motivation of this analysis.

With the use of different tools such as Seurat, significant data was found which could be used to further understand the implications of fibrosis and the different characteristics of the macrophages. This process provides insight into certain macrophage patterns from the specialized graphs produced by our pipeline.

After performing the analysis, a group of plots were formed showing similarities and differences between feature expression in different macrophages. These results were then compared to the actual study done1. From this it can be determined which macrophages would be more equipped to handle lung fibrosis.

**Methods:**

Data Accessibility and Tools:

To perform the analysis, Seurat was used specifically in Rstudio, which was the package that had to be installed from CRAN3. Macrophage data was studied by looking at a PBMC sample2 that was sequenced and downloadable as a folder with .tsv and .mtx files. The data was contained in .gz form so to put it in the correct format for Seurat, so the gunzip command had to be used to extract the files.

Filtration:

After installation, the data was read using the Read10x command which was then used to create a Seurat object with minimum cells of 3 and minimum features of 200. To begin analysis, the genes that are mitochondrial need to be filtered out. This can be done by the PercentageFeatureSet function looking for genes with the “-MT” flag indicating it is a mitochondrial gene. To visualize the distribution of the features, RNA count and percentage of mitochondrial genes, a violin plot was created (Figure 1). The violin plot gives some indication of what the threshold for filtering the data should be. The next step in the analysis was to normalize the data to keep data consistent and get rid of outliers. Based on the violin plot the thresholds can be determined as features > 200, RNA features < 2500, and percent mitochondria < 5%. Additionally, the data was normalized on a log scale. An additional step is to scale the data. This is done to reduce the dimensions of the data to fit the PCA plot format. After scaling, PCA (Figure 2) is done which calculates features that show high cell-to-cell variation. Specifically looking at these genes and features downstream can help illuminate the biological signal of the dataset.

Plotting:

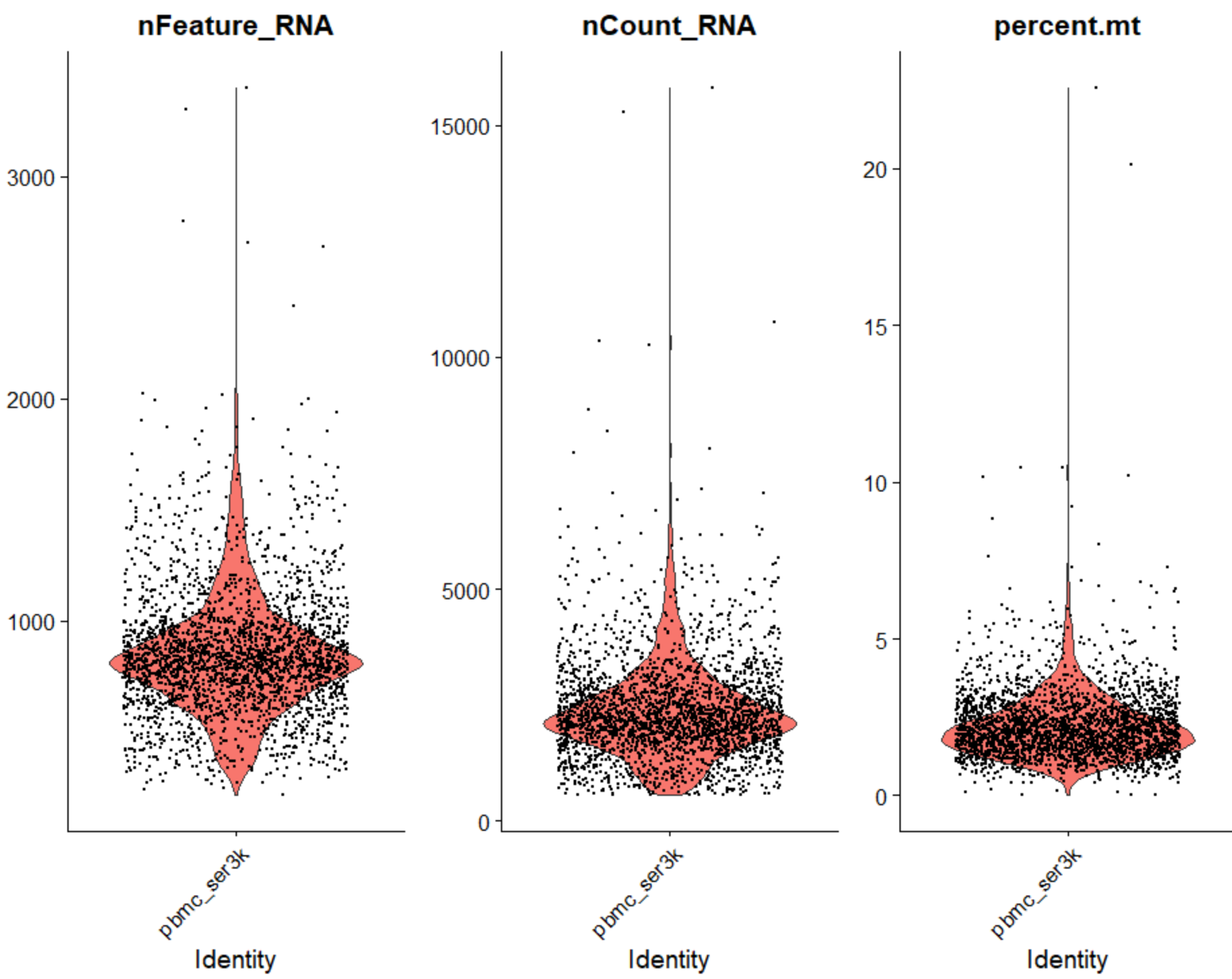
To finally understand the dataset, it is important to see what cell types cluster together indicating similarity in expression. tSNE plots are used to visualize this and it was created by the Run\_tSNE command (Figure 3). To further see association between different macrophages, a FeaturePlot is constructed and similar colored regions indicate similar expression levels (Figure 4). Specifically, the FeaturePlot focused on markers, "CD3E", "CD4", "CD8A", "CCR7", "SELL", "IL7R", "FOXP3", "GNLY", "NKG7”.

**Results:**

1. *Violin plots*

The violin plot produced aims to understand the distribution of the data being analyzed as well as the outliers in the data. This was important to understand the further filtration needed before continuation of plotting.

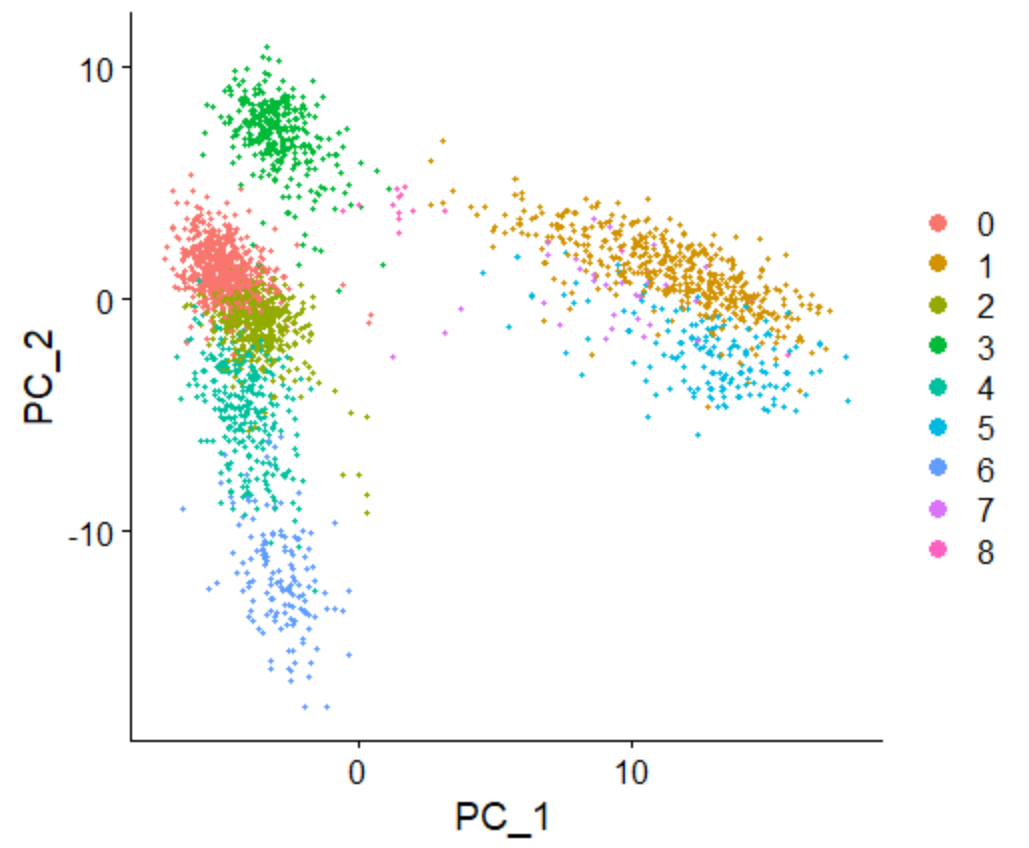
*Figure 1(a,b,c):*



The figure above plots the quality measures4 of features (a), counts (b), and the percentage mitochondria RNA plots (c) and plots the different quality measures of the three different categories (a,b,c). Further, these plots exemplify the discussion of distribution where the bulk of the data can be found near the median, while few outliers are found past the upper and lower quartiles. Data shown therefore indicated the need to further filter the data having unique feature counts over 2500 counts and less than 200. Similarly, this also indicated the use of filtering out the data with less than 5% mitochondrial counts.

1. *PCA*

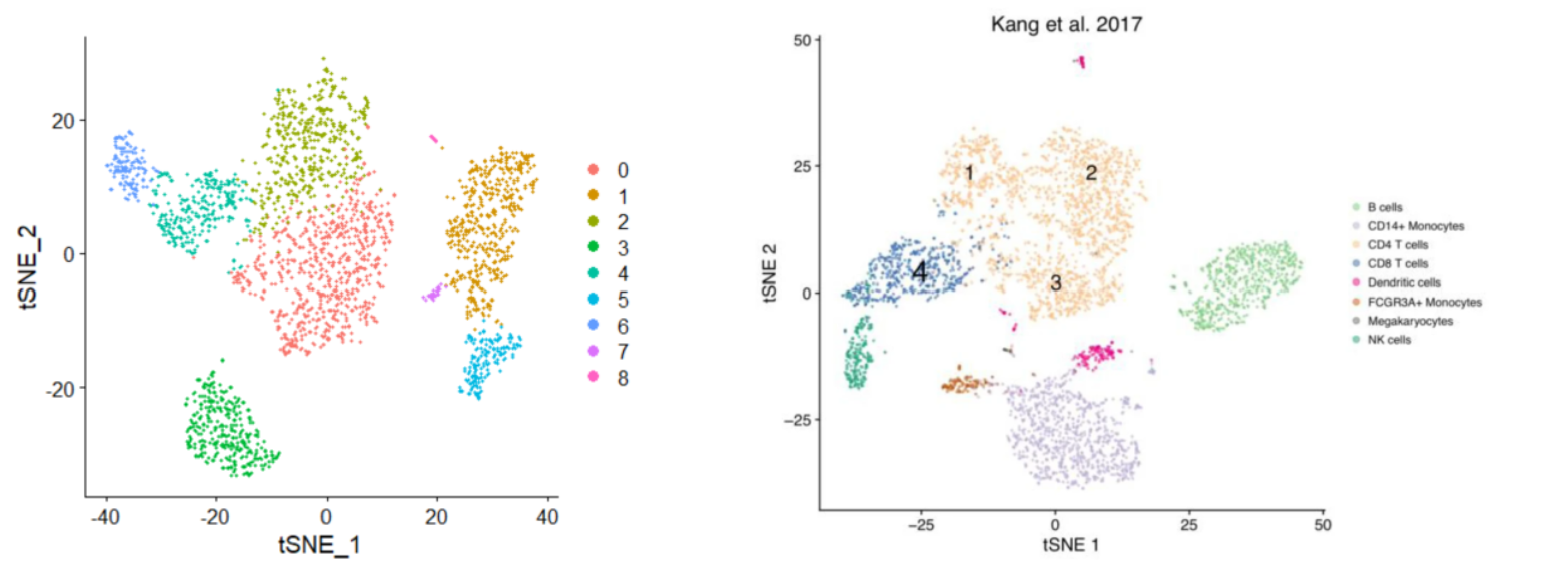
The PCA (principal component analysis) plot generated was used to better understand the patterns and trends in the dataset without omitting any of the data and which would in turn influence the plotting and clustering of the data. The main goal of including a PCA plot in this pipeline was to understand the different cells that exist in the dataset and the combination of features in the set that explains the most variation in the data used. Batch effects were not seen in this plot due to the usage of only one dataset.



After scaling the data and further filtration, the relationship between features was analyzed and reproduced to view the different clusters, neglecting batch effects.

1. *tSNE*

The tSNE plot is used to visualize the distribution of data through clustering where different colors represent different macrophages. This is done to visualize more clusters than a PCA would show and specify the clusters.

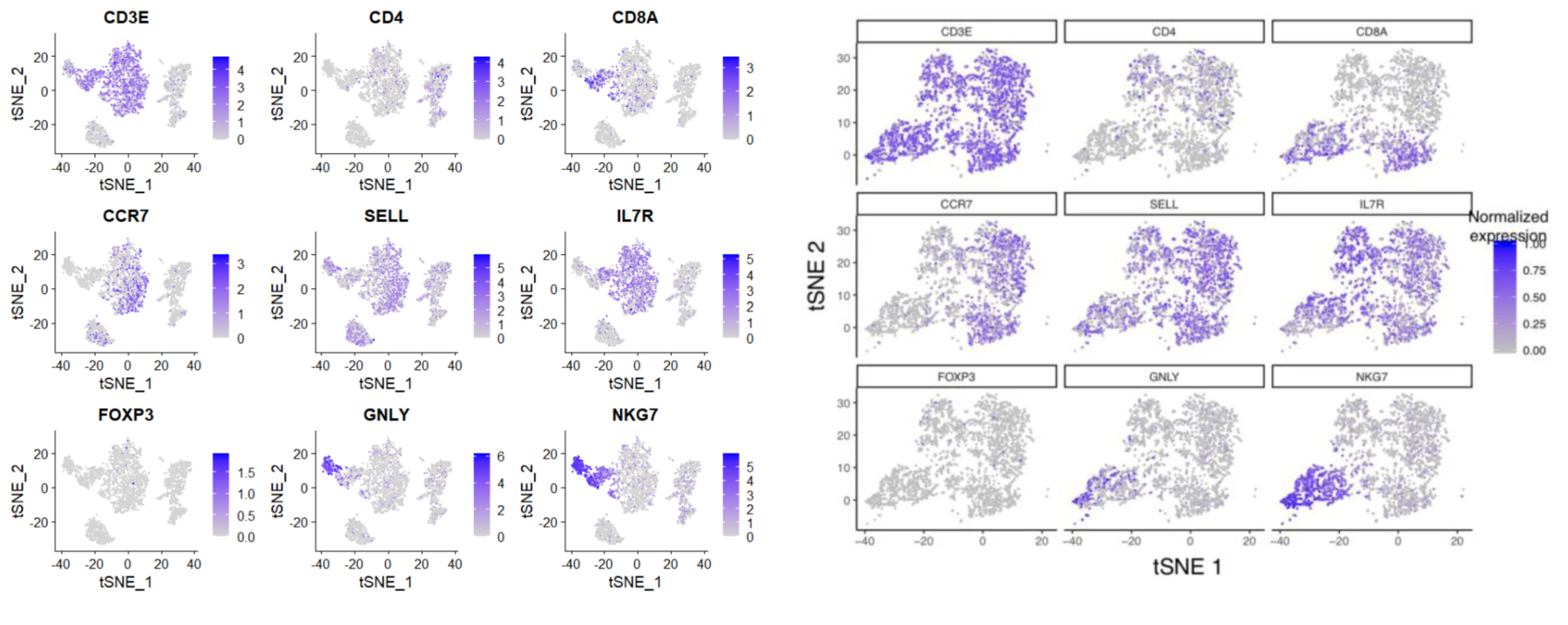
*Figure 3(a,b):* 

The plot in 3a is the one that was created through this analysis and the plot in 3b is the plot created by the actual study1 done using SingleR. Even though the shapes differ, the general clusters are relatively similar. Certain clusters can be matched to the referenced paper to determine what cell types fall into those regions.

1. *FeaturePlot*

The feature plot is a tSNE plot built specifically for each macrophage that is being observed. Each tSNE plot behaves similarly to Figure 3, however colors in this case indicate expression levels of features. Darker shades indicate more feature expression while lighter shades indicate less.

*Figure 4(a,b):*



With macrophages CD3E, SELL and IL7R having relatively similar regions of expression, it can be concluded that they are similar in their expression levels and may have similar functions to some level. Additionally macrophages GNLY and NKG7 have similar regions of expression so it can be deduced that they are similar as well. The difference in expression and shape between the plot produced in our analysis (3a) and the plot produced by the reference paper’s analysis (3b) can be correlated to the difference in the color scale as well as extra analysis done by the reference paper.

**Discussion:**

1. Limitations
2. The inability of access singleR:

Single cell recognition (singleR), the newly and privately developed software by the authors of the paper which was developed as a computational tool for the unbiased annotation and clustering of scRNA-seq data in the dataset. This software was used to add identifiers for the single cell transcriptomes which was done by comparing to the reference data set of the control (pure cell types) which was sequenced by RNA-sequencing.

We were unable to access this software and use it along with Seurat which caused discrepancies due to the extra sub clustering allowed due to the singleR normalization which was not done to correct our raw data. Therefore, as seen, there are differences in the plots generated in this graph and the original plots from the paper.

1. Subjectivity of data:

There are no current methods generated for the identification of the single cell types other than manual labor to annotate every cell type dealt with, which results in human error or subjectivity to each of the cell types. One may declare a cell type differently than another, which is something we also are victim to, resulting in the differences in the plots from the sample.

1. Challenges:
2. The first challenge faced in this pipeline was the usage of Seurat as a new package for the implementation of the plots. Seurat was not easily downloadable on JupyterLab due to permission problems; however, it was easier to download and install this software on R studio rather than jupyterLab or personal computer terminals.
3. The dataset was very large to run (about 5gb) which made the filtration and normalization of the data very tedious processes. The plotting of the data was also very time consuming due to the slow parsing of the data. This required the running of cells on JupyterLab for hours and restarting dead kernels to generate plots.
4. Another personal challenge was finding a viable paper with accessible data to use. Every paper found and curated was ultimately not a viable option due to the privacy terms of the datasets (like inability to access data due to the google cloud commands).
5. Improving the Pipeline:
6. To improve the pipeline, further understanding of the SingleR software would make the plots generated more similar to the paper. If we understand the usage of singleR, and what was done using singleR, it would be possible to recreate those commands and apply that in R studio using seurat to generate better plots.
7. In addition, it could have been possible to use another dataset that is similar to the data collected from PMBC but maybe in a different stage of maturation. Perhaps studying the growth of the macrophage into its differentiation can provide some further insight into the heterogeneity of macrophages. Doing this would require the use of Harmony which is needed to correct batch effects produced by running PCA on a combined dataset. Batch effects are systemic differences in the data that may be due to external factors. Removing these from the data analysis would ensure that no external variables affected the analysis.

**References:**

1Aran, Dvir, et al. "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage." Nature immunology 20.2 (2019): 163-172.

2<https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#normalizing-the-data-1>

3CRAN: <https://cran.r-project.org/web/packages/Seurat/index.html>

4Ilicic, Tomislav, et al. ``Classification of low quality cells from single-cell RNA-seq data." Genome biology 17.1 (2016): 1-15.

**Appendix:**

Code and Data: Gautham wrote the code and Nidhi helped in the data process and wrote the README.

Proposal: Gautham and Nidhi wrote the proposal.

Presentation: Gautham and Nidhi made the presentation, Je An presented the introduction, Gautham presented the data, and Nidhi presented the methods and limitations.

Report: Je An and Gautham wrote the Introduction, Gautham wrote the methods and half of the results section, and Nidhi wrote the limitations and half the results section